Data Science at the Singularity

David Donoho[†,*]

[†] Stanford University

ABSTRACT. *Something* fundamental to computation-based research has *really* changed in the last ten years. In certain fields, progress is simply dramatically more rapid than previously. Researchers in affected fields are living through a period of profound transformation, as the fields undergo a transition to *frictionless reproducibility* (FR). This transition markedly changes the rate of spread of ideas and practices, affects scientific mindsets and the goals of science, and erases memories of much that came before.

The emergence of FR flows from 3 data science principles that matured together after decades of work by many technologists and numerous research communities. The mature principles involve data sharing, code sharing, and competitive challenges, *however* implemented in the particularly strong form of frictionless open services.

Empirical Machine Learning is today's leading adherent field; its hidden superpower is adherence to frictionless reproducibility practices; these practices are responsible for the striking and surprising progress in AI that we see everywhere; they can be learned and adhered to by researchers in whatever research field, automatically increasing the rate of progress in each adherent field.

Keywords: Reproducible Computational Research. Challenge Problems Paradigm. Frictionless Reproducibility. Frictionless Research Exchange. Emergent Superpower. AI Singularity.

MEDIA SUMMARY

Overheard in conversation...

(Outsider): Things are happening awfully fast these days! I frequently see advances in data processing – text, video, tabular data – coming at a rate I hadn't been expecting to see. This makes it look like there's a superpower out there. Maybe it's AI?

(Insider): No, actually it's data science.

(Outsider): What's the data science superpower?

(Insider): Okay, on the fingers of one hand:

(1): In the last decade, frictionless services became available thanks to the modern smartphone information ecosystem

(2): Those frictionless services were applied by Scientists and Technologists to Data Sharing, Code Sharing and Challenges

(3): Some communities of researchers started frictionlessly sharing research artifacts – code, data, results – and building on each others' work.

(4): Involved research communities are progressing much faster. It's night and day.

(5): Soon it will *really* speed up, as scientists write computer programs to exploit all the digital artifacts from previous research.

(Outsider): But wait, I'm still reading everywhere about AI.

(Insider): AI is one of those communities where people are working this way, that's why you're hearing so much news about it.

(Outsider): But I heard that there's a singularity, that AI is going to unemploy everybody and may kill us all.

(Insider): There is a singularity (= new superpower) but not in AI. It's a superpower allowing us to do data science research faster and better. It's not dangerous. You'll just be hearing more and more frequently about advances, because research goes faster and faster.

1. INTRODUCTION

The last 10-15 years witnessed a dramatic transformation of information technology, as the smartphone spread across the planet, spawning omnipresent internet connections, and always-on cloud computing capacity. The result was a unified global computing/communications resource now driving nearly every industry and transforming nearly every human sphere of activities.

Information is now flowing to peasants in developing economies in a fashion inconceivable not long ago; this is driving historically unprecedented declines in global poverty and ignorance.

Our collective intellectual life is also transforming rapidly. Scientists today have a completely different set of ideas about what can be learned from data, and how to go about learning it, than only ten years ago. There are many prominent stories of dramatic recent changes in science's ability to make sense of data, the most well-known being in the fields of computer vision, natural language translation, or protein folding, but examples are springing up nearly everywhere. We all know something has happened, and there's no going back; but we may not yet perceive the larger, implicit, emergent phenomena, causing so many dramatic changes, at the same time, in so many fields.

The media and scientific public view these many signs of progress as signs that **mysterious superpowers** have somehow **emerged**, but are secreted away in hegemon labs, and need immediate political regulation. Memes of 'AI Singularity' are widely disseminated, often alongside hints of 'AI Doom'. Politicians on the global stage are now worried.

In a companion paper, I argue that "AI Singularity" offers a misleading narrative, which can't explain a wide range of new developments and which hides from view the true driving forces. In this paper, I argue for a different narrative: that computation-driven research really *has* changed in the last ten years, driven by 3 principles of Data Science, which, after longstanding partial efforts, are finally available in mature form for daily practice, as frictionless open services offering **data sharing**, **code sharing**, and **competitive challenges**.

Researchers in fields adhering to these principles find themselves living through a period of profound, rapid transformation. Research life after adopting these principles operates at a faster pace: good ideas spread and are adopted, and improved upon, seemingly with very little friction (eg human labor) compared to earlier decades. We are entering an era of frictionless research exchange, in which research algorithmically builds on the digital artifacts created by earlier research, and any

good ideas that are found get spread rapidly, everywhere. The **collective behavior induced by frictionless research exchange** is the **emergent superpower** driving many events that are so striking today.

To discuss these data science principles and their singular effects, I first discuss the larger civilizational changes enabling them; I then review the recent public discussion of singularity for eventual comparison with my thesis; and then discuss the dramatic acceleration in research progress which happens when a research discipline adopts the new practices and crosses this singularity.

I then mention actions individual readers can take in response to the arguments presented here, including in research and teaching; and I make predictions about future developments on the other side of the singularity.

2. Explosion of accessible compute power

Moore's "Law" was, from 1950-2000, an engine of continual improvement in computing speed and power, but then it stalled out, post 2005, as decreases in the size of transistors no longer allowed reductions in processor cycle times. Amazingly, progress in compute power *accelerated* since the 'death' of Moore's law.

The powerful instigator was a marked one-time 'smartphone transition' dividing human history into periods 'before' and 'after' smartphones. Over the first fifteen years of the smartphone era, smartphones proliferated rapidly; today purportedly 80%+ of all adult humans have access to this technology either directly or through friends and associates[1]. Smartphones rely upon a global computational and communications infrastructure whose rapid construction has been a major civilizational achievement, with myriad consequences and corollaries. These include undersea fiber, satellite internet, arctic data centers, cloud computing, novel database technologies and a massive reconfigurable global workforce of software developers. This workforce envisioned and created a range of organizational tools orchestrating virtual armies of compute servers; as a result, the compute capacity that could be easily thrown at a single task grew perhaps 100,000 times from when the smartphone era began[2].

The implications for computational science were unprecedented. Even a negligible fraction of existing global compute capacity could suddenly allow computational scientists to be ambitious at previously unimaginable scales. But beyond this unexpected windfall, computational science independently and organically drove additional bounties; for example, hardware hegemons invested heavily in GPUs for numerical computing, pushing stock market valuations beyond anyone's expectations. Combined with novel machine learning models and a whole industry category – MLOps[3]

---

[1]There are many sources in the same ballpark, from 90%+ penetration by cellphones of all different generations to 68% penetration by latest generation smartphones. An example consistent with this figure is given by the blog post of Josh Howarth, Dec 14, 2023: https://explodingtopics.com/blog/smartphone-stats

[2]This is wild understatement. Indeed, discussions of the Dask open source scheduler advertise control of 100K cluster nodes without trouble; such clusters exist at the big public clouds; the nodes in such clouds have perhaps 20x more core per node than in the mid 2000's and perhaps 20x more RAM; along with better memory caching.

[3]Machine Learning Operations; from `Databricks.com`: MLOps stands for Machine Learning Operations. MLOps is a core function of Machine Learning engineering, focused on streamlining the process of taking machine learning models to production, and then maintaining and monitoring them.

– for specifying and deploying machine learning workflows, a further 1000-fold increase in compute power has been delivered by new ways to organize machine learning computations[45].

Importantly, all this infrastructure is now available immediately, for a price, to anyone who will pay.

This new world order in computing offered all of us a vast expansion in functionality and scale - while remaining true to the original purpose of the smartphone era: to enable humans to spontaneously share information among themselves. The new world order was committed to frictionless open sharing of digital artifacts and substantial computing in response to spontaneous user requests. This **frictionlessly available data and compute infrastructure** could be ingeniously adapted for many purposes, including computation-driven research.

3. Data Science Matures

The new world order in computing crystallized into maturity three long-ongoing sociotechnical initiatives spreading throughout science. The three initiatives have been aspirations of many data scientists across decades, and the coming-to-fruition moment, where everything needed is now in place and immediately accessible, has been reached by the hard work of many technologists.

The initiatives are related but separate; and all three have to come together, and in a particularly strong way, to provide the conditions for the new era. Here they are:

[FR-1: Data]: datafication of everything, with a culture of research data sharing. One can now find datasets publicly available online on a bewildering variety of topics, from chest x-rays to cosmic microwave background measurements to Uber routes to geospatial crop identifications.

[FR-2: Re-execution]: research code sharing including the ability to exactly re-execute the same complete workflow by different researchers.

[FR-3: Challenges]: adopting challenge problems as a new paradigm powering methodological research. The paradigm includes: a shared public dataset, a prescribed and quantified task performance metric, a set of enrolled competitors seeking to outperform each other on the task, and a public leaderboard. The paradigm can also include *virtual challenges* lacking the formal leaderboard, in which authors still attempt to publish a new state-of-the-art result going beyond previously recorded/published performance levels on a given dataset/performance metric. Thousands of such challenges and virtual challenges, with millions of entries, have now taken place, across many fields.

   Note: the term **benchmarks** is frequently used to describe what we here call virtual challenges; but it can be unclear whether this term refers to the process of quantifying performance, or instead to the specific value of certain quantitative measures in a particular case.

Each initiative addresses aspirations in scientific research as old as data-driven science. Researchers have always wanted to access the same data used by their predecessors in earlier studies;

---

[4]For example a single Dask array can implicitly distribute a single virtual array across many physical nodes, implementing a single matrix add operation (or other) virtually across a whole cluster. In my Stats 285 class at Stanford we used Dask to reach matrix sizes by 1000 beyond the capacity of a single node.

[5]Nvidia's Jensen Huang: "Accelerated computing looks at the entire stack, algorithms, software and processor. ... New software systems make the application go faster, not just the chip."; Moore's Law is dead.

or apply the same algorithms from those studies; or to score performance of new algorithms in later studies according to the same metrics used in earlier studies. However, those wishes were generally frustrated in various ways in earlier generations.

In the past decade, all of these initiatives, in their best implementations, became **frictionless open services**, essentially offering immediate, permissionless, complete, access to each relevant digital artifact, programmatically from a single line of code. Not all friction is yet gone, but the trend is clear: what frictions still remain will vanish soon enough.

Today's researchers will know many specific examples of initiatives [FR-1]+[FR-2]+[FR-3] coming together as frictionless open services in the last decade. Notable ones include:

[FR-1]: Research Data repositories: Open Science Collaborative[6], Nightingale Open Science [7];

[FR-2]: Research code re-execution environments like Jupyter [8]; and

[FR-3]: Challenge problem platforms such as CodaLab Competitions [9].

(Prominent commercial examples for [FR-1]-[FR-3] exist as well; eg. for [FR-3], Kaggle[10], Wazoku Crowd (originally InnoCentive)[11], Sage Bionetworks[12]; for [FR-2], Posit Cloud, Azure Notebooks, Google Colab. Commercial frictionless data access: Data.World (general); Bloomberg (financial); MapBox (street maps); Maxar (Satellite Imagery). Non commercial, non-research examples: [FR-1] Data.Gov (US); European Data; Our World in Data).[13][14]

What were once mostly-frustrated wishes of researchers are today achieved realities in several fields, supported by habits and institutions. Most importantly, where the behaviors are not yet

---

[6]https://osf.io/

[7]https://www.ngsci.org/

[8]http://jupyter.org

[9]https://codalab.lisn.upsaclay.fr/

[10]See for example https://www.kaggle.com/code/ashagutlapalli/evolution-of-kaggle, accessed December 21, 2023

[11]See for example [Wilan, 2007]

[12]https://sagebionetworks.org/

[13]Each of these commercial platforms has a large footprint outside of the scientific methodology research we focus on here. Although they are important resources for FR as discussed here, they are of interest for many other reasons. For example, Kaggle and Google Colab are very important for education and for training. Wazoku Crowd can accommodate challenges outside the computational methodological realm - including challenges involving materials or processes in chemistry, manufacturing, environmental, medicine, or biology.

