RONNY KOHAVI

LENNY'S PODCAST

BILINGUAL TRANSCRIPT

ORIGINAL BY

Lenny Rachitsky

@lennysan • x.com/lennysan

ANALYSIS BY

@Penny777 • x.com/penny777

Generated: 2026-01-16

Ronny Kohavi - 双语对照

Lenny's Podcast: Ronny Kohavi (Bilingual Transcript)

[00:00:00] Ronny Kohavi

English:

I'm very clear that I'm a big fan of test everything, which is any code change that you make, any feature that you introduce has to be in some experiment. Because again, I've observed this sort of surprising result that even small bug fixes, even small changes can sometimes have surprising, unexpected impact.

中文翻译:

我非常明确，我是“测试一切”的坚定支持者。也就是说，你所做的任何代码更改、你引入的任何功能，都必须放在某种实验中。因为我再次观察到了这种令人惊讶的结果：即使是微小的错误修复，甚至是微小的改动，有时也会产生令人惊讶、意想不到的影响。

[00:00:22] Ronny Kohavi

English:

And so I don't think it's possible to experiment too much. You have to allocate sometimes to these high risk, high reward ideas. We're going to try something that's most likely to fail. But if it does win, it's going to be a home run.

中文翻译:

因此，我不认为实验会做得“过多”。你必须有时为这些高风险、高回报的想法分配时间。我们要尝试一些极有可能失败的事情，但如果它真的成功了，那将是一个“全垒打”（巨大的成功）。

[00:00:38] Ronny Kohavi

English:

And you have to be ready to understand and agree that most will fail. And it's amazing how many times I've seen people come up with new designs or a radical new idea. And they believe in it, and that's okay. I'm just cautioning them all the time to say, "If you go for something big, try it out, but be ready to fail 80% of the time."

中文翻译:

你必须准备好去理解并接受大多数实验都会失败的事实。令人惊讶的是，我见过无数次人们提出新设计或激进的新想法，他们对此深信不疑，这没关系。但我一直告诫他们：“如果你想做大的改动，去尝试吧，但要做好

80% 的时间都会失败的准备。”

[00:01:05] Lenny

English:

Welcome to Lenny's Podcast, where I interview world-class product leaders and growth experts to learn from their hard win experiences building and growing today's most successful products.

中文翻译:

欢迎来到 Lenny 的播客。在这里，我会采访世界级的业务负责人和增长专家，学习他们在构建和增长当今最成功产品过程中积累的宝贵经验。

[00:01:14] Lenny

English:

Today my guest is Ronny Kohavi. Ronny is seen by many as the world expert on A/B testing and experimentation. Most recently, he was VP and technical fellow of relevance at Airbnb where he led their search experience team. Prior to that, he was corporate vice president at Microsoft, where he led Microsoft Experimentation Platform team. Before that, he was director of data mining and personalization at Amazon.

中文翻译:

今天的嘉宾是 Ronny Kohavi。Ronny 被许多人视为 A/B 测试和实验科学领域的全球专家。最近，他担任 Airbnb 的副总裁兼相关性技术院士，领导搜索体验团队。在此之前，他是微软的公司副总裁，领导微软实验平台团队。更早之前，他是亚马逊数据挖掘和个性化部门的总监。

[00:01:38] Lenny

English:

He's currently a full-time advisor and instructor. He's also the author of the go-to book on experimentation called Trustworthy Online Controlled Experiments. And in our show notes, you'll find a code to get a discount on taking his live cohort-based course on Maven.

中文翻译:

他目前是一名全职顾问和讲师。他还是实验科学领域的权威著作《值得信赖的在线对照实验》（Trustworthy Online Controlled Experiments）的作者。在我们的节目介绍中，你可以找到一个优惠码，用于参加他在 Maven 平台上的直播训练营课程。

[00:01:53] Lenny

English:

In our conversation, we get super tactical about A/B testing. Ronny shares his advice for when you should start considering running experiments at your company, how to change your company's culture to be more experiment driven, what are signs your experiments are potentially invalid, why trust is the most important element of a successful experiment, culture, and platform. How to get started if you want to start running experiments at your company. He also explains what actually is a P value and something

called Twyman's law, plus some hot takes about Airbnb and experiments in general. This episode is for anyone who's interested in either creating an experiment driven culture at their company or just fine-tuning one that already exists. Enjoy this episode with Ronny Kohavi after a short word from our sponsors.

