USC University of Southern California

AI-enabled Discovery:
The Digital Scholarly Record

Victoria Stodden
University of Southern California

Joint Quantitative Brownbag

The Ohio State University, University of Maryland, University of North Carolina at Chapel Hill, Notre Dame University, University of Virginia, Vanderbilt University, and University of South Carolina

September 9, 2024

USC

Agenda

1. Effects of leveraging AI on scientific research and discovery

2. *Disruption* in scientific norms: transparency, accountability, reproducibility

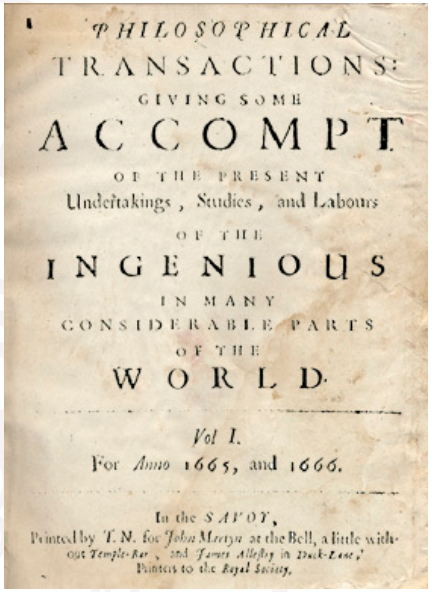
3. The emergent *Digital Scholarly Record* and why it's important

USC

The Scientific Record: Touching the *Spring of the Air*

The Royal Society of London founded 1660 (the "Invisible College")

- members discussed Francis Bacon's "new science" from 1645,
- Society correspondence reviewed by the first Secretary, Henry Oldenburg, who became the founder, editor, author, and publisher of the first scientific journal in 1665; Philosophical Transactions.


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol. I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

Boyle wrote, "**It is much more difficult than most men can imagine, to make an accurate Experiment**" (*Certain Physiological Essays And Other Tracts: Written at Distant Times, and on Several Occasions By the Honourable Robert Boyle,* 1673)

In *New Experiments Physico-mechanicall, Touching the Spring of the Air and its Effects* (1660) Boyle set the standard for scientific communication:

1. Enough detail on equipment, material, and procedures, for reproducibility
2. "Communal witnessing"
3. Exhaustive details on experimental settings, false starts, failures, etc.

USC

Science 2024: Unstoppable Forces

Highly complex computation, data, and integrated scientific workflows:

- Open, transparent, re-executable machine learning pipelines, shared on common infrastructure
- Open Data

Deeply disruptive innovation in scientific discovery:

- New discovery methods: *Common Task Framework*
- Leveraging LLMs

- ➢ Research shared natively digitally or not at all
- ➢ Publishing a pdf is an afterthought…

Assertion 1: Boyle's standards for publication are defunct.

USC

A Thought Experiment..

Say:

- LLMs are now leveraged for metadata completion for integrated and accessible datasets
- Standard benchmarking of a “clear and concise definition of a solution”
- Black box pipelines are routinely crawled to find “best” solutions

Is the solution correct? Well, just ask an LLM.

Assertion 2: We can no longer judge correctness since we no longer understand the chain of reasoning that gave the results [1].

[1] Stodden, V. (2024). On Emergent Limits to Knowledge—Or, How to Trust the Robot Researchers: A Pocket Guide. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.dcaa63bc

Corollary:

A De Facto Digital Scholarly Record (for the robots)

A vision: In 2050 the scholarly record will be a detritus of organically connected frameworks, training data, and leaderboard results.

Correctness will be established not by transparency and human verification, but by direct checks in a new meta-Bacon methodology:

1. verification of conclusions by direct experiment **on black boxes**,
2. discovery of truths unreachable by other approaches,
3. investigation of the secrets **of black box discovery methods**, opening us to a knowledge of past and future.

We don't Touch the Spring of the Air, but instead generate a collection of useful results that work.

Idea: The chain of logic behind the discoveries is accessible via LLMs and impenetrable code (to humans) comprising yet-to-be-developed specialized scientific discovery pipeline querying tools.

USC

Change 1. Wither Authorship

Since Boyle, authorship has been an important part of accountability.

LLMs and other large model leveraged discoveries break the chain of authorship:

- Who's the author?
- Whose ideas does this work build on?
- What ideas does the result build on?

It no longer matters whether the reasoning is human accessible. As long as we can query to obtain the results..

USC

Change 2. Knowledge as Utility not Understanding

In an AI-enabled future, researchers (and everyone else) believe results are correct since they trust their (opaque) discovery process:

- meaningful precision of the problem description,
- correctness of the benchmarks,
- appropriateness of the data.

A Scientific Method for Challenges?

- What problems lend themselves best to this approach?
- What data attributes are essential for success?
- Who wins who loses?

A Successful Machine Translation Research Program

(adapted from [2] Liberman 2011)

In 1986 Charles Wayne, a DARPA program manager, reinvigorated the defunct machine translation program with:

1. a well-defined, *objective evaluation metric* applied by a neutral agent (NIST) on shared data sets, to protect against “glamour and deceit” from “mad inventors” and “untrustworthy engineers” (Pierce, 1969);
2. *participants revealing their methods* to the sponsor and to one another when the evaluation results are revealed, to ensure that “simple, clear, sure knowledge is gained” (Pierce, 1969).

The Common Task Framework

(adapted from Liberman 2011)

A detailed evaluation plan:

- developed in consultation with researchers
- and published as the first step in the project.

