# Preface: How I gave birth to Evaluatology

By Dr. Jianfeng Zhan

This morning, I woke up from a dream where I spent almost two hours writing the preface for this book. I feel very regretful. If I'd woken up earlier, I could have finished this challenging but enjoyable task.

In 2009, I was given the task of writing a technical report on information technology infrastructures for emerging computing, like Internet Services, Cloud Computing, and Big Data. At the time, my boss was Prof. Ninghui Sun, who introduced me to Prof. Kai Li, a well-known professor from Princeton University.

In the 1980s, Prof. Li graduated from the Institute of Computing Technology (ICT) at the Chinese Academy of Sciences, where I was an Assistant Professor since 2002, an Associate Professor since 2004, and later promoted to a Full Professor in 2012.

Kai impressed me in three key ways. First, when discussing innovations, he always starts by calculating the cost using current best practices—I found this truly amazing. In China, many scientists often see cost calculation as boring or even pointless, so his approach stands out.

Second, Kai has created several highly influential benchmark works. One is PARSEC, a well-known CPU benchmark. Another is ImageNet, which he co-developed with Professor Feifei Li. The AI community widely credits ImageNet as one of the key drivers behind the AI boom.

Lastly, Kai is incredibly successful in business. His startup, DataDomain, was acquired by EMC for nearly one billion US dollars, which speaks volumes about his achievements.

I admire Kai's influence. Looking back, I believe my conversation with Kai was the starting point for both Evaluatology and this book. I define Evaluatology as "the science of uncovering the effects of everything." In this book, I use the same methodology to trace the people and events that influenced both the book's creation and me while I was writing it.

During this journey, two things stand out—one was a stroke of luck, and the other was a bit unfortunate.

The first thing I want to share is BigDataBench, our first influential benchmark for Big Data. I worked on it in 2013 with Mr. Lei Wang, who was my Ph.D student at the

time. Surprisingly, we finished this work in just two weeks.

Eventually, our article was accepted by HPCA 2014. I remember clicking the submission button to upload our paper to the HPCA conference while sitting in the flight cabin. My hands were shaking. Five minutes later, my flight took off from San Francisco to Beijing.

We were very lucky—our paper received a high score, and the industry chair of HPCA was incredibly supportive and encouraging. Dr. Zhen Jia, my previous Ph.D. student, later told me that when a famous professor asked Prof. Kai Li to recommend a big data benchmark, Kai recommended our BigDataBench.

At the same time, I submitted another OS-related article to top conferences like ASPLOS, SOSP, and OSDI. But this time, we spent six years without the article being accepted. I still remember one year at ASPLOS, where one reviewer even gave us a "strong accept," which is quite rare in computer systems conferences. In the end, I decided not to publish this article.

The second thing I want to share is AIBench, another benchmark project focused on AI. I worked on it with Miss Wanling Gao, who was also my Ph.D student at the time. She's a very smart and hardworking person.

Wanling submitted the AIBench paper to the HPCA 2020 conference. She hadn't slept for almost three days before she finally clicked the submission button. I felt confident about the paper's chances, and it turned out we received very high scores: four accepts and one weak reject.

However, I noticed something unusual. The reviewer who gave a weak reject mentioned that MLPerf was already enough, and there was no need for AIBench. MLPerf is a collaborative project involving major U.S. companies and universities, making it a competitor to our work.

I decided to write an email to the industry chair, thanking him for the effort and reminding them that there's a competition between AIBench and MLPerf. I also emphasized that having two independent benchmarks would benefit the entire community.

The chair never replied to my email, and I couldn't help but feel that something was happening behind the scenes.

It's no surprise that our article was rejected for a non-technical reason—someone was clearly manipulating the process. I felt furious and complained to the conference organizers and high-level committees. But in the end, the decision still favored the industry chair of HPCA 2020. The chair sent me an email apologizing for making me feel unfairly treated.

I think Dr. Wanling Gao felt very depressed after this. She took several years to recover from the experience. Meanwhile, the MLPerf article got accepted about half a year later. Our article received a high score at Micro 2020 but was still rejected. Later, at PACT 2021, our article was initially deemed rejected. I wrote an email to the chair, a French scientist, and he felt our treatment was unfair. He asked another reviewer to recheck the article, and it was finally accepted.

By that time, MLPerf had become the "superstar" in the field, which made the whole situation even more frustrating.

This is a case study that showcases the toughness of our science and technical society and our guys. When you compare the outcomes (or "fates") of two articles and two young scientists under nearly identical visible conditions, it's like a quasi—experiment, which is one topic of this book.

Many unknown or known but hidden factors dominate the outcomes, which is what we have to face. I'm no longer angry about it. Instead, I've learned to focus on more meaningful work to overcome these struggles. One such work is Evaluatology, a fascinating and thought-provoking field I'm now pursuing.

In 2021, I realized that benchmarks, while widely used across many fields, actually have no rigorous methodology. This is true even for famous benchmarks like ImageNet. I learned this from Prof. Kai Li's open lecture. He mentioned that when he and Dr. Feifei Li (who was a young assistant professor at the time) applied for funding, some reviewers even laughed at their idea. At that time, Kai was already a very senior and well-known professor at Princeton. Because of this, I wrote an article in my launched journal TBench to call for establishing the benchmark sciences and engineering.

But what exactly is a benchmark? My former Ph.D. student, Tang Fei, now working in a famous Chinese company, once told me that when someone asked about his research field, he felt embarrassed to say he was working on benchmarks. It seemed to him like a low-status and uninteresting area of research.

In 2022, I finally understood the link between benchmarks and evaluation. I asked myself a critical question: Why are rigorous methods like Randomized Control Trials (RCT) used to evaluate drugs, while in computer science, people still rely on empirical methods like SPEC CPU?