中文翻译:

在我们的对话中，我们深入探讨了 A/B 测试的实操战术。Ronny 分享了关于何时应该开始在公司运行实验、如何将公司文化转变为实验驱动型、实验可能失效的迹象、以及为什么“信任”是成功的实验文化和平台中最重要的元素的建议。他还介绍了如果你想在公司开始运行实验该如何起步。此外，他还解释了 P 值（P value）的本质以及所谓的“特怀曼法则”（Twyman's law），并分享了一些关于 Airbnb 和实验科学的犀利观点。本期节目适合任何想要在公司建立实验驱动文化，或优化现有文化的人。在听完赞助商的简短介绍后，请欣赏与 Ronny Kohavi 的精彩对话。

[00:02:39] Lenny

English:

This episode is brought to you by Mixpanel. Get deep insights into what your users are doing at every stage of the funnel, at a fair price that scales at you grow. Mixpanel gives you quick answers about your users from awareness, to acquisition, through retention. And by capturing website activity, ad data, and multi-touch attribution right in Mixpanel, you can improve every aspect of the full user funnel. Powered by first party behavioral data instead of third party cookies, Mixpanel is built to be more powerful and easier to use than Google Analytics. Explore plans for teams of every size and see what Mixpanel can do for you at mixpanel.com/friends/lenny. And while you're at it, they're also hiring. So check it out at mixpanel.com/friends/lenny.

中文翻译:

本期节目由 Mixpanel 赞助。以随业务增长而扩展的公平价格，深入洞察用户在漏斗每个阶段的行为。Mixpanel 为你提供从认知、获取到留存的用户快速解答。通过在 Mixpanel 中直接捕获网站活动、广告数据和多触点归因，你可以改进整个用户漏斗的各个方面。Mixpanel 由第一方行为数据而非第三方 Cookie 驱动，旨在比 Google Analytics 更强大且更易于使用。前往 mixpanel.com/friends/lenny 探索适合各种规模团队的方案。顺便提一下，他们也在招聘，请访问该链接查看详情。

[00:03:27] Lenny

English:

This episode is brought to you by Round. Round is the private network built by tech leaders for tech leaders. Round combines the best of coaching, learning, and authentic relationships to help you identify where you want to go and accelerate your path to get there, which is why their wait list tops thousands of tech execs. Round is on a mission to shape the future of technology and its impact on society. Leading in tech is uniquely challenging, and doing it well is easiest when surrounded by leaders who understand your day-to-day experiences. When we're meeting and building relationships with the right people, we're more likely to learn, find new opportunities, be dynamic in our thinking, and achieve our goals. Building and managing your network doesn't have to feel like networking. Join Round to surround yourself with leaders from tech's most innovative companies. Build relationships, be inspired, take action. Visit round.tech/apply and use promo code Lenny to skip the wait list. That's round.tech/apply.

中文翻译:

本期节目由 Round 赞助。Round 是一个由技术领袖为技术领袖建立的私人网络。Round 结合了顶尖的教练辅导、学习和真实的社交关系，帮助你明确目标并加速实现路径，这就是为什么他们的候补名单上有数千名技术高管。Round 的使命是塑造技术的未来及其对社会的影响。在技术领域担任领导者极具挑战，而当你身边环绕着理解你日常经历的领导者时，你会做得更出色。当我们与对的人建立联系时，我们更有可能学习、发现新机会、保持思维活跃并实现目标。建立和管理你的网络不一定非得像在“社交”。加入 Round，与来自最具创新精神的技术公司的领导者为伍。建立关系，获取灵感，采取行动。访问 round.tech/apply 并使用促销代码 Lenny 跳过候补名单。

[00:04:30] Lenny

English:

Ronny, welcome to the podcast.

中文翻译:

Ronny，欢迎来到播客。

[00:04:33] Ronny Kohavi

English:

Thank you for having me.

中文翻译:

谢谢你的邀请。

[00:04:34] Lenny

English:

So you're known by many as maybe the leading expert on A/B testing and experimentation, which I think is something every product company eventually ends up trying to do, often badly. And so I'm excited to dig quite deep into the world of experimentation and A/B testing to help people run better experiments. So thank you again for being here.