[14]Two of the commercial challenge providers – Wazoku Crowd (ex InnoCentive) and Sage Bionetworks – are spinouts of the Pharmaceutical industry. There are interesting origin stories; InnoCentive spun out of Eli Lilly, where Alpheus Bingham was a PhD entrepreneur who more than 20 years ago saw the scientific and commercial potential in challenges [Wilan, 2007]; while Stephen Friend and Eric Schadt formed Sage Bionetworks with backing from Merck; they considered themselves on a mission to save pharmaceutical research integrity through challenges.

dependably present, there is no essential obstacle to turning them into everyday habits, except the interest and diligence of participating researchers[1516].

This is a big change from the recent past! The last few decades had been a period of experimentation and prototyping, trying out various approaches to sharing research data, challenge problems, and research code. Sometimes, in principle, we could reproduce the research of others. But in various ways there was still clumsiness and friction.

Then came the breakneck development of information technology driving our post-2007 smartphone era. The emergence of 'The Cloud' gave us tools for web-accessible database management, and source code management for workgroups which made it possible for information technology organizations to be distributed globally yet work together.

Building on the new capabilities, developers implemented globally accessible code and model repositories, data repositories, and, eventually, globally visible task metric leaderboards.

These services and their use patterns are well known to modern data scientists. Research communities can now organize themselves to model their research publication process around the use of these services.

Among research communities, Empirical Machine Learning (EML) has gone the farthest, its members today can use accepted commercial services including Kaggle, GitHub, HuggingFace, and Weights and Biases for storing and accessing these artifacts. Quite literally, today's model for academic EML research requires accessing the various artifacts created under [FR-1]+[FR-2]+[FR-3] and then modifying some of them in some way and then sharing those modified artifacts. Articles describing such research, which we discuss later below, are simply highly stereotyped press releases advertising the availability of these artifacts.

Step back from the EML context for the moment. The same process is ready for immediate deployment by computational scientists more generally.

4. Emergence of Frictionless Reproducibility

The maturation of data science initiatives [FR-1: Datafication and Data Sharing], [FR-2: Code Sharing and Re-execution], and [FR-3: Challenge Problems] in their current form as essentially frictionless services brings us to a phase transition in computation-driven research, to a new era of *frictionless reproducibility* in data-driven methodological development. This new research era features unprecedented rates of advance.

Indeed, today, a computational scientist working in one of those fields where all three practices are standard, can, with a few mouse clicks, and maybe a little extra typing, access everything needed to reproduce some recently reported experimental result from some other institution: code, data, experimental analysis. The scientist can be running some other team's original experiment or a modified computational experiment within minutes or even seconds of hearing about it. In many

---

[15]A referee points to data sharing obstacles in some fields; for example in clinical medical research, where patient privacy laws act as constraints. However, even in such cases, initiatives spring up to bridge the obstacles. Nightingale Open Science https://www.ngsci.org/ extendsthe benefits of the 'standard' FR-1 data sharing model to the protected health information space.

[16]A referee points out obstacles to code sharing - such as the profit motive. (See the discussion of *Proprietary Methods Challenges* in Section 4 below). As in the previous footnote, initiatives have sprung up to bridge such obstacles. The AlphaFold protein folding system won two recent CASP protein folding challenges; while the code was not released to the public, *its folding predictions are available* **as a frictionless service** hosted through the European Microbiology Laboratory https://alphafold.ebi.ac.uk/.

cases, the scientist doesn't even really need to be manually involved; a predefined script can do all the work, completely automating the data and code access, and running the experiment – possibly with novel modifications or measurements layered on top.

In contrast, progress seems notably slower in a field where only two (or one, or none) of these initiatives have yet come to fruition, rather than three[17].

The most common leave-one-out setting is surely **Reproducible Computational Science** (RCS) where we combine [FR-1: Data Sharing] and [FR-2: Code Sharing], without [FR-3]. Here there is a scientific question but no underlying challenge problem being considered; we might simply be doing an exploratory data analysis and reporting what we saw, and giving others access to the data and the analysis scripts. RCS lacks the ability to focus attention of a broad audience on optimizing a performance measure.[18]

A less common but very famous and important setting is the **Proprietary-Methods Challenge** (PMC), with [FR-1: Data] and [FR-3: Challenges] but not widespread code sharing by participant teams, even after the fact. The \$1 Million Netflix Challenge (2009), and many large-dollar prize challenges since, are examples.

Least common by far are **Bring-Your-Own-Data Challenges** (BYODC), ie. offering [FR-3: Challenges] - [FR-2: Code Sharing] without [FR-1: Data Sharing]. This can happen in clinical medical settings where the research focuses on predicting patient outcomes but the underlying data is private and only a few credentialed researchers ever get to see it, under nondisclosure. Here, the research publications may share task definitions and results and also algorithms, without sharing the data. Clinical medical research overall is a field with about 1 million research publications per year.

Without all three triad legs [FR-1]+[FR-2]+[FR-3], FR is simply blocked. In two of the above leave-out-one scenarios, this is obvious. Without [FR-1 – Datafication and Data Sharing] we would be missing the data; without [FR-2 – Code Sharing and Re-execution] we would be missing the opportunity to inspect and build upon the workflow.[19]

Less clear is what we might be missing without [FR-3 – Challenges]. We would be missing the task definition which formalized a specific research problem and made it an object of study; the competitive element which attracted our attention in the first place; and the performance measurement which crystallized a specific project's contribution, boiling down an entire research contribution essentially to a single number, which can be reproduced. The quantification of performance – part of practice [FR-3] – makes researchers everywhere interested in reproducing work by others and gives discussion about earlier work clear focus; it enables a community of researchers to care intensely about a single defined performance number, and in discussing how it can be improved.[20]

---

[17]The next 3 embolded terms have been coined to simplify this exposition; I make no claim that they are widely accepted.

[18]This setting is often called *Open Science*.

[19]Again we insist on programmatic access to digital artifacts.

[20]We may still get the benefit of challenges, without having a formal leaderboard and participant registration; this can happen when we have a *virtual challenge*, with public data, code and benchmark criterion, as can happen with formal publication of a research article aiming to establish a new state-of-the-art benchmark result, and adhering to open science practices.

If we only have...	We are blocked, because	Example
[FR-1] + [FR-2]	No defined task	Exploratory Data Analysis
[FR-1] + [FR-3]	Can't build on code of others	Netflix Challenge; DARPA Biometric Challenges
[FR-2] + [FR-3]	No Common Dataset	Human Subjects Clinical Research

Table 1. Leave-One-outs, and what is blocked

Of course, not every field works this way today[21], but those that do, commonly benefit from very high velocity of progress. Frictionless reproducibility spontaneously spawns groups of inspired researchers to a tight loop of iterative experimental modification and improvement, often engaging many scientists. This tight loop per researcher, combined with iterative feedback from cross-researcher competition, results in a rapid series of improvements in the performance metric. The outcome of the process is much higher performance on the task metric.

An emergent property is manifest here. The researchers in an adherent field, understanding that they are all subscribed to the FR triad, think differently about what they do/should be doing, behave differently and develop habits and practices that they wouldn't have outside the FR setting. We see a *new institution* arising spontaneously; let's call it a **Frictionless Research Exchange** (FRX).

FRX is an exchange, because participants are constantly bringing something (code, data, results), and taking something (code, data, new ideas), from the exchange; and various globally visible resources - task leaderboards, open review referee reports - broadcast information to the whole community about what works and what doesn't. Of course, this is a very different type of exchange from those involved in financial markets; it involves intellectual engagement, not money. Financial exchanges produce price discovery. Frictionless Research Exchanges produce *community critical review*.

Within an FRX, each researcher expects to find out what others have done, how those results worked, and then can build directly and explicitly on the selected work of others; seeking tweaks and improvements over their work. A community of researchers working this way initiates a chain reaction of reproductions followed by furious experimental tweaking, with each new ascent up the leaderboard exciting responses by others and further tweaks and improvements.

Empirical Machine Learning (EML) is, par excellence, the field which has most nearly adopted the FR triad in its research model; which has most completely shaped itself to have the habits and attitudes that create an FRX; and has benefited the most from FR practices.

Many computational experiments in EML follow a very uniform structure; understood by everyone in the field, the sites and tools (see eg. CodaLab on the noncommercial side; Weights and Biases, HuggingFace, Github on the commercial side) offer easy immediate uniform access to archives of canonical data, models, performance metrics, and all the other paraphernalia needed to begin reproducing results as they are announced, or developing modified experiments which build on those announcements.

EML research has delivered the numerous stories of remarkable 'AI progress' I noted earlier: in machine vision, natural language processing, gameplay, and even protein structure prediction. In my opinion, without the last decade's (near-) Frictionless Reproducibility, this progress could never

---

[21]Many fields are just sleepy and haven't had their wake-up call; some feel constrained by code or data ownership (but may not know about the proven ways to mitigate those concerns).

have been delivered on the decadal time scale which we have all just witnessed. The rate of progress on each problem would have been much smaller, because the number of meaningful contributors would be much smaller, and uncertainty about the meaning of individual experimental reports would have exerted a significant drag on the energy and enthusiasm of the research community. But also, without the right habits and engagement by researchers in the field, the full FRX experience would not have emerged.

EML, by its embrace of frictionless exchange, has made a very valuable discovery of new habits and mindsets in research, which can benefit us all, and has revealed the spontaneous emergence of Frictionless Research Exchanges, a **research superpower** at the heart of the EML story[22].

The triad of FR practices is not "the same thing as what EML is doing"; The EML community seems not inherently committed to the whole FR Triad; some hegemons but not others play nice with these practices [23].

Protein structure prediction began using the challenge problem paradigm with the CASP challenges 3 decades ago; spent a great deal of research effort over the years in developing insights into *identifying scientifically relevant performance measures.* and benefited as we all have from the increasing tendency towards data sharing and code sharing.

Other fields might follow these practices and benefit correspondingly. Scientific fields recently implementing challenges include NMR Structure Determination [Rosato et al., 2009], and Computational Biology [Blasco et al., 2019]. Recent large-scale funding for new challenges include an initiative in High Energy Physics [24] and in Applied Mathematics (Dynamical Systems)[25].

For three more extended examples, see Section 6.

5. A Leap Beyond

In my view, FR and FRX mark a leap well beyond their three well-known predecessors.

5.1. **Beyond In-Principle Reproducbility.** Traditionally, computational scientists sometimes offered In-Principle Reproducibility (IPR) in their publications, sharing only some code and data publicly, with additional details embedded in the paper, so that after careful reading and collating of information from various original sources, a researcher could eventually roughly reproduce another's work. While IPR is hardly frictionless, it stood for some time as a worthy standard to emulate, deserving of approval where achieved.

In the experience of researchers who lived under both regimes, IPR contrasts dramatically with FR. The common experience under IPR was that during the effort to reproduce, there would be: (a) discovery of many undocumented steps that might offer 'stoppers'; and (b) many places for previously undisclosed human input and decision where there could be misunderstanding or miscommunication. Consequently, the prospect of success in reproducing reported results was forever in doubt, in turn spawning near-universal skepticism on the part of experienced researchers

---

[22]Yet, an unsung achievement. Paradoxically, an unsung superpower.

[23]At the time of writing, Meta AI has been releasing many open AI tools and artifacts (not always full open source code but sometimes open source weights)) Meta's AI research group wants to keep releasing models for free despite criticism that Llama 2 isn't open enough, while some other hegemons are not currently releasing even those artifacts.

[24]"DOE awards $6.4 Million for AI benchmarks for HEP"

[25]https://dynamicsai.org/, University of Washington: "... new NSF institute ...to understand dynamic systems."

of results reported by others (the well-known "Not Invented Here" attitude). FR, by removing such undisclosed and undocumented components, makes it much, much easier for researchers to verify and therefore build trust in another group's work.