Automatic evaluation software:

- written and maintained by NIST
- and published at the start of the project.

Shared data:

- Training data is published at start of project
- Test data is withheld for periodic public evaluations.

Integration of Data, Code, Results

Create a verifiable and extensible base in a systematic and open way, facilitating:

1. regeneration of a computational results/models;
2. comparisons and reconciliations of different hypotheses;
3. the reimplementation of methods on new data and update methods;
4. the generation and evolution of benchmarks and standardized testbeds for the assessments of models and inference methods;
5. the development and application of appropriate policies regarding data privacy, ethics, and meta-research on the scholarly record.

A Computable Digital Scholarly Record II

Such an entity acts as a traditional scholarly record:

- forms a locus for a research community to share ideas, get feedback, improve their work, agree on priorities, and resolve debates.

Integrates AI enabled affordances:

- Document and trace contributions and authorship,
- Expose pipelines and (automated) testing.

USC

No One Knows How and Nobody Cares

1. Models routinely contain billions of parameters, so a direct interpretation of the mechanism of response creation is far out of reach.
2. A focus on producing results that satisfy widely recognized benchmark performance goals, not a cognitively tractable explanation of the underlying phenomena.
3. The guardrails for knowledge production processes become incredibly important, as we can no longer solely rely on our usual mechanisms to assess research: peer review, disclosure of transparent methods for (human) verification, and the independent reproduction of findings.

A Scientific Method for Challenges

Our task is to adapt the standards of the scientific discovery process to decide whether:

1. The quantitative goal of the challenge problem is sufficient to support scientific conclusions?
2. Can we understand why the winning method wins?
3. Can we understand how and when to generalize challenge problem findings?

The Data

Information about the data can increase trust in the result:

- With how much fidelity does the data span the true underlying population of interest?
- Every data set is a sample, and so effective measures of representativeness can guide where we expect applications of the resulting model to be reasonable, in other words, trustworthy.
- Was the test set sampled from the input data or is it a new sample, likely with greater variability?

Open Code and Re-execution of Pipelines

1. To what extent are the results and models architecture dependent?
2. How can effective software tests be designed and routinely expected to increase trust in the functioning of the code?
3. Can aspects of the software testing or model parameter estimates be shared in a verifiable way, since often these models and pipelines require significant resources at scale to estimate and retrain?
4. What aspects of the implementation pipeline and the resulting trained model *must* be made available to the research community for verification purposes?

Challenge Problem Approach

1. Are there researchers who will be left out of participating in knowledge production?
2. Does the objectivity of accepted results change?
3. How do we give credit to discovery pipeline creators, especially when solutions and code are reused and extended for new challenge problems?
4. How does the challenge approach intersect with the traditional rationale for scientific discovery: a love of wonder and the excitement of explaining and understanding the natural world?

Who wins? Who loses? How do we know?

- Who has access to requisite compute power? Models? Weights?
- Who engages in competition style discovery?
- What's the motivation for researchers?
- ?

The Emergent Digital Scholarly Record

Integrate computational knowledge for:

- a verifiable and extensible open knowledge base,
- synthesizing computationally and data-enabled discoveries,
- reusable discovery pipelines,
- regeneration of a computational results or models,
- comparison and reconciliation of different conclusions,
- reimplementation of methods on new data,
- the generation and evolution of benchmarks and standardized testbeds,
- development and application of appropriate policies regarding data privacy, ethics.

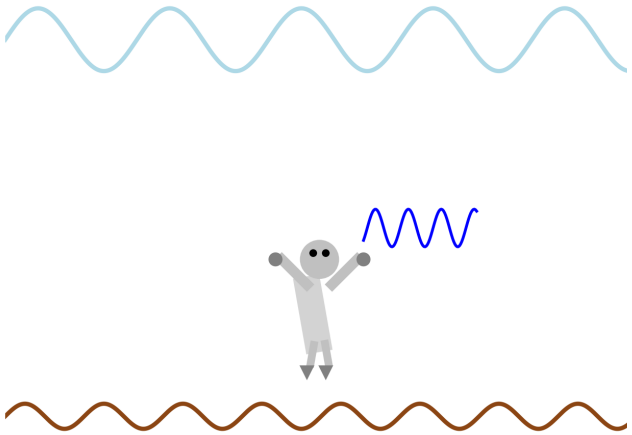
All perhaps in automated ways…

Conclusions

Challenge problems and large opaque models are coming to dominate scientific research, leverage massive compute and dat infrastructure.

Traditional ideas foundational to science break:

- Authorship and accountability,
- Human accessible chains of scientific rea
- An emergent digital scholarly record.



Here's a surrealist-style drawing of a robot Touching a Spring in the Air. The robot is depicted with abstract and distorted features, set against a dreamlike landscape with flowing, surreal shapes. The spring maintains a twisted, helix-like form to add to the surreal ambiance. (ChatGPT4)

USC

References

[1] Stodden, V. (2024). On Emergent Limits to Knowledge—Or, How to Trust the Robot Researchers: A Pocket Guide. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.dcaa63bc

[2] Mark Liberman (2011) Lessons for Reproducible Science from DARPA's Progams in Human Language Technology, The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer, AAAS Annual Meeting 2011.

[3] Donoho, D. (2024). Data Science at the Singularity. *Harvard Data Science Review*, 6(1). https://doi.org/10.1162/99608f92.b91339ef

USC