中文翻译:

许多人认为你是 A/B 测试和实验科学领域的顶级专家。我认为每家产品公司最终都会尝试做这件事，但通常做得并不好。所以我很高兴能深入探讨实验和 A/B 测试的世界，帮助大家运行更好的实验。再次感谢你的到来。

[00:04:54] Ronny Kohavi

English:

That's a great goal. Thank you.

中文翻译:

这是一个伟大的目标。谢谢。

[00:04:56] Lenny

English:

Let me start with kind of a fun question. What is maybe the most unexpected A/B tests you've run or maybe the most surprising result from an A/B test that you've run?

中文翻译:

让我从一个有趣的问题开始。你运行过的最出乎意料的 A/B 测试是什么？或者说，你运行过的 A/B 测试中，最令人惊讶的结果是什么？

[00:05:06] Ronny Kohavi

English:

So I think the opening example that I use in my book and in my class is the most surprising public example we can talk about. And this was kind of an interesting experiment. Somebody proposed to change the way that ads were displayed on Bing, the search engine. And he basically said, "Let's take the second line and move it, promote it to the first line so that the title line becomes larger."

中文翻译:

我想我在书里和课堂上用的开篇案例，就是我们可以谈论的最令人惊讶的公开案例。这是一个非常有趣的实验。有人提议改变 Bing 搜索引擎上广告的显示方式。他基本上是说：“让我们把第二行挪到第一行，提升它的位置，这样标题行就会变得更大。”

[00:05:37] Ronny Kohavi

English:

And when you think about that, and if you're going to look in my book, or in the class, there's an actual diagram of what happened, the screenshots. But if you think about it, just realistically it looks like a meh idea. Why would this be such a reasonable, interesting thing to do? And indeed, when we went back to the backlog, it was on the backlog for months, and languished there, and many things were rated higher.

中文翻译:

当你思考这个想法时——如果你看我的书或课程，会有实际发生的图表和截图——但从现实角度看，这听起来像个平庸的主意。为什么这会是一件合理且有趣的事情呢？事实上，当我们查看待办事项列表（backlog）时，它已经在那里躺了好几个月，无人问津，许多其他事情的优先级都比它高。

[00:06:05] Ronny Kohavi

English:

But the point about this is it's trivial to implement. So if you think about return on investment, we could get the data by having some engineers spend a couple of hours implementing it.

中文翻译:

但关键在于，它的实现非常简单。所以如果你考虑投资回报率（ROI），我们只需要让几名工程师花几个小时来实现它，就能获得数据。

[00:06:19] Ronny Kohavi

English:

And that's exactly what happened. Somebody at Bing who kept seeing this in the backlog and said, "My God, we're spending too much time discussing it. I could just implement it." He did. He spent a couple of days implementing it, as is the common thing at Bing, he launched the experiment.

中文翻译:

事实也确实如此。Bing 的某个人一直看到这个待办项，然后说：“天哪，我们讨论它花了太多时间，我直接把它做出来得了。”他确实这么做了。他花了两天时间实现它，然后按照 Bing 的惯例启动了实验。

[00:06:37] Ronny Kohavi

English:

And a funny thing happened. We had an alarm. Big escalation, something is wrong with the revenue metric. Now this alarm fired several times in the past when there were real mistakes, where somebody would log revenue twice, or there's some data problem. But in this case, there was no bug. That simple idea increased revenue by about 12%.

中文翻译:

有趣的事情发生了。我们收到了警报。严重的升级提醒，收入指标出问题了。以前这种警报响过几次，通常是因为真的出错了，比如有人重复记录了收入，或者存在数据问题。但在这种情况下，没有 Bug。那个简单的想法让收入增加了约 12%。

[00:07:01] Ronny Kohavi

English:

And this is something that just doesn't happen. We can talk later about Wyman's law, but that was the first reaction, which is, "This is too good to be true. Let's find a bug." And we did. And we looked for several times, and we replicated the experiment several times, and there was nothing wrong with it. This thing was worth $100 million at the time when Bing was a lot smaller.