5.2. **Beyond Open-Source Software.** Open-Source Software (OSS) has been a major presence in the world of information technology for 30 years[26]. In its best implementation, which has been widely practiced for decades, one indeed obtains OSS through frictionless services. OSS is thus an enabler of FR, an important step in the right direction. It enables software developers to build their work on the coding of others; this is a crucial part of the usual path to FR, but *only one part*. Re-executability of a published computational experiment requires **much more** than open access to some key software components that were used in that experiment; generally requires the exact reconstruction of an entire computational environment, and an entire orchestrated workflow. An important corollary of the cloud information technology era was ubiquitous virtualization and containerization[27], enabling specific machine instances, and even specific virtual cluster architectures. These technologies matured long after the OSS movement.

In my mind, the Open-Source movement takes us beyond In-Principle Reproducibility but not yet to full Frictionless Reproducibility, and also not to the emergence of a Frictionless Research Exchange. For example, a researcher absorbed in solitary pursuits can share code and data, it can even be available frictionlessly, but without engagement by a community of other researchers that cares deeply about assimilating the best contributions, building upon them, and outdistancing them, we will not see emergence of rapid progress.

5.3. **Beyond Competitive Challenges.** Challenges alone, without *each* of [FR-1: Data Sharing] and [FR-2: Code Sharing] *and consequent research community engagement*, also can't produce the advantages we are discussing. The typical missing ingredient is [FR-2], reverting us to what we called the Proprietary-Methods Challenge.[28] In such circumstances, we miss the sharing and cross-fertilization of knowledge, we don't as a community know what has been done to get a method which bests the competitors, and we don't get the chance for *critical assessment* of the methods. It is possible that a 'method' is just a set of random tweaks with no rhyme or reason, yet can win a particular competition at a particular moment essentially due to random noise – some competitions are just that close. Competitions don't, inevitably, produce lasting results or knowledge. *Community review is what ensures that important advances are recognized and propagate.*

Certain communities already take 'Open Source' as a given, meaning data sharing and code sharing. In such situations, [FR-3: Challenges] seem to be the only missing ingredient. It is then natural to think 'Oh, it's all challenges'. Fair enough. But if that psychology takes hold, by a 'slippery slope effect', we may forget to secure the assumed givens (Data, Code, Community) and fail to produce critical assessments.

---

[26]https://en.wikipedia.org/wiki/Open-source_software

[27]https://en.wikipedia.org/wiki/Virtualization

[28]Current LLM development involves *frontier LLM models* developed at hegemons suxh as OpenAI and Google; these companies may share descriptive meta-information about the data they use (or they may not) and they may share benchmark performance numbers. There is an implicit competition on these benchmarks which resembles a (virtual) Proprietary Methods Challenge. In contrast, the developers of the so-called *open-source LLMs* (mostly derived from LLAMA), are typically playing according the full FR stack [FR1]-[FR3]. These developers also share the community aspect, so FRX as well.

6. FR Everywhere!

Crucially, other research fields can benefit from adoption of FR practices and institutions just as well as EML – or maybe more. The FR triad can be, and is being, practiced outside the empirical machine learning world.

To see this, we do pattern recognition, noting analogies where certain elements in the triad are, although different in detail, *structurally* identical.

For example, the notion of challenge could change. Here the data and code ingredients [FR-1]+[FR-2] might be as before, but the Challenge task being solved per [FR-3] might no longer be what EML calls a prediction task. Mark Liberman, a Natural Language Processing expert at University of Pennsylvania, has been using **Common Task Framework** to label challenges [FR-3] more neutrally, without implying that they are prediction challenges[29].

Consider the field of Optimization. There are many algorithms for solving well-posed optimization task instances such as linear programming and quadratic programming, and traditionally one developed personal preferences for one algorithm or another based on intellectual perspectives and mathematical properties. What matters is the speed of an algorithm and the accuracy of the solution. The data science approach in this context would: [FR-1] share standard datasets which will be used in defining specific optimization problems, [FR-2] share code defining reference implementations, and [FR-3] define metrics (running time, solution accuracy) and maintain leaderboards. Benchopt[30] does exactly this, focusing particularly on the algorithms of interest in machine learning; but instead of asking for the prediction performance, as in much EML research, it asks for timing or solution accuracy. French research funding agencies supported Benchopt with personnel funding and compute access.

Consider an example from Statistical Methodology. Stephen Ruberg and co-authors [Ruberg et al., 2022] conducted a challenge to promote development of 'subgroup identification methodologies' ultimately intended for the analysis of data in pharmaceutical clinical trials. Here *synthetic data* are used, with ground truth known. The authors construct many specified generative models, whose details are (let's say) sequestered from view, then sample data from those models. Task performance metrics included the ability to correctly identify the subgroups of treatment responders, because the underlying ground truth synthetic model was known to the organizers; such metrics couldn't be defined in the usual EML framework. The challenge itself was hosted on the InnoCentive platform, and offered a prize of $30,000 for first place and $15,000 for second place. More than 100 competitors vied to predict properties of the generative model. This work and its publication were supported by Eli Lilly.

In physical chemistry, Adam Schuyler and co-authors [**NUScon**] conducted the NUScon challenge, seeking reconstruction methods for non-uniformly sampled NMR Spectroscopy. NUScon was hosted using the NMRbox methods platform [**maciejewski2017NMRbox**], which allows the creation of workflows using a large number of different software tools developed by the research community over the years, as well as extensive 'workflow glue' that allows chaining together such methods and tweaking their operation. Synthetic data were created using NMRbox based on underlying physical models. Competitors used NMRbox to code their entry workflows, and administrators used NMRbox and the NMRbox compute cluster to implement the contest evaluation. The platform provides an (almost) fully-automated workflow for running contestant reconstruction recipes

---

[29]https://www.simonsfoundation.org/event/reproducible-research-and-the-common-task-method/

[30]https://benchopt.github.io

on contest data and evaluating solutions. Several thousand dollars in cash prizes were awarded in a NUScon prize ceremony at the annual meeting of the Experimental Nuclear Magnetic Resonance conference (ENC). The contest produced community engagement with, and adoption of, methods for nonuniform sampling, which allow very substantial speedups in data acquisition for high-dimensional NMR experiments. It documented in a fair and objective way the measured performance of a variety of methods on a common platform. In addition, NUScon improved awareness of, and adoption of, the NMRbox platform, and thereby created a more capable and productive community of researchers.

In each of these three examples, the problems being attacked are not one-one mappings of standard ML problems; but some variant of common task framework can be defined and implemented.

7. The Roots in Data Science

The maturation of [FR-1]+[FR-2]+[FR-3] and emergence of FRX did not spring out of a vaccum. Nor out of slidedecks presented to Silicon Valley VC's who funded Github, Kaggle, and Hugging Face. Nor out of today's hegemon research labs. Rather, they developed organically from efforts by data scientists and technologists across at least 4 decades, witnessed by my own eyes.

7.1. **Reproducible Computational Science.** RCS has been developing steadily for decades[31]. A few examples: [Peng and Hicks, 2021]

- Geophysicist Jon Claerbout began advocating and practicing[32] Reproducible Computational Research in the early 1990's, changing the standards in his field and changing publication practices to enable provision of digital research artifacts.
- Over the last 20 years, Computational Biologists have been increasingly sharing code and data, including genome sequences[33] [Benson et al., 2013], gene expression data[34] [Ringwald et al., 2000], phenotypic data[35], and algorithm libraries like BioConductor [36], and workflow frameworks [Grüning et al., 2018].
- In the last 10 or 15 years, numerous professional scientific journals and societies have acknowledged the need for, and developed accommodations for, reproducibility of computations [Peng, 2009]. Cross-science surveys of this activity are available from reports of the National Academies of Science, Engineering and Medicine; see for example the recent 2019 consensus report *Reproducibility and Replicability in Science.*[37]

The intent has consistently been to share data and code using the technological capabilities and scientific awareness of the day. As those capabilities advanced, the comprehensiveness of the reproducibility and immediacy of the access also advanced. In my view, today's RCS capabilities,

---

[31]I offer myself as a witness who has seen and participated in data-driven methodological development first-hand for almost 50 years, including through multiple stints in industry.

[32]We find the following at https://sepwww.stanford.edu/sep/jon/reproducible.html: "In 1990 I set ... a goal of reproducible documents and results. The basic idea we had is that anyone should be able to reproduce our research results with the software on our CD-ROM ... Many people thought we were reaching too high. I recall some thought we were braggarts, liars, or crazy. Reality showed us it wasn't as easy as it seemed."

[33]https://www.ncbi.nlm.nih.gov/genbank/

[34]https://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml

[35]https://www.ncbi.nlm.nih.gov/gap/

[36]https://www.bioconductor.org/

[37]NASEM: "Reproducibility and Replicability in Science"

though they would of course not been predicted in detail by scientists of two or three decades ago, would have been clearly anticipated at a high level. A hypothetical scientist of those times might even have expected them to come together long before 2023[38].

Actually, though, many many scientists and technologists had to march in the direction of RCS in order to create the culture, customs, and tools we have today. Moreover, RCS still doesn't happen in every case. Incredibly to me, advocacy of RCS is spotty in certain fields.

Advocacy of RCS, following RCS practices, and implementation of RCS therefore deserve our praise. I won't enumerate all those I consider instrumental. The already-mentioned report *Reproducibility and Replicability in Science* [National Academies of Sciences Engineering and Medicine, 2019] presents an extensive cross-section sampled across many scientific fields.

Still, for this article, there's one person I'd like to spotlight as a standout advocate and implementer.

> **Yann LeCun:** Turing Prize winner; Chief AI Scientist of Meta; Silver Professor at NYU; and important thought leader in today's AI information space. Among technical leaders in the trillion dollar market cap group, LeCun, to me, stands alone as a highly vocal and consistent advocate of code and data sharing. Over the last decade, Meta produced, shared, and supported fundamental tools like PyTorch[39] and Llama [40], which are integral to reproducible computational science in many fields today.

As the last touches were being placed on this manuscript, LeCun spoke to the US Senate Committee on Intelligence, which wanted to hear about AI and preserving American Innovation [41]. As one can imagine, many politicians and corporate leaders would be talking in such a setting about placing technology under lock-and-key. Instead (my observation in italics):

> *LeCun chose to use his precious moments of testimony to emphasize the importance of Code and Data Sharing. He explicitly mentioned these factors as* **crucial to the recent rapid advances in AI.**

In this paper's language, LeCun is implying that our secret sauce is the frictionlessness of building on the work of others. This is our economic and intellectual engine.

I quibble only in whether RCS alone – [FR-1]+[FR-2] – is enough to get all the rapid advances. Against my inclinations, I came to believe one extra factor is essential - [FR-3: Challenges] – to produce full FR, and after that, we also need the engagement of a whole community of researchers with a new type of institution – the FRX – for emergence of all the rapid progress we are seeing.

7.2. **The Challenge Paradigm.** From my perspective, the *I-didn't-already-see-this-coming-40-years-ago* ingredient in the FR triad has been the impact of [FR-3: Challenges]. For at least 50 years it was widely understood that collecting data, and sharing it, could drive data analysis challenges.[42] However, the view of data analysis 'challenge' in early years would have been extremely broad and

---

[38]I published in 1995 an article advocating widespread adoption of RCS [Buckheit and Donoho, 1995] and in 2009 one reviewing 15 years of experience [D. L. Donoho et al., 2009]

[39]https://pytorch.org/   Ars Technica: "Meta Spins off Pytorch"

[40]Wired: "Open-Source Llama upsets AI Horse Race"

[41]Senate Hearing, Sept. 19 2023

[42]A dataset I manually gathered drove such a challenge problem at the Joint Statistical Meetings in 1983; I have seen challenges with my own eyes for 40 years.