Everyone loves to report a CPU performance number using SPEC CPU. But my Ph.D. student, Chenxi Wang, my colleagues, Dr. Lei Wang, and Dr. Wanling Gao, proved convincingly that for the same CPU, performance numbers can vary by tens or even hundreds—showing how unreliable this method can be.

I joked: Well, reporting a CPU number won't kill anyone, unlike reporting a drug's performance. But this shouldn't be the case! We're part of a science and engineering community.

In many other areas, like university rankings, the situation is even worse. A University ranking made many young boys and girls, and even their parents, feel unhappy or even depressed. How ironic is that!

In 2023, I came up with the term Evaluatology and wrote an article titled Evaluatology: The Science and Engineering of Evaluation. In the first draft, I used the term Evaluationology, but my Ph.D student consulted a native English speaker, who suggested the shorter and more natural-sounding name Evaluatology. I happily adopted the idea.

I sent the article to many scholars, and Prof. David Lilja responded with a warm and encouraging message. He said the article was very interesting and especially appreciated my evaluation axioms. He also suggested I explore Design of Experiments (DoE) in more depth. I'm truly grateful for his insightful feedback.

I also received positive feedback from several professors, including Prof. Weiping Li

from the Civil Aviation Flight University of China, Prof. Aoying Zhou, Prof. Weining Qian, and Prof. Wei Wang from East China Normal University. Their encouragement meant a lot to me.

In 2024, we organized a workshop on Evaluatology in Guangzhou, where we discussed the science and engineering of evaluation with experts and researchers. It was a great opportunity to share ideas and learn from others.

From 2024 to 2025, I dedicated almost two years to writing this book without taking a single day off. Initially, I planned to work on it alone, but I soon realized how challenging it would be to handle everything by myself. I also understood that leading my colleagues and students to work together would make the process much valuable.

That's why I invited two of my colleagues, Dr. Lei Wang and Dr. Wanling, along with one Ph.D. student, Hongxiao Li, to join me. However, after a month, we fell behind schedule, so I asked Mr. Chenxi Wang and Dr. Fanda Fan (my postdoc) to join the team as well.

Just last month, my postdoc, Dr. Guoxin Kang, developed a strong interest in Evaluatology-based AI and put in a lot of effort. I think it's only fair to invite her to join us, too!

During the two-year process of writing this book, four events stand out as worth mentioning.

One person I know well attended my public presentation about Evaluatology and got inspired by it. He quickly wrote an article and published it in a famous magazine in a short time window. Several ideas in his work were clearly inspired by my talk, and some even directly derived from it, which was published a month earlier.

He felt embarrassed and texted me to explain two things: First, he had mentioned my work without naming me in another article. Second, he extended an invitation for me to author an article for a themed section of the magazine he oversees. I turned down his offer, but I didn't complain to the committees handling this issue. I didn't want to hurt his career.

Throughout my career in science and technology, I've encountered many disappointing situations. For example, I once wrote a technical report, and someone asked me if I had already published it. If I hadn't, he planned to write a book based on my report by himself alone.

I didn't like this behavior at all. That's why I added several footnotes in this book, clearly showing how my ideas were inspired by others' work. I wanted to make it clear that I respect everyone's contributions.

Second, many people don't take evaluation seriously—they think it's a "soft" field. Personally, I don't agree with this view at all. I believe Evaluatology is just as hard as design, and it might even help us create a new AI paradigm.

During a group meeting, Dr. Chunjie Luo made a very convincing point: evaluation and design are actually two sides of the same coin, the so-called dual problem. His presentation was so persuasive and compelling that it really made me think.

Third, during this process, I crossed paths with Mr. Hedong Yan. He had initially planned to join my research group as a Ph.D. student, but we didn't work well together.

After a year, he left. I don't plan to share the details of our differences, but I do want to thank him for one thing—he provided many valuable references for Part IV, even though he didn't contribute directly to the work.

Fourth, one day in 2025, I suddenly thought: Evaluatology could be defined as the science of uncovering the effects of things. I got this idea, inspired by Dr. Judea Pearl and Dana Mackenzie, after reading their book: The Book of Why.

Last year, a well-known professor joked with me, "Why haven't you been fired by ICT, Chinese Academy of Sciences?" I could mention his name, but I won't—out of respect for him.

He explained the reason. Many scientists are busy applying for funding and gaining official recognition, like distinguished young scientists. It seems that I feel no interest in such things. His point is that I should be fired. That is one of the reasons that I am very grateful for the support from ICT.

When I moved from the Advanced Computer Systems Research Center to the Distributed Systems Research Center, both as director, ICT granted me one million yuan in research funding as an unsolicited gift—I never even had to apply. For this kind support, I am profoundly grateful to Prof. Xilin Chen, Director of ICT, and Prof. Ninghui Sun, Academician of the Chinese Academy of Engineering.

I'd like to end by sincerely thanking my family. Looking at it through the lens of Evaluatology, my wife has clearly been the person who has influenced me most over the past two decades.

We first met in October 2001 while hiking on Vigilance Mountain. A year later, we got married without the usual wedding celebrations. We embarked on a honeymoon journey to the enchanting and mystical Jiuzhai Valley, a place renowned for its breathtaking beauty. After that, we enjoy a simple life happily.

My daughter is my beloved treasure. I miss you very much. May you find joy in your life in Boston. Every life is unique, each carrying its own inherent dignity. This dignity is not defined by appearance but resides in the mind.

I also want to thank the many small animals, trees, and flowers I've cared for this past year. Whenever I felt completely drained, spending time with you—just watching and tending to your growth—brought me deep peace and renewal.