Data Science in Focus

Andrew Stewart

2025-04-12

Table of contents

Data Science In Focus 4

Data Science In Focus 5

Refining Practice, Deepening Understanding 5

Framing and Foundations 6

What Is Data Science 7

A Brief History of Data Science 8

Where Data Science Fits in Technology Organizations 9

Data Science as Applied Systems Science 10

Skillsets and Team Structure 11

A Refined Venn Diagram of Data Science 12

The T-Model of Skill Development 13

Data Science as a Team Sport 14

Research Management and Process 15

Managing Data Science Through Research Programs 16

The Scientific Method in Data Science 17

Using the Capability Maturity Model in Data Science 18

Tooling and Lab Practice 19

The Data Scientist's Lab Workbench 20

Documentation and Knowledge Repositories 21

Statistical Thinking 22

Probability and Statistical Inference 23

Data Science In Focus

Data Science In Focus

Refining Practice, Deepening Understanding

Data Science In Focus is a reflective essay series on data science as a scientific discipline—what it is, how it's practiced, and what it has become. As the field matures beyond its early hype cycles and into a coherent form of applied research, this series aims to sharpen our collective understanding of the work itself.

Where *Fundamentals of Data Science* laid the groundwork for a newly forming field, these essays revisit core questions with the benefit of hindsight:

- What does it mean to practice data science as science?
- How should teams, tools, and systems support inquiry over output?
- What kind of knowledge does data science produce—and for whom?

Rooted in the scientific method, structured around the research lifecycle, and steeped in the evolving norms of modern tech orgs, this series puts the discipline itself into focus.

Framing and Foundations

What Is Data Science

(Full draft previously provided. You may replace the placeholder with the detailed version.)

A Brief History of Data Science

Data science didn't emerge from a single lineage—it was born from a convergence. Statistics, scientific computing, database management, and machine learning all played formative roles in shaping the field. During the early 2000s, as computational infrastructure and open-source tools proliferated, organizations began to realize that "data" wasn't just a byproduct of digital systems—it was a source of insight.

The rise of the "data scientist" as a role came partly from pragmatism: organizations needed generalists who could both analyze and code. But this era was also defined by confusion. Data science became a catch-all title encompassing analysts, engineers, ML researchers, and more. The result was a temporary inflation of the field—a golden age of generalists with ambiguous scope.

Now, a decade later, the landscape is maturing. Specialized engineering roles have splintered off. Machine learning engineering, analytics engineering, and decision science have defined clearer scopes. What remains in the heart of "data science" is its original essence: a scientific discipline grounded in inquiry, system behavior, and the pursuit of explanatory knowledge.

Where Data Science Fits in Technology Organizations

Data science occupies a unique niche within technology orgs. It bridges engineering, product, and strategy—but it is fundamentally distinct from all three. Its core deliverable is understanding, not execution.

The organizational structure of data science varies:
- **Centralized teams** emphasize consistency and shared standards.
- **Embedded models** prioritize domain intimacy and responsiveness.
- **Hybrid approaches** attempt to balance autonomy and alignment.

Tensions often arise when data science is treated as a service function, expected to deliver dashboards or one-off analyses on demand. But a scientific function thrives on longer-term research questions, context-rich collaboration, and space to explore uncertainty. Successful teams recognize this and give data scientists both embedded partnerships and protected research capacity.

The most effective data science orgs invest in career ladders, program-based workstreams, and a clear cultural distinction between scientists and engineers—while encouraging tight collaboration across them.

Data Science as Applied Systems Science

At its core, data science is the study of complex, adaptive systems. These systems—markets, platforms, recommendation engines, networks—are not static. They evolve, respond to feedback, and often exhibit emergent behavior.

To understand such systems, data scientists borrow heavily from adjacent fields:
- **Control theory** to manage dynamic processes
- **Cybernetics** to study feedback loops
- **Complexity science** to model emergence
- **Information theory** to reason about signal and noise

This systems-thinking mindset distinguishes data science from business analytics or product instrumentation. It frames metrics as proxies, not truths. It treats models as lenses, not deliverables. And it embraces the recursive nature of systems—where measurement affects behavior, and knowledge must continually update in response.

Skillsets and Team Structure

A Refined Venn Diagram of Data Science

Earlier attempts to define data science visually—such as Conway's Venn diagram—centered on hacking skills, statistics, and domain knowledge. While influential, that framing is now outdated.

A more precise picture situates data science at the intersection of:

- **Statistical modeling** — tools for inference, uncertainty, and causality
- **Scientific computing** — numerical methods, simulations, and computation
- **Systems research** — studying complex, interactive software environments

This triad places data science alongside fields like econometrics, computational physics, and quantitative social science. It emphasizes rigor and replicability. And it highlights that data science is not just about *using* data, but about *understanding the systems* that produce it.

The T-Model of Skill Development

Data scientists develop along a T-shaped trajectory:
- The **horizontal bar** represents breadth—exposure to tools, methods, and adjacent domains.
- The **vertical bar** represents depth—specialization in one or more areas like causal inference, simulation, or optimization.

This model supports differentiated roles:
- A generalist might help design broad research programs.
- A specialist might focus on methodological innovation or platform-level modeling.

Title tracks should reflect this diversity. Rather than forcing all scientists into a "full-stack" mold, orgs can recognize distinct paths:
- Researcher
- Methodologist
- Domain expert
- Tool-builder

This structure also helps build effective teams—ones that combine varied skill sets across the disciplinary landscape.

Data Science as a Team Sport

(Organizing scientists into labs with complementary strengths across the venn.)

Research Management and Process

Managing Data Science Through Research Programs

(Differentiating research from projects; long-term, iterative, institutional memory.)

The Scientific Method in Data Science

(Hypothesis-driven study design, using Jira and other tools to support structured inquiry.)

Using the Capability Maturity Model in Data Science

(Applying the CMM to assess and scaffold experimentation and system design.)

Tooling and Lab Practice

The Data Scientist's Lab Workbench

(SQL as data collection, Python as instrumentation, Jupyter/Markdown/Quarto as lab notebooks.)

Documentation and Knowledge Repositories

(Writing for different audiences, synthesis vs analysis, memory and communication.)

Statistical Thinking

Probability and Statistical Inference

(Quantifying uncertainty, validating observations, and the logic of belief.)