I would say 'humanistic', and could tolerate a wide range of challenge deliverables, include the challenge of 'understanding' what's in the data, or the challenge of 'revealing an interesting view of the dataset', or even the challenge of 'redefining our ideas of what can be done with the dataset'. A popular term at the time was 'exploratory data analysis', which well evokes the possibility of fun and discovery that inspired many. From this vantage point, what came next was a real surprise.

A very different *engineering* view of challenges became formalized[43] in the mid-1980's for research in natural language processing and biometrics. Researchers were asked to submit entries scored by specified, predefined metrics. Leaderboards were instituted and winners declared algorithmically. This approach to challenges was radical in its clarity and simplicity, but also in its narrowness of focus.[44]

In a series of projects, many funded ultimately by the Pentagon (DARPA) in the 1980's through the 1990's and beyond, challenges were mounted in speech processing, biometric recognition, facial recognition, and other fields. Funding, often to NIST [45], covered data collection and curation, and challenge contest administration. Sometimes, winners of challenges were selected for follow-on DARPA grants and research contracts.

In a major departure from the principles advocated in this article, some DARPA challenges followed the Proprietary Challenge model, which allowed contestants to keep their code and working methods confidential, essentially meaning that contestants learn leaderboard results, but not necessarily about the winning models or fitting procedures. Such an implementation of [FR-3: Challenges] is incompatible with [FR-2: Code Sharing], and hence with frictionless reproducibility, showing it makes a difference when all three of [FR-1]+[FR-2]+[FR-3] are specifically present. This longstanding tension is important at the present moment in empirical machine learning, where some internet hegemons take a proprietary line.

Other data science fields instituted challenges organically without Pentagon instigation. The CASP protein structure prediction problem is the longest-lasting scientific recurring challenge problem venue[46]. The sequence of a certain protein is known; predict its 3D structure. This competition has been around roughly as long as the world-wide-web and web browsers; its website today at https://predictioncenter.org/ documents the (peri-) biennial competitions since 1994 and the evolution of predictive solutions. Many outside-of-proteins data scientists' first inkling of this competition followed a recent episode of hegemon engagement; DeepMind entered into recent CASP competitions, deployed overwhelming computational resources, got their wins, and issued their corporate PR[47], to the applause of the capital markets[48]. But the competition has been ongoing for 3

---

[43]Ben Recht and Moritz Hardt [Hardt and Recht, 2022] have pointed out that Bill Highleyman, already in the late 1950's, had gathered data and shared it with others and that some of those he shared with were implicitly competing in a challenge, even though the notion hadn't yet been formalized: http://www.argmin.net/2021/10/20/highleyman/

[44]I recommend Mark Liberman's presentation Reproducible Research and the Common Task Method

[45]https://www.nist.gov/biometrics

[46]A referee points out that NIST/DARPA speech recognition challenges began already in the 1980's: Early NIST Speech Recognition challenges; CASP is going strong even today, while speech to text challenge goals and definitions have mutated and the venues have changed.

[47]Including expensive documentaries using Oscar winning talent.

[48]Venture Beat: "AlphaFold wins CASP"

decades, adapting and developing in the direction of perfecting the challenge model and the community that engages in it and gradually improving protein folding; without a corporate PR budget then or now.

Many data scientists ought to be mentioned as heros of the challenge paradigm, across science and technology, and across decades. I list 3 important figures, in last-name-alphabetical order.

- **Isabelle Guyon** of Université de Paris–Saclay [49], who collaborated on the original UNIPEN dataset [Guyon et al., 1994] and co-authored a foundational paper on computer handwritten digit recognition [Bottou et al., 1994]. Over the last 20 years she has been heavily invested in many prediction challenges [Guyon et al., 2006, 2008, 2011, 2015; Jiang et al., 2020; Z. Liu et al., 2021]. A recent example is the 2 Million Euro Horizon prize for big data technologies[50], organized by her and colleagues at Chalearn http://www.chalearn.org.
- **John Moult** of the University of Maryland, a molecular biophysicist who has been a sparkplug of the CASP protein structure prediction contests since the very beginning, i.e. the early 1990's [Kryshtafovych et al., 2010, 2014, 2021; Moult, 2005, 2006; Moult et al., 1995, 2018].
- **Jonathon Phillips** of NIST, who played a crucial role on many of the pioneering '90s and 00's NIST/DARPA challenges in vision and biometrics, including data collection and contest organization [**Phillips1998feret**; Phillips et al., 2000, 2002, 2005, 2009], with challenge involvement continuing up to the present time.

Although they have very different professional backgrounds, they have in common an exceptional clarity on the importance of challenges and exceptional persistence in promoting this paradigm in their domains.

From a later generation, I will also mention:

- **Percy Liang** of Stanford; he had the farsighted idea in the 2000's to create a sort of 'operating system for conducting challenge problems' that is today's CodaLab worksheets [51]. An offshoot of this, managed by Isabelle Guyon and collegues is today's CodaLab competitions [Pavao et al., 2023] and Codabench systems [Xu et al., 2022], which administer a large number of academic challenges annually [52].

These data scientists, and many others I don't have space to mention, have shown how to make data science a **science, full stop**. Their vision is that data science research projects making full use of [FR-1]+[FR-2]+[FR-3] are truly empirical science; while research projects without all three will fall short in some way. In fact, Isabelle Guyon, in her 2022 keynote address at NeurIPS [53], has chosen to emphasize exactly this: through challenges, Machine Learning has become an empirical science.[54]

---

[49]https://guyon.chalearn.org/

[50]Horizon Prize

[51]https://worksheets.codalab.org/

[52]https://codalab.lisn.upsaclay.fr/

[53]Isabelle Guyon: NeurIPS Keynote

[54]Many mathematical scientists ask me: what is *conceptually* new in EML, really outdistancing the various contributions that mathematical scientists put forward long ago? In my view it is easiest to explain this as the consequences of a community embracing the FR triad and reaping the consequences.

So the social practice, of challenges propelling research, wasn't born yesterday; it has been developing and refined across four decades. Pioneers saw early on, and soon made others understand, that this would be a transformative development.

7.3. **Surprising Reactions.** In a sign of its potency, [FR-3: Challenges] elicits a range of surprising reactions.

7.3.1. *Reaction 1: Wild Enthusiasm.* Mark Liberman has made clear that already in some of the earliest contests, the challenge paradigm tapped into a very special, previously unexpressed human energy. Leaderboards game-ified research and exerted a deep grip on the moment-by-moment attention of many participants of early challenges. They eagerly dropped into a tight loop of tweak model and data, submit new answers, try again. It was understood that performance improvements were happening at a much faster pace, even in the humble early days of this paradigm. It was also clear from this early experience that, if data and code sharing grew, and leaderboards went global, this human energy would increase in staggering ways. Exactly as happened in EML over the last decade.

7.3.2. *Reaction 2: Shock and Disbelief.* Quite a few mathematical scientists (including myself) have found it initially difficult to accept that a great deal of work in the empirical machine learning community was not supported by careful formal analysis and derivations e.g. theorems from mathematics or from CS-theory. Many math scientists (also including me, intially) had difficulties understanding that a whole scientific community could be based entirely on the foundations of reproducible computations and challenge leaderboards.[55]

In processing my own sense of dislocation, I turned to the famous article of Alon Halevy, Peter Norvig, and Fernando Pereira *The Unreasonable Effectiveness of Data* (2009) [Halevy et al., 2009]. This squarely took aim at Eugene Wigner's *The Unreasonable Effectiveness of Mathematics in the Natural Sciences* (1960) [Wigner, 1960], an article of some veneration among those trained in the mathematical sciences over the last 60 years. Halevy et al.'s article is best viewed as a provocation rather than an attempt to convince doubters, but it does brutally explain the new rules of the game.[56]

8. ACCELERATION TOWARDS SINGULARITY

The argument so far is an argument that [FR-1]+[FR-2]+[FR-3] have recently come together, after lengthy gestation in plain sight, and this combination - in the presence of the right research community practices - unleashes an acceleration of computational-methodology research. But acceleration is everywhere, not just in research; for inhabitants swimming through the 2020's decade, it surrounds us like water.

---

[55]i.e. literally an epistemology based on RCS=[FR-1]+[FR-2] as well as Challenges [FR-3]. See Section 9.1 below. Many also initially miss the emergent superpower, the community behavior that sets in as FR becomes FRX.

[56]Matan Gavish and I are preparing a lengthy discussion about the contrast between Digital and Mathematical worldviews that aims to be more instructive, and less merely provocative, than Halevy et al..

8.1. **Acceleration Everywhere.** Innovations of all sorts are today adopted more easily and spread more rapidly than ever. The spread of memes on the internet was recognized as a new phenomenon already in the 1990's; the sudden virality of slogans, posters, or videos that began to be seen at that time is today a dependable feature of the information space; each day on social media, the 'trending now' feature presents that day's viral topics, likely to be replaced tomorrow by other viral topics, often unpredictably.

At the heart of viral meme spread is the 'single-click transmission' of ideas. Some saying or image is appealing, and the ease of transmitting that message to others is so effortless, just a few clicks; the viewer overcomes inertia, and forwards it to others; initiating or continuing a chain. This is then repeated, in some cases almost endlessly.

Some observers see a biological drive at work; it is common to say that each participant in the chain gets a 'dopamine hit' from her/his 'discovery' of the content in their in-box, followed by a fulfilling 'action' of reposting or forwarding[57]. Some participants speak darkly of their 'addiction' to participating in such chains as caused by 'dopamine craving' [Lustig, 2018]. Others praise, and seek out, the 'flow state' induced by participating in the single-click transmission of ideas.

We are now so used to frictionless spread of sayings and images, some quite arresting, entertaining, and shocking, that many of us expect a daily parade of amazement.

Mass culture has been practically overwhelmed by this development. Billions of people have smart phones, which expose them to social media for hours a day[58], during which they experience the cresting and breaking of waves of viral memes.

This has upset traditional activities like politics [59]. Popular 'grass roots' movements can grow virally and crest beyond the reach of any traditional debate or clash of ideas. The attention-grabbing capabilities of such viral waves are breathtaking. Originally "fringe" ideas can suddenly emerge, seemingly from nowhere, to capture significant mindshare.

The recent unprecedented dynamism of mass culture is happening all around us, all the time; but we may forget that it is rooted ultimately in the technological possibility for delivering (almost) frictionless spread of sayings and images. Today's frictionless regime is invisible to us fish; it is water.

8.2. **Friction spoils everything.** In contrast, friction – drag on the effortless sharing of information – is really, really visible to us. It breaks the flow, stops the dopamine. Today, many have no patience for friction, of any kind.

One can see complaints about this on social media: 'receipts or it didn't happen', 'URL or I don't care'. Commenters want the experience of accessing digital content with a single click. If there is no digital content to access efficiently, participants are frustrated and share their frustration, vocally; perhaps even with digital tokens of their frustration, such as emojis or GIFs.

Putting content behind 'paywalls' has the same effect. Some participants will react strongly against it. Some may not want to pay, but many simply don't want friction!

8.3. **Virality in Data Science.** Frictionless reproducibility of scientific computations is the analog, for research, of the frictionless transmission of internet memes.

---

[57]Harvard Science in the News: "Dopamine and Smartphones"

[58]https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

[59]Brookings Institution: "Tech and Political Polarization"

FR aligns powerfully and naturally with the habits and practices each of us experiences in our experience of digital culture at large. A new methodological tool published in a way that subscribes to [FR-1]+[FR-2]+[FR-3] may spread across a field like wildfire.

I have seen this up close in single-cell RNA-Seq data analysis. A paper [Gribov et al., 2010] in a leading journal used a new software package, Seurat, and the data and code were packaged in a very clean way[60], with a user interface that made it extremely easy to reproduce their results – but also to apply the same methodology to fresh data from another project. The package spread across the RNASeq research landscape rapidly and it soon became de rigueur for visualizing RNASeq data in publications.