中文翻译:

这种情况简直闻所未闻。我们稍后可以谈谈特怀曼法则（Twyman's law），但当时的第一反应是：“这好得让人难以置信，肯定有 Bug。”于是我们去找 Bug。我们找了好几次，重复实验了好几次，结果发现完全没问题。在 Bing 规模还比较小的时候，这个改动价值 1 亿美元。

[00:07:22] Ronny Kohavi

English:

And the key thing is it didn't hurt the user metrics. So it's very easy to increase revenue by doing theatrics. Displaying more ads is a trivial way to raise revenue, but it hurts the user experience. And we've done the experiments to show that. In this case, this was just a home run that improved revenue, didn't significantly hurt the guardrail metrics. And so we were in awe of what a trivial change. That was the biggest revenue impact to Bing in all its history.

中文翻译:

关键在于，它没有损害用户指标。通过耍花招来增加收入很容易，比如多展示广告就是一种增加收入的简单方法，但它会损害用户体验。我们做过实验证明了这一点。但在这种情况下，这简直是一个“全垒打”，它增加

了收入，却没有显著损害护栏指标（guardrail metrics）。我们对这样一个微小的改动感到敬畏。那是 Bing 历史上对收入影响最大的一次改动。

[00:07:57] Lenny

English:

And that was basically shifting in two lines, right? Switching two lines in the search results.

中文翻译:

那基本上就是调换了两行，对吧？在搜索结果中调换了两行。

[00:08:02] Ronny Kohavi

English:

And this was just moving the second line to the first line. Now you then go and run a lot of experiments to understand what happened here. Is it the fact that the title line has a bigger font, sometimes different color? So we ran a whole bunch of experiments.

中文翻译:

是的，只是把第二行移到了第一行。然后你会去运行大量的实验来理解这里到底发生了什么。是因为标题行的字体更大了，还是有时颜色不同了？所以我们运行了一大堆实验。

[00:08:16] Ronny Kohavi

English:

And this is what usually happens. We have a breakthrough. You start to understand more about, what can we do? And there suddenly a shift towards, "Okay, what are other things we could do that would allow us to improve revenue?" We came up with a lot of follow on ideas that helped a lot.

中文翻译:

这就是通常会发生的情况。我们有了一个突破，你开始更深入地理解：我们还能做什么？然后突然间，重心转向了：“好吧，还有哪些事情可以让我们增加收入？”我们提出了很多后续想法，非常有帮助。

[00:08:34] Ronny Kohavi

English:

But to me, this was an example of a tiny change that was the best revenue generating idea in Bing's history, and we didn't rate it properly. Nobody gave this the priority that in hindsight, it deserves. And that's something that happens often. I mean, we are often humbled by how bad we are at predicting the outcome of experiments.

中文翻译:

但对我来说，这是一个微小改动成为 Bing 历史上最佳创收想法的例子，而我们当时并没有正确评估它。事后看来，没有人给它应有的优先级。这种情况经常发生。我的意思是，在预测实验结果方面，我们经常因为自己的无能而感到谦卑。

[00:09:01] Lenny

English:

This reminds me of a classic experiment at Airbnb while I was there, and we'll talk about Airbnb in a bit. The search team just ran a small experiment of what if we were to open a new tab every time someone clicked on a search result, instead of just going straight to that listing. And that was one of the biggest wins in search-

中文翻译:

这让我想起我在 Airbnb 时的一个经典实验，稍后我们会聊到 Airbnb。搜索团队做了一个小实验：如果用户每次点击搜索结果时都打开一个新标签页，而不是直接跳转到该房源页面，会怎么样？那是搜索领域最大的胜利之一——

[00:09:18] Ronny Kohavi

English:

And by the way, I don't know if you know the history of this, but I tell about this in class. We did this experiment way back around 2008 I think. And so this predates Airbnb. I remember it was heavily debated. Why would you open something in a new tab? The users didn't ask for it. It was a lot of pushback from the designers. And we ran that experiment. And again, it was one of these highly surprising results that made it that we learned so much from it.

中文翻译:

顺便说一下，我不知道你是否了解这段历史，但我会在课堂上讲。我们在 2008 年左右就做过这个实验。所以这比 Airbnb 还要早。我记得当时争议很大。为什么要用新标签页打开？用户又没要求这么做。设计师们也强烈反对。但我们运行了那个实验。同样，那也是一个非常令人惊讶的结果，让我们学到了很多。

[00:09:49] