Notes on 10x-100x Research Opportunities

Ravi B. Sojitra
Stanford University

First draft: April 9, 2025
This draft: April 12, 2025

Summary

- Last week, I attended two Stanford Artificial Intelligence (AI) workshops: one on AI Alignment for Businesses and one on AI and Experiment Design for Bioengineering.
- My goal was to understand which applied statistics, causal inference, and machine learning contributions would accelerate the pace of innovation and scale of impact.
- I documented impressions and observations based on content, talks, and discussion. They are broadly applicable (e.g. to marketing, product, search, ranking, targeting).
- Submit feedback by creating a GitHub Issue at the `public-notes` repository. Thanks!

Motivation

Eleven years ago, I shifted my focus from biology and neuroscience to applied statistics, causal inference, and machine learning. I did this because I love the latter's pace of innovation and real world, world-wide impact via industrial adoption. I also enjoy the work.

One recent example is the impact of generative models on biology research and development (R&D). More specifically, de novo drug design, protein folding prediction, protein synthesis, genomics, and transcription. At a Stanford workshop on AI and Experiment Design for Bioengineering, I learned that R&D in machine learning (e.g. diffusion models), data curation efforts (e.g. the Protein Data Bank), and computing (e.g. software, hardware) *already* unlocked hundreds of product categories, technological capabilities, and regulatory challenges. I have a feeling that I am understating the impact potential.

Question: How can we bring the future to the present, faster?

I typically attend workshops to identify methodological research directions that could accelerate innovation. Last week, my focus was on agentic AI products and bioengineering. I attended two AI workshops at Stanford: One on AI alignment for Businesses and one on AI and Experiment Design for Bioengineering.

Positive Answers (10-100x Opportunities)

Automate and Scale *Custom* Data Collection and Experiment Design

When empiricists learn more about mechanisms in an experiment, the next steps are usually to measure new or different things with greater precision. Methods for automatically designing cheap and fast measurement tools (e.g. imaging or data logging) given a user's needs are game changers.

One recent example of a small success is related to detecting and measuring content policy violations on digital platforms. Across multiple workshops at Stanford last year, I learned that companies are consistently finding that Multimodal Large Language Models can be cleverly used to improve measurement and detection of content policy violations. Violations that were previously difficult to detect and label are now being caught at scale. I am willing to bet this has become or will become common practice because the impact is outstanding.

In general, measurement remains a major friction to innovation even at well resourced internet companies. For example, content ranking teams frequently run experiments to evaluate ranking algorithm changes. If we suspect that a treatment arm in an experiment has a negative impact on user experience even though engagement and revenue increased, we should check whether there was an increase in content reported by users or rates of content deletion. However, if this information was not logged ahead of time, a nontrivial amount of engineering investment and statistics expertise is required to log the necessary data before the experiment can be re-run. We have to make sure it is done correctly and with imperceptible latency for users.

In theory, "cloud laboratories" could fulfill this function, but the current quality and scale of services are unsatisfactory.

Build Better AI Agents for Tasks that Distract from Intellectual Work

We should be striving to reallocate time from mundane tasks (e.g. diagnosing contamination issues in experiments, exploratory data analysis) to planning higher quality and more ambitious research designs. The latter is a better use of scientists' time and expertise.

Existing Agentic AI solutions are not (yet) reliable, but there exist compelling proofs of concept. There is currently lot of discussion about Agentic AI solutions for research use cases across academia and Silicon Valley, so I did not learn anything new. I suppose this is one of the privileges of being a graduate student at Stanford University right now.

Plan for and Shape the Future of Compliance and Privacy Engineering

I learned a lot of new things. The most important lesson for me is that there is no clear industry standard (yet) on how to approach compliance and privacy engineering for Agentic AI products. The main paying customers are businesses, so meeting their compliance and privacy needs is necessary for generating revenue. I documented more notes below.

This theme is consistent with my impression about machine learning and AI products in general. Moving forward, I expect methodological contributions that raise fewer red flags during legal, policy, and privacy reviews will have faster and broader adoption.

Shape the Future of Patents and Regulatory Frameworks

Experts noted that patent offices are not prepared for emerging challenges. For example, since it has become easier to synthesize artificial proteins, it has become easier to copy patented molecules while being "different enough".

I imagine companies are being founded with the intention of patenting millions (or more) of the most promising biological and chemical products and licensing them for revenue later on. The return on investment would be tied more closely to the quality scientific evaluation compared to startups we are used to seeing. As a scientist, it is exciting to imagine a future where scientific discovery is the core business function, but I am worried that the current thinking about patents has put us on the wrong path.

Scientists need to actively inform paths forward (e.g. introducing quantitative definitions of novelty and ownership to support qualitative desiderata). There is already public discourse about artwork and copyright, but I have heard little about novel physical product categories.

Negative Answers ($<$ 2x Opportunities)

Increase Statistical Power ("Sample efficiency")

This is a favorite topic among methods researchers. Algorithms from active learning and Bayesian optimization research are tried and true methods for cost cutting. However, the impact of ***new*** methods on the pace of innovation appears to be qualitatively small. This suggests to me that other blockers and frictions are the limiting factors. For example, academic and industrial norms or underinvestment in planning quality research designs.

Raise the (Publication) Bar

One discussion point was whether the research publication bar ought to be raised. In particular, the question was whether randomized experiment results should be required for publication. The rationale is that the volume of papers is unmanageable and accelerating in growth, but most of it is underwhelming, unimpactful, and dishonest about the facts. Nevertheless, the consensus is usually that these efforts will slow the pace of innovation.

Notes on AI Alignment for Businesses

Compliance with regulation and company policies was a major theme. Providing a commercial Agentic AI product is challenging when you need to control who is allowed access to what information. For example, you may not want your performance evaluation or social security number available to everyone who uses the company's Agentic AI tools.

Alignment was defined more precisely than I have seen in engineering and computer science departments. Not surprisingly, in practice, this is defined roughly as "alignment with respect to the company's governance structure".

Example desiderata of a generative AI based search or agent system.

- You want the Agentic AI system to respect governance structures. For example, decision making permissions and information access for different groups of employees.
- You may not want to disclose information to people outside your reporting chain.
- You want to be aware of the documents and people you are elevating as experts when providing AI search results.
- You want to remove political bias from internal AI systems.

Technical challenges in Agentic AI systems.

- How should you ingest data from various sources (e.g. Google Docs, Microsoft Word, Overleaf, text files, PDFs, data lakes, other tools)?
- How should you create datasets from conversations between employees (e.g. discussions between networks of people on Slack across different chat channels)?

Notes on AI and Experiment Design for Bioengineering

The conference program highlighted work on AI agents, Active Learning, Bayesian Optimization, Diffusion Models, and Transformer based Sequence Models. Other topics were discussed, but not emphasized. For example, there were also presentations on mathematical optimization and conformal inference. There was surprisingly little discussion about compliance, regulation, and property rights.

Overall, there was a lot of excitement about generative models. Diffusion models have more attention, empirical evidence, and impact than I expected. Conditional sampling using diffusion models modeling stood out to me in particular. I am not sure to what extent this excitement is due to audience selection bias, but the empirics are impressive.

Empirical Observations Noted by Experts

- The Reinforcement Learning From Human Feedback (RLHF) paradigm works when combined with data from experiments testing which structures repel and bind to one another (potentially in the presence of other structures).
- Genotype to phenotype mapping is a problem where the former has a lot of cheaper data and the latter has little data.
- Cell state trajectory can be represented in low dimensions using cytometry data in silico.