In a series of interactions with the computational biologist I was working with, the narrative quickly went from 'what is this new tool? Can I trust it?' in an initial meeting to, a few meetings later, 'I was just at a conference, Seurat is everywhere; you now have no choice, you must now present your data using Seurat, or no one will take you seriously'.

After diving into Seurat's technical ideas, I saw that a key driver of such viral spread must be simply the ease of adoption of the package and the attractiveness of the produced plots, almost divorced from other considerations. The package was very well designed for 'unboxing' by newcomers, effortlessly delivered colorful presentations of some supplied datasets, and following the rules for creating general purpose software for use with other datasets. In my opinion, this ease of adapting the tool to other datasets definitely propelled Seurat's adoption.

By the way, Seurat was not the lone option; many competing methodological tools were available. I studied some of them and came to the conclusion that Seurat offered the least 'reproducibility friction', and so it spread.

8.4. **Onset of Frictionless Reproducibility as a Singularity.** Today's onset of Frictionless Reproducibility will have major consequences for methodological research. To see this, consider lessons from *The Singularity is Near*, [Kurzweil, 2005] a prominent influence in discussions about modern AI.

In 2005, computer technologist Ray Kurzweil famously forecast that "The Singularity" would take place around 2030, after a factor 10^{14} more compute would become available to humanity, compared to its 2005 level. This forecast was derived by stacking many fascinating speculations, at the top end approaching the spiritual and religious. Kurzweil's audacity firmly planted the idea of "The Singularity" in the minds of intellectuals, after which Computer Scientist Eliezer Yudkowsky[61] and other adherents of the "Less Wrong" online community[62] focused attention on the problem of "AI Alignment"[63] – the worry that AI will one day surpass us, and then, afterwards, be cruel to us. The most vocal worriers are "AI Doomers"; they fear the singularity will arrive with a "hard takeoff" that will overwhelm us with destructive consequences before we even understand what is happening.

Kurzweil quotes one of the 20th century's most prominent mathematicians, John von Neumann:

> The history of technology ... gives the appearance of approaching some essential singularity in the history of the race, beyond which, human affairs, as we know them, cannot continue.
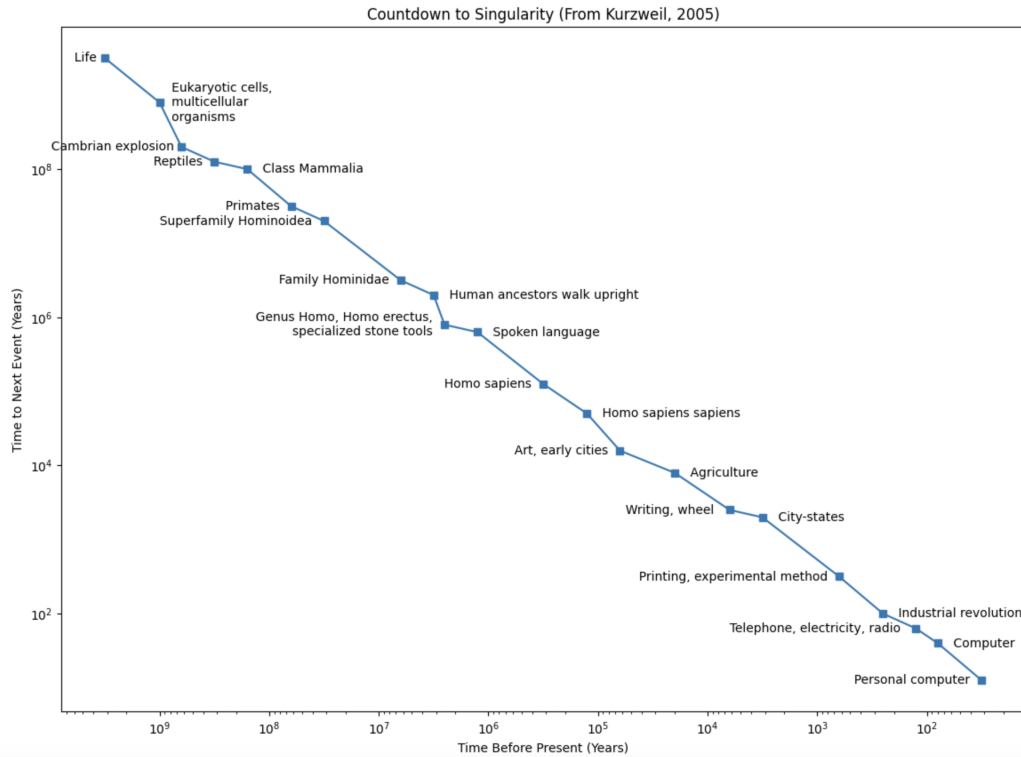
---

[60]http://seurat.r-forge.r-project.org/

[61]https://en.wikipedia.org/wiki/Eliezer_Yudkowsky

[62]https://en.wikipedia.org/wiki/LessWrong

[63]https://en.wikipedia.org/wiki/AI_alignment

Von Neumann introduces the idea that a singularity is coming.[64] But when? Kurzweil presents Figure 1.

Figure 1. Ray Kurzweil's 'Countdown to Singularity' (Kurzweil, 2005)

Here Kurzweil's vertical axis presents "time between notable events" and the horizontal axis shows time before the present day. The interval between notable events is dropping towards zero, signaling acceleration in the pace of human affairs.[65]

Combining the last two exhibits, Kurzweil's rhetoric suggests that:

when the time between notable events **drops to zero**, *we are at* **singularity**.

---

[64]Von Neumann died in 1957. He was presumably at the time of this quote gloomy from his own impending early demise and from the Soviet Union's apparent willingness to use nuclear weapons including H-bombs to advance its imperial ambitions. Von Neumann makes two unnecessary leaps: (a) the terminology 'essential singularity' - borrowed from complex analysis - overdramatizes, (b) the concept of 'could not continue' - also borrowed from complex analysis - again goes too far. Von Neumann's fears have not yet materialized; the global diplomatic order has been able to inject *friction* into the development of nuclear weapons, including outright prohibitions on testing, and sanctions on nuclear violator countries, such as North Korea. A more objective and less dramatic statement would have been: "The history of technology gives the appearance of approaching ... *a new era of unprecedented rapid change*." But of course if von Neumann's phrasing had been less dramatic we would not be discussing it today.

[65]Dr. Michael Feldman points to the arbitrary character of the 'notable events' in Kurzweil's list, suggesting that their selection might be biased to prove the rhetorical point being made.

The 'reaching zero moment' is quite explicit in both Von Neumann's and Kurzweil's account; it involves observables we can directly sense. I adopt this as my *criterion for 'dating a singularity'*.[66],[67],[68]

Compare to Frictionless Reproducibility in computational science. FR heralds a drop to zero of the human effort required to reproduce a computational result. This micro-phenomenon (time for a researcher to reproduce one result dropping, essentially to zero) in turn drives a macro-phenomenon – the time for a field to globally adopt a new dominant methodology – also dropping, essentially to zero. As we saw in the rapid embrace of Seurat for Single Cell RNA data analysis.

I see analogies to the physicist's idea of superconductivity, where an electrical conductor's resistance drops to zero. Many have explained that easy access to superconductivity could one day have amazing consequences for the energy industry and the world economy[69]. I likewise sense that frictionless reproducibility can have, and already is having, impressive consequences.

Everywhere we look, progress in science and technology is speeding up. At the same time, we see a transition to frictionless exchange of digital research artifacts. At the macro-level, speedup; at the micro-level, frictionlessness.

The mRNA vaccine story of 2020 is a well-known emblem of science acceleration. Early in the COVID-19 pandemic we were told that novel vaccine development has never taken place in less than a decade, and yet in about a year hundreds of millions of people were already vaccinated, in the US mostly with mRNA vaccines [Bourla, 2022]. A key part of the story, often left out, is the frictionless spread of information about the SARS-COV-2 virus, most importantly its sequence. The virus' RNA sequence data were published by Chinese scientists on virology websites in mid-January 2020 and within days virologists all over the world were analyzing the sequence and very soon had vaccine candidates[70]. The new mRNA technology fit perfectly with this situation as it allowed virologists to design a vaccine candidate directly from digital sequence – a comparatively frictionless process. The friction in the end-to-end process was all in the traditional political, regulatory, manufacturing, and public outreach realms. Companies like BioNtech Pfizer and Moderna are justly proud of the

---

[66]This notion of singularity differs from another one of Kurzweil's, mentioned earlier. That other one points to the moment when the amount of non-biological computation first exceeds biological computation; but this seems not directly observable, while "time between major innovations reaches zero" seems something we can all sense, and hence more empirically valid. In Kurzweil's AI context the difference between the two criteria is perhaps a decade.

[67]Kurzweil is completely correct to call our attention to the trend towards a decreasing time interval between key innovations, and to attach importance to the fact that this time interval is currently approaching a vanishing point. However, after following along on Kurzweil's voyage this far, I jump off his train, and so should you.

[68]Also important to my thinking: singularity can have a neutral interpretation in mathematics. It doesn't necessarily involve the world blowing up or ending abruptly. In Fourier analysis, a function $f_\alpha(x)$ of a real variable such as $\text{sgn}(x-s) \cdot |x-s|^\alpha$ for some fixed α can be said to have a singularity at point s. This includes $\alpha \geq 0$ which does not 'blow up' as $x \mapsto s$. Singularities don't have to diverge. There just has to be a noticeable break in behavior, before versus after. (Of course, in complex analysis the term singularity does generally involve 'something blowing up'.)

Perhaps the reader thinks that 'inflection' makes more sense than 'singularity'. And yet. Is the recent rapid change in machine learning performance, across many tasks and data modalities, a mere inflection? It seems we saw a *rupture* in performance between the adjacent decades of the 2000's and 2010's.

[69]For example: Article: "Room-Temperature Superconductors could Revolutionize Electronics"

[70]Cell Host Microbe: "Genome Composition of Novel Coronavirus Originating in China"

amazing operational job they did in immunizing billions of people in 2020-2021; this practical effort was unprecedented. The scientific and technological translation was, comparatively speaking, instantaneous[71][72][73].

Key enablers of this rapidity were [FR-1] data sharing (of the virus' RNA sequence) and widespread [FR-2] code sharing (of algorithms that could analyze and translate that sequence in various ways).

We earlier discussed the staggering rate of recent progress in large language models. The sudden emergence of GPT-4 and chatGPT in last year's public discourse spotlights one company's work; but it is better viewed as a evidence that the NLP field as a whole has made very dramatic progress across many tasks and challenges in the last decade. Consistent with our theme, the habits and institutions of [FR-1]+[FR-2]+[FR-3] are, as we expect, very strong in Natural Language Processing (NLP). The availability of massive language datasets ([FR-1]), the sharing of architectures and fitted models ([FR-2]), and the popularity of challenges ([FR-3]) are heavily present in this field; their synergy and associated institutions, including CS conferences[74] and Open Review of submissions[75] shaped the community of NLP researchers into a Frictionless Research Exchange.

Again, frictionlessness at the micro level, rapid progress at the macro.

Combining the above comments, and applying the criterion given above, I propose we are at a singularity - the **reproducibility singularity**. I don't propose dramatic claims of mass unemployment and human extinction. I do propose instead fundamental changes affecting computation-driven research and its rate of progress.

9. Revolutions at the Singularity

The crossover to FR upends our expectations about research and researcher behaviors, in two ways.

9.1. **Computational Epistemology.** In the 'comprehensive FR' regime, where every computation can be reproduced, adherent researchers no longer depend on hypotheticals. They instead limit themselves to documenting facts:

What:: a specific workflow does;

On:: a specific publicly available dataset;

According to:: a specific task performance measurement;

Using:: publicly available code.

This is an extremely limited agenda for research discussion! In exchange for the limitations, one gets two benefits:

---

[71]Nature: "Lightning Fast Covid Vaccines and what it means for other Diseases"

[72]Of course, this progress relied on lots of prior scientific and technological developments, for example rapid sequencing and sequence analysis and manipulation. But that had all been done and already made frictionlessly available – in time to be deployed for a global pandemic.

[73]Admittedly, it took some months to collect the clinical trials data, and analyze it. However, the existence of rapid translation of data to vaccine candidates motivated subsequent efforts at 'clinical trials process optimization' leading to much faster progress in conducting clinical trials: Accelerating Vaccine Trials. FR is influential in the sense that it changes attitudes about things outside the data science field.

[74]https://2023.aclweb.org/

[75]https://openreview.net/   https://aclrollingreview.org/

Epistemic modesty.: The research literature is transparent about what is being asserted and under what conditions.

Full computational reproducibility. : The research literature becomes a transcript of actual code executions, and so is, simply, true.

This approach to epistemology is completely a creature of the Frictionless Reproducibility triad and would make no sense without it. One is only making claims that are immediately verifiable because of the assumed [FR-1]+[FR-2]+[FR-3] setting. This intertwining is spelled out in the table below:

Facet Documented	Facet explanation	Triad Elements Involved
What:	Specific Workflow Does	[FR-1]+[FR-2]+[FR-3]
On:	Specific, Publicly Available, Dataset	[FR-1: Data Sharing]
According to:	Specific Task Performance metric	[FR-3: Challenges]
Using:	Specific Publicly Available Code	[FR-2: Code Sharing]

Table 2. Epistemology and FR Components

What is obviously gained is efficiency in article composition, and efficiency in article evaluation. Less obvious is that *this epistemology enables emergence of an FRX*, which is a research community exchanging digital research artifacts rather than more nebulous entities.

What is potentially gained compared to traditional scientific publication is unprecedented virality. In EML, some articles showing a fertile new architectural concept breaking some performance barrier can garner massive impact quickly (eg. ResNet, Transformers), by being rapidly incorporated in thousands of works of other community members.

Look at the ranking of top journals sciencewide, ranked by accepted citation metrics. Over the last decade, coincident with the maturation of the data science components [FR-1]+[FR-2]+[FR-3], the journal impact leaderboard, long heavy with incumbent biology, medicine, and general science journals, has been disrupted; 3 of the top 10 journals are today from engineering fields that follow this paper layout[76]. Ten years ago, these venues were nowhere to be seen in the journal leaderboard. A true revolution.

It's easy to see why. Traditional science epistemology was heavy with counterfactuals. Authors discussed consequences of some hypothetical intervention on hypothetical measurements and compared those with observed measurements. In fact, over the last century, human intelligence was more or less equated with the articulate use of counterfactuals, conditionals, and hypotheticals in discourse. The new context removes the burden of articulate deployment and parsing of counterfactuals.

9.2. **Research Mindset.** There is also a 'disruption' in inter-researcher professional discussion. When I observe today's younger data scientists, I note that their typical interactions exchange information about their information technology stacks. Typical questions include:

- What's your package name?
- What's your URL? QR Code?
- Is package $< X >$ on your stack? ($< X >$ varies from conversation to conversation)

---

[76]https://scholar.google.com/citations?view_op=metrics_intro&hl=en, Accessed September 29, 2023. The new entrants are IEEE Conf. CVPR, NeurIPS, ICLR. Their new peers are Nature, New Eng. Journ. Medicine, Science, Lancet,Adv. Materials, Cell

More generally, they are often asking for details underlying the FR Triad: They want to know are you Data Sharing? Using some new shared dataset I haven't previously heard of? are you Code Sharing? Using some new shared codebase I haven't previously heard of? Is there some new specific numerical performance measure? Is there a challenge that just dropped? How can I get single-click access to the work you are describing? These are all signs that **these researchers are participants in an FRX** - a research community exchanging digital research artifacts.

Traditional intellectual discourse was heavy with conditionals and counterfactuals – i.e. thought experiments. Computation was so difficult that the ability to run mental simulations of what computations might deliver, if we could conceivably do them, was a sign of real intellectual penetration. Today it's a big turn-off to propose thought experiments and impossible hypotheticals like 'you could potentially try to do such-and-such'. One might as well suggest a fly-by of Venus.

The currency of discourse today is Frictionless Replications. Researchers are thinking: *Can't I just try this now?*

10. Actions Readers can take Immediately

Here are some action items implied by the argument so far.

10.1. **Exemplify Frictionless Reproducibility.**
In line with the spirit of the times, try to make your own research work capable of viral spread among data scientists. You can best do this by having it be single-click reproducible.

Suppose you are a mathematical scientist who thought of your job as mathematically describing a data science procedure and creating a theorem to probe its behavior. That approach worked for my generation, but it runs afoul of the 'URL or it didn't happen' mindset of modern life. Today, you might also implement your tool computationally, share code, evaluate it on shared data, and make formalized quantitative comparisons of your work with standard baselines. Implicitly in each research product you can be either participating in an existing challenge, or creating a new reproducible challenge problem and proposing its first entry. If there are today no relevant datasets or baselines, you can productively work to create them. Creating a dataset or challenge can have much more impact than merely theorem-izing a specific tool. If you don't immediately see a way to integrate the three elements [FR-1]+[FR-2]+[FR-3] into your work, sense this as a puzzle to unlock an opportunity.

Alternatively, suppose you work in a non-mathematical field, let's say in medicine, that does not currently use one or more legs of the triad. Work to change this, for example instituting the sharing of data and code and documenting performance. Such work can be visionary and foundational, and there can be grants to support it. This has been proven for example by the Nightingale Open Science project. Medicine is just one example; this pattern can repeat throughout science and technology.

10.2. **Research Opportunities.**
Dataset Bias. Are the shared public databases we depend on for challenges a 'reasonable representation' of the dataset universe, or are selective/opportunistic publication in some way skewing our methodological research? Are there ways to better incentivize Datafication? Data Sharing? Are there negative consequences of this approach, perhaps ones we can guard against or mediate?

Are there types of datasets that are flatly incompatible with this approach? Data that are changing? Datasets that are too small? Too Large? Are there adaptations of the whole paradigm that should be adopted? See for example [Varoquaux and Cheplygina, 2022].

Availability Bias. I perceive a heavy bias of researchers towards certain popular approaches over others; often this is *because of* [FR-2: Code Sharing]. Thus, deep learning is considered 'first priority' by today's empirical ML researchers whereas, traditional tree-based prediction methods like CART and random forests are considered deprecated, perhaps because the shared code resources for deep learning became more **convenient to access** - *although they are much more computationally expensive to apply.* For many non-image, non-language so-called 'tabular' datasets, the tree-based methods make more intuitive and practical sense. But if few researchers want to use them in challenge entries, we face ignorance about whether those tools work, or how they compare.

Challenge Bias. In the new regime, we have groups of researchers competing to improve performance in task performance on shared public databases. There are predictable consequences of this approach, such as *overfitting* but perhaps also other negative ones we hadn't thought of? Are there novel computational tools to ward off the negative effects [Neto et al., 2016], or adjust for them?

Challenge Design. How do we design better Challenges? How do we invent performance criteria for important problems that never were subject to challenges before, thereby enabling access to a new paradigm. [J. Liu et al., 2022] gives an intriguing new avenue: *reproducibility challenges.* Can we design challenges to promote *interpretability.*

Exchange Design. How do we entice a group of researchers to participate in our challenge? What practices work and don't work? How do we ignite a group of researchers into a dynamic, productive FRX? How do we catalyze FRX takeoff?

Resource Blindness. A particular feature of the new regime is that doing things just about as well but dramatically more efficiently is typically not at all valorized. Have brutal scaling and blind focus on performance obscured tremendous efficiencies lying in plain sight? What would the more efficient methods be? How could uncovering them be valorized effectively? Can we design challenges to promote more efficient algorithms?

10.3. **Teaching Opportunities.** *Push Back against Faulty Narratives!* This paper, together with its companion, exposes a clash of narratives, between the highly visible 'AI Singularity', and 'Brutal Scaling' narratives and the contrasting narrative that I prefer, the little-known but historically grounded 'Data Science finally takes hold'. The AI Singularity narrative promotes powerlessness and is demoralizing to the young. Data science instruction can proclaim an accurate and hopeful counter-narrative.

This narrative will better equip students for the future and induce more respect of students for the material they are learning and for the value of their own educations. It can be injected everywhere in academic data science curriculum.

New Advanced Courses. In Fall 2023, Xiaoyan (XY) Han and I taught Statistics 335 at Stanford on the Challenge Problems Paradigm, its components and it consequences, both in ML and elsewhere. As with other 300-level courses in my department, this course may have PhD students and ambitious masters students enrolled. We engaged students in many of the issues discussed in this paper, in particular, to thinking about the history to date of this approach, some of its greatest achievements and some variations and modifications.

New Undergraduate Courses. The game has changed for data-driven methodological development, but actual instruction has not adapted. The National Academies conducted a study [National

---

Narrative: Data Science Takes Hold

(a): A remarkable confluence of developments took place in the last decade, creating a powerful information technology facilities stacked on a global compute, storage, and communications resource.

(b): Because of these new facilities, the global economy is going through a multiple-decades-long total digital transformation, which will gradually change everything: business, government, education, science.

(c): In particular, there are many exciting new careers and opportunities adapting these technologies to human needs; in these careers we deploy these new facilities to develop new models and processes.

(d): The data science principles [FR-1]+[FR-2]+[FR-3] increasingly become the go-to paradigm for research-based model building. Let's learn this paradigm, learn some of its history and successes, and follow it in our own careers.

Academies of Sciences Engineering and Medicine, 2018] of the undergraduate data science curriculum in 2017-2018[77]; this already looks in need of severe revision. Many of today's students simply wouldn't be able to get value from statistics courses taught the way they were taught before 2010, yet some courses at some institutions haven't really adapted. The profusion of publicly shared datasets and data analysis scripts, along with on-line teaching resources such as youtube videos, has already changed the expectations of students, about what they want to learn and how; in particular students have the expectation of frictionless access to tools and other artifacts and traditional coursework defeats this. Probably we don't yet agree on what comes next, but our students are in a completely new mindset and completely new approaches are needed.

10.3.1. *Software Developer opportunities.* Frictionless reproducibility is here, now; in the future we will see more and more research fields following its discipline, meaning there is an opportunity for services and tools to support and build upon the triad [FR-1]+[FR-2]+[FR-3], including outside the context of empirical machine learning. To me the big opportunity lies in making some of the predictions listed below, come true.

11. Predictions

This paper began life for a session on 'The Next 50 years of Data Science' at the 2023 Joint Statistical Meetings. Implicitly, we were being asked for predictions; here are several. See also [D. Donoho, 2017].

11.1. **Computation on Research Artifacts (CORA).** Frictionless Reproducibility in computational research is now essentially here. As a side effect this makes available a vast array of digital research artifacts generated by earlier research. Soon, those byproducts will be viewed as digital gold.

Here's how. In many research disciplines, reports spotlight tables which constitute the main deliverables of the work being presented. Such tables bring together chosen data and algorithms and specific task performance metrics. In methodological work, they can often be viewed as what an

---

[77]NASEM: "Envisioning the Data Science Discipline"

EML researcher would call a 'mini-challenge leaderboard'; I am speaking here inclusively, broadly, outside the EML context, and allow an enlarged definition of task and task performance.

In the near future, researchers will understand that future researchers, digesting the given research work, would wish to directly probe the computations that generated the numbers in the researcher's tables, as a way of engaging with the researcher's result and possibly building upon it. Whole communities will expect their researchers to expose the data and workflow that produced the surfaced numerical results.

New tools will soon enough become available to automatically enable frictionless access to the underlying digital artifacts that went into any relevant table in a study. This will happen seamlessly, as part of ordinary research and publication; standard research computing and publication computing environments, with little or no effort, will automatically expose the underlying data, code and performance measurements that delivered these results.

Exposing such artifacts enables other data scientists who read a published report to build upon the research that has been done in that report: to first reproduce and, later, to possibly modify some underlying algorithms or some underlying observables and recompute the table, thereby producing a new research project.

This enters us all into a new era, where **computing on** the digital **research artifacts** created by previous research computing (for which we adopt the label CORA) allows us to write algorithms to inspect and generalize the workflows used in previous research, and thereby algorithmically obtain new research results. We will begin to naturally think and operationalize at a new level as follows: "Take project X's workflow and everywhere replace algorithm A with algorithm B, call the result 'project Y's workflow'. Execute the full workflow of project Y, computing all the same observables as you did for Project X."

As a vision, CORA has been clear to many data scientists for some time. The Quantitative Programming Environment Mathematica and its associated notebook system delivered some of this functionally already, at least 30 years ago. Stephen Wolfram in several venues has evoked his vision of **computing on anything**, broad enough to include CORA[78].

CodaLab[79],[80] is a platform for running challenge problem contests that can inherently perform CORA, for certain specific tasks (e.g. swap the public test dataset that was used by a certain contest entry, for a certain private dataset sequestered from public view, and re-evaluate specific performance metrics after the swap).[81].

My prediction is that CORA will inevitably emerge and grow in scope, to outstanding effect, increasing the pace of scientific advance. My thinking derives from (practically unknown) papers by the author and Matan Gavish [Gavish and Donoho, 2011, 2012] in which we described the notion of *Verifiable Computational Results* (VCR), and three dream applications.

VCR, where implemented, ensures that digital artifacts emerging during a computation are automatically captured and made frictionlessly available to future researchers; and dream applications

---

[78]To share the 'vibes', I change one word in this 'typical' Wolframism: '... we are finally now getting into a position where we can take all that *[RESEARCH]* that we as humans have accumulated in our civilization, and encapsulate it in an active computational form. And when we do this, we make it possible to dramatically extend and generalize all our achievements so far.' (almost-quoted-from) Stephen Wolfram, *On the Quest for Computable Knowledge, 2017.*

[79]https://worksheets.codalab.org/

[80]https://codalab.lisn.upsaclay.fr/

[81]The NMRBox system discussed in Section 6 has some analogous features

exploit these artifacts to automatically generalize upon earlier work (eg swap this computational method for the one originally used in prior research, and show me the new table that results) or automatically evaluate earlier research (eg bootstrap this entry in that previously published table and show me the histogram that results).

It seems to me that full implementation of CORA, when it happens, will be a transformative achievement! Also, that we as a scientific computing world, are headed towards that achievement, and its fulfillment dates *the actual moment of the reproducibility singularity.*

11.2. **CORA and the 'End Times'.** Kurzweil calls attention to a very dramatic moment, coming soon – yielding priority to IJ Good, *Speculations concerning the first ultraintelligent machine* [Good, 1966]:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

Good was an amazing intellectual [Good, 1983], who served as Alan Turing's research assistant and who 'Kept Bayes Alive' during its exile from academic favor [Neapolitan, 2008]. As you can see, he was a stylish writer and his formulation is charismatic and compelling. It's easy to see why Kurzweil latches onto it and how it might inspire "AI Doom" influencer threads; Eliezer Yudkowsky coined[82] the notion of a "hard takeoff" or FOOM[83], where machine intelligence starts to expand its own powers hyper-exponentially. Good seems to anticipate hard takeoff.

Back to Earth. Let's acknowledge that, in certain ways, performance on challenges is inevitably and automatically going to improve markedly post-transition to FR. Indeed, it is a simple matter to take any given frictionlessly reproducible workflow and modify it by trial and error to produce a new workflow that may improve it. Consider two meta-operators on workflows:

**Transformation 1: : **Wrappers
- Adopt a given existing workflow as a starting point.
- Prepend a new workflow element to preprocess the inputs
- Append a new workflow element to postprocess the outputs
- Obtain a new workflow

**Transformation 2: : **Tweaks
- Adopt a given existing workflow as a starting point.
- Assuming this workflow exposes internal hyperparameters that were previously unexplored:
    - Systematically vary those unexplored hyperparameters
    - Extract the optimal hyperparameter combination
    - Obtain a new workflow

Much of the EML research activity we have seen in recent years reports discoveries of successful applications of these two strategies, in many cases adding new layers/gadgets to a multilayer architecture (Transformation 1); or modifying an existing architecture by, let's say, adding extra width

---

[82]https://www.lesswrong.com/posts/tjH8XPxAnr6JRbh7k/hard-takeoff

[83]https://www.lesswrong.com/tag/ai-takeoff

(Transformation 2). The new models produced by such meta-operations can then be evaluated robotically to find one which improves performance.

This has been automated in empirical machine learning (eg. with famous systems like AutoML [Hutter et al., 2015; Hutter et al., 2019; Karmaker ("Santu") et al., 2021], and Vizier [Golovin et al., 2017; Song et al., 2023]); the process points in the direction of Jack Good's vision. We avoid human guesswork and blind luck to get good models; hyperparameter search is done to a certain degree robotically.

These meta-operators involve operations on workflows [Drori et al., 2021] – digital artifacts produced by previous research computing. Hence these are instances of what has earlier been discussed and labelled CORA. In any field that invests in the FR regime, the digital artifacts are there, as a side effect of publication, available frictionlessly to be exploited. It is just a matter of adopting the mindset to exploit the artifacts to do future research.

As individual fields invest in FR, they will transition to CORA; the transition may even enable algorithmic improvement of models and produce a hard takeoff of model performance. Maybe the last decade's steep increase in performance in EML is an approximation to hard takeoff; the approximation manifesting the 'still-slightly-frictioned', 'still-human-handwork-was-involved' 'CORA-was-still-in-its-infancy' situation over the last decade.

Will CORA make machines hyper-intelligent by itself? No. All we will get via automated application of strategies like those listed here is incremental bumps to performance. We will painlessly and rapidly mine out the value of each human-supplied idea.

Peter Huber, in his 2012 book *Data Analysis: What can be learned from the last 50 years* [Huber, 2012] contrasts strategy (high-level wisdom) and tactics (low-level, step-by-step). Using his terminology, we predict that, as FR takes hold in a field, and CORA becomes ubiquitous, humans will increasingly focus on strategic choices about model architectures and types of hyperparameters; while increasingly, information technology will automate the tactical details of finding the best instances within those strategic constraints.

11.3. **What's Next.** Certain features of research life going forward from that point are very clear.

- *Irreversibility.* The transition to FR is a one-way transition: the wishes underlying [FR-1]+[FR-2]+[FR-3] have been around as long as there has been data-driven science; the ability to fulfill those wishes is now provided by a complete technology stack which came into existence to service the global communications/computation infrastructure. That infrastructure is here to stay. The desire for FR won't go away; the ability to offer FR won't go away either.

- *Acceleration.* Research fields that have transitioned to the Frictionless Reproducibility regime will progress even faster in the future as the habits get even more ingrained, producing even higher quality with ever higher velocity. Further, a community of researchers working together according to this pattern will recognize and seize new opportunities as a result.

- *Stagnation.* Some fields won't or can't practice mature forms of the three data science initiatives we have discussed. Barriers to transition can include: inhibitions against data sharing, for example, because of confidentiality restrictions; inhibitions against code sharing, for example because of proprietary restrictions; inhibitions against prediction challenges, for example because of a preference for theoretical derivations disconnected from empirical measurements.

Such fields won't transition to a Frictionless Reproducibility regime. They will be noticeably lagging behind in rate of progress. Those fields may soon enough be recognized as backwaters.

- *Resource Allocation.* Funding agencies will recognize that they get higher rates of return on investments in research communities which have transitioned to Frictionless Reproducibility; they will invest more heavily in such fields over time. Research talent will follow the dollars. The separation between fields, initially from reproducibility practices, will then involve funding, compounding the contrast between accelerating and stagnating fields.

Some other features of our future are less obvious.

- *Opacity.* A key point made by Kurzweil in "The Singularity is Near" is that progress "on the other side" of "The Singularity" will be so amazing, we can't even imagine what will be accomplished. This seems fair to me. Seniors among us became used to the idea that certain research topics progress slowly or not at all; are best learned by dint of long study and philosophical reflection; and research contributions happen on the time scale of entire careers. Minds trained in such a regime might not adjust to how rapidly things will soon be happening.

    Of course some research topics are, so far, untransformed by the three data science initiatives. But eventually some will be so transformed, and research in such areas will rapidly move far beyond what experts of earlier generations would have thought possible, or even conceivable. Others may be, essentially forgotten. Hence, the Reproducibility Singularity is indeed opaque.
- *Amnesia.* The opacity of the Reproducibility Singularity works in reverse. Much of the work pre-reproducibility transition will be forgotten. Listening to the young makes this clear.

    Current Stanford Statistics Ph.D. student Apratim Dey told me about a Stanford CS Class on "Distribution Shift" in Spring 2023. In that class, the initial course reading was a more or less traditional academic Statistics journal article from 20 or so years ago that established the field (more or less). That paper posed a generative probabilistic model for the distribution shift problem and performed a formal analysis using it. Later in the term, papers came from leading CS conferences and followed a more empirical bent, using shared data and code.

    The CS students in this course apparently were not fans of the academic statistics paper, although they were told it was fundamental. In discussing this paper, they pursued critiques along these lines:
    - It's based on a theoretical generative model
    - I don't know how to tweak or improve a theoretical generative model
    - I don't know if this generative model describes any real data
    - I don't know if the phenomena it describes occur on any real data
    - If I do research on real data, I can share results with others, publish in a CS conference.
    - If I instead do research with a generative model, as is done here: I have no immediate outlet for my work, no readership, no follow-up. It could take years and years to publish in a traditional stats journal.

    Maybe these are just comments about one paper; but possibly they are comments about papers of this type. Namely, papers that don't hew to today's empirical data science

style, which involves building data and code and empirical performance on the work of others. The CS students don't see the older foundational research as connecting to today's frictionlessly reproducible regime; it doesn't offer the opportunity for single-click adoption and code re-execution by students. Engaging with such non-conforming work disrupts the students' state of 'flow'; hence such papers are seen as frustrating. Such work confronts students, in their view, with gotchas and obstacles. Also they don't see how engaging with it allows them to efficiently progress to publication in the outlets they prefer. Such pre-FR work offers friction, not fulfillment.

This younger CS generation is steeped in the recent empirical machine learning literature, much of which is powered by FR/FRX; it seemingly doesn't warm to the style of mathematically powered discourse practiced previously, which is often based on hypotheticals and counterfactuals; just maybe, it doesn't see the need to learn how to engage with such work.

We may be starting a rapid de-skilling transition, in which new researchers initially don't want to, and later simply can't, onboard the lessons of symbolic analysis of generative models.

12. Conclusion

A data-science-driven phase transition, crossing a kind of singularity, is happening now.

Not the 'AI singularity' — the moment when AGI is achieved — this hasn't happened. Rather, some fields in computation-driven research are approaching a Reproducibility Singularity, the moment when all the ingredients of frictionless reproducibility (FR) of computational research are assembled in a package that makes it, essentially, immediate to build on the work of other researchers on an open, global stage. A research community working fully within this regime creates and exchanges digital artifacts following new models of scientific publication and epistemology; this is a research superpower enabling strikingly rapid progress.

This new regime has been best approximated within the field of empirical machine learning (EML), which indeed has made very rapid progress over the last decade in the fields of computer vision and natural language processing.

The FR regime is best thought of, not as 'property' of EML, but instead as a more general achievement of dedicated Data Scientists working for decades, working in many research disciplines, including those where empirical machine learning and AI historically plays little role.

Today's 'AI Singularity' narrative – that the last decade was the decade where AI was achieved – is premature; this was instead the decade when data science matured, where a global computing and communication infrastructure came together, and frictionless delivery of research artifacts emerged, enabling computational science as a solid new avenue to scientific validity.

The next generation of data scientists can leverage this understanding in building their own careers to identify interesting questions to work on and to make their research impact more viral.

whom I learned a great deal about the global, sciencewide nature of the changes going on in science and the advances in reproducibility. Also, I thank my fellow members on the NASEM study on Envisioning the Data Science Discipline: The Undergraduate Perspective(David Culler, Chair) for making me aware of the many currents of change in Data Science education.

I had the fascinating chance to witness 'from inside' the NUScon challenge emerge in NMR Signal Processing; many thanks to Adam Schuyler and Jeffrey C. Hoch (University of Connecticut HealtH Sciences Center).

Certain 'insiders' corrected various comments I made about their work in an earlier draft. Thanks for these corrections are due to: Isabelle Guyon (Université de Paris), John Moult (University of Maryland), Jonathon Phillips (NIST), Benjamin Recht (University of California, Berkeley), and Steven Ruberg (Analytix Thinking).

The Editor of *Harvard Data Science Review*, Prof. Xiao-Li Meng, and the anomymous reviewers have materially improved the manuscript through several revisions. I am deeply grateful for their willingness to engage in this process with full attention and interest.

Brian Wandell (Stanford) provoked the title. In alphabetical order, Milad Bakhshizadeh(Stanford), Apratim Dey (Stanford), Andrew Donoho (Donoho Design Group), Xiaoyan Han (Cornell), Vardan Papyan (Toronto), Elad Romanov (Stanford) and Yu Wang (Stanford) made many helpful comments. This is closely related to work in progress by the author and Matan Gavish (Hebrew University); I also want to acknowledge Victoria Stodden (University of Southern California) for many conversations on reproducibility over the years.

REFERENCES

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Res*, *41*(Database issue), D36–42. https://doi.org/10.1093/nar/gks1195

Blasco, A., Endres, M. G., Sergeev, R. A., Jonchhe, A., Macaluso, N. J. M., Narayan, R., Natoli, T., Paik, J. H., Briney, B., Wu, C., Su, A. I., Subramanian, A., & Lakhani, K. R. (2019). Advancing computational biology and bioinformatics research through open innovation competitions. *PLoS One*, *14*(9), e0222165. https://doi.org/10.1371/journal.pone.0222165

Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Jackel, L. D., LeCun, Y., Muller, U. A., Sackinger, E., Simard, P., et al. (1994). Comparison of classifier methods: A case study in handwritten digit recognition. *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, *2*, 77–82.

Bourla, A. (2022). *Moonshot: Inside Pfizer's nine-month race to make the impossible possible*. Harper Collins.

Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and Reproducible Research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and Statistics* (pp. 55–81). Springer New York. https://doi.org/10.1007/978-1-4612-2544-7_5

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766.

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., & Stodden, V. (2009). Reproducible Research in Computational Harmonic Analysis. *Computing in Science and Engineering*, *11*(1), 8–18. https://doi.org/10.1109/MCSE.2009.15

Drori, I., Krishnamurthy, Y., Rampin, R., de Paula Lourenco, R., Ono, J. P., Cho, K., Silva, C., & Freire, J. (2021). AlphaD3M: Machine learning pipeline synthesis.

Gavish, M., & Donoho, D. (2011). A universal identifier for computational results. *Procedia Computer Science, 4*, 637–647.

Gavish, M., & Donoho, D. (2012). Three dream applications of Verifiable Computational Results. *Computing in Science & Engineering, 14*(4), 26–31.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., & Sculley, D. (2017). Google Vizier: A service for black-box optimization. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1487–1495.

Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers* (pp. 31–88, Vol. 6). Elsevier.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications.* U of Minnesota Press.

Gribov, A., Sill, M., Lück, S., Rücker, F., Döhner, K., Bullinger, L., Benner, A., & Unwin, A. (2010). SEURAT: Visual analytics for the integrated analysis of microarray data. *BMC Medical Genomics, 3*(1), 21. https://doi.org/10.1186/1755-8794-3-21

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., Goecks, J., Backofen, R., Nekrutenko, A., & Taylor, J. (2018). Practical computational reproducibility in the Life Sciences. *Cell Systems, 6*(6), 631–635. https://doi.org/https://doi.org/10.1016/j.cels.2018.03.014

Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Ho, T. K., Macià, N., Ray, B., Saeed, M., Statnikov, A., et al. (2015). Design of the 2015 ChaLearn AutoML challenge. *2015 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Guyon, I., Dror, G., Lemaire, V., Taylor, G., & Aha, D. W. (2011). Unsupervised and transfer learning challenge. *The 2011 International Joint Conference on Neural Networks*, 793–800.

Guyon, I., Gunn, S., Hur, A. B., & Dror, G. (2006). Design and Analysis of the NIPS2003 challenge. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds.), *Feature extraction: Foundations and applications* (pp. 237–263). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35488-8_10

Guyon, I., Saffari, A., Dror, G., & Cawley, G. (2008). Analysis of the IJCNN 2007 agnostic learning vs. prior knowledge challenge. *Neural Networks, 21*(2-3), 544–550.

Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., & Janet, S. (1994). UNIPEN project of on-line data exchange and recognizer benchmarks. *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5), 2*, 29–33.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE intelligent systems, 24*(2), 8–12.

Hardt, M., & Recht, B. (2022). *Patterns, predictions, and actions: Foundations of machine learning.* Princeton University Press.

Huber, P. J. (2012). *Data Analysis: What can be learned from the past 50 years.* John Wiley & Sons.

Hutter, F., Kégl, B., Caruana, R., Guyon, I., Larochelle, H., & Viegas, E. (2015). Automatic machine learning (AutoML). *ICML 2015 Workshop on Resource-Efficient Machine Learning, 32nd International Conference on Machine Learning.*

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. Springer Nature.

Jiang, Y., Foret, P., Yak, S., Roy, D. M., Mobahi, H., Dziugaite, G. K., Bengio, S., Gunasekar, S., Guyon, I., & Neyshabur, B. (2020). NeurIPS 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*.

Karmaker ("Santu"), S. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2021). AutoML to date and beyond: Challenges and opportunities. *ACM Comput. Surv.*, *54*(8). https://doi.org/10.1145/3470918

Kryshtafovych, A., Fidelis, K., & Moult, J. (2010). *CASP: A driving force in protein structure modeling*. Wiley.

Kryshtafovych, A., Fidelis, K., & Moult, J. (2014). CASP10 results compared to those of previous CASP experiments. *Proteins: Structure, Function, and Bioinformatics*, *82*, 164–174.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1607–1617.

Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. Penguin.

Liu, J., Carlson, J., Pasek, J., Puchala, B., Rao, A., & Jagadish, H. V. (2022). Promoting and Enabling Reproducible Data Science Through a Reproducibility Challenge. *Harvard Data Science Review*, *4*(3). https://hdsr.mitpress.mit.edu/pub/mlconlea

Liu, Z., Pavao, A., Xu, Z., Escalera, S., Ferreira, F., Guyon, I., Hong, S., Hutter, F., Ji, R., Junior, J. C. J., et al. (2021). Winning solutions and post-challenge analyses of the ChaLearn AutoDL challenge 2019. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(9), 3108–3125.

Lustig, R. H. (2018). *The Hacking of the American Mind: The science behind the corporate takeover of our bodies and brains*. Penguin.

Moult, J. (2005). A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, *15*(3), 285–289.

Moult, J. (2006). Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1467), 453–458.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics*, *86*, 7–15.

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods.

National Academies of Sciences Engineering and Medicine. (2018). Envisioning the Data Science Discipline: The Undergraduate Perspective.

National Academies of Sciences Engineering and Medicine. (2019). Reproducibility and Replicability in Science.

Neapolitan, R. E. (2008). A Polemic for Bayesian Statistics. In D. E. Holmes & L. C. Jain (Eds.), *Innovations in Bayesian Networks: Theory and applications* (pp. 7–32). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-85066-3_2

Neto, E. C., Hoff, B. R., Bare, C., Bot, B. M., Yu, T., Magravite, L., Trister, A. D., Norman, T., Meyer, P., Saez-Rodrigues, J., Costello, J. C., Guinney, J., & Stolovitzky, G. (2016). Reducing overfitting in challenge-based competitions. *arXiv preprint arXiv: 1607.00091*.

Pavao, A., Guyon, I., Letournel, A.-C., Tran, D.-T., Baro, X., Escalante, H. J., Escalera, S., Thomas, T., & Xu, Z. (2023). CodaLab Competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, *24*(198), 1–6. http://jmlr.org/papers/v24/21-1436.html

Peng, R. D. (2009). Reproducible Research and Biostatistics. *Biostatistics*, *10*(3), 405–408. https://doi.org/10.1093/biostatistics/kxp014

Peng, R. D., & Hicks, S. C. (2021). Reproducible Research: A Retrospective [PMID: 33467923]. *Annual Review of Public Health*, *42*(1), 79–93. https://doi.org/10.1146/annurev-publhealth-012420-105110

Phillips, P. J., Flynn, P. J., Beveridge, J. R., Scruggs, W. T., O'toole, A. J., Bolme, D., Bowyer, K. W., Draper, B. A., Givens, G. H., Lui, Y. M., et al. (2009). Overview of the Multiple Biometrics Grand Challenge. *Advances in Biometrics: Third International Conference, ICB 2009. Proceedings 3*, 705–714.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., & Worek, W. (2005). Overview of the Face Recognition Grand Challenge. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, 947–954.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(10), 1090–1104.

Phillips, P. J., Sarkar, S., Robledo, I., Grother, P., & Bowyer, K. (2002). The Gait Identification Challenge Problem: Data sets and baseline algorithm. *2002 International Conference on Pattern Recognition*, *1*, 385–388.

Ringwald, M., Eppig, J. T., & Richardson, J. E. (2000). GXD: Integrated access to gene expression data for the laboratory mouse. *Trends Genet*, *16*(4), 188–190. https://doi.org/10.1016/s0168-9525(00)01983-1

Rosato, A., Bagaria, A., Baker, D., Bardiaux, B., Cavalli, A., Doreleijers, J. F., Giachetti, A., Guerry, P., Güntert, P., Herrmann, T., Huang, Y. J., Jonker, H. R. A., Mao, B., Malliavin, T. E., Montelione, G. T., Nilges, M., Raman, S., van der Schot, G., Vranken, W. F., . . . Bonvin, A. M. J. J. (2009). CASD-NMR: Critical assessment of automated structure determination by NMR. *Nat Methods*, *6*(9), 625–626. https://doi.org/10.1038/nmeth0909-625

Ruberg, S., Zhang, Y., Showalter, H., & Shen, L. (2022). A platform for comparing subgroup identification methodologies. *Biometrical Journal*, *n/a*(n/a), 2200164. https://doi.org/https://doi.org/10.1002/bimj.202200164

Song, X., Perel, S., Lee, C., Kochanski, G., & Golovin, D. (2023). Open Source Vizier: Distributed infrastructure and api for reliable and flexible blackbox optimization.

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine*, *5*(1), 48. https://doi.org/10.1038/s41746-022-00592-y

Wigner, E. P. (1960). The unreasonable effectiveness of Mathematics in the Natural Sciences. *Communications on Pure and Applied Mathematics*, *13*(1), 1–14. https://doi.org/https://doi.org/10.1002/cpa.3160130102

Wilan, K. (2007). Profile: Alpheus Bingham. *Nature Biotechnology*, *25*(10), 1072–1072. https://doi.org/10.1038/nbt1007-1072

Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.-W., Yao, Q., Zhao, H., & Guyon, I. (2022). Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, *3*(7), 100543. https://doi.org/https://doi.org/10.1016/j.patter.2022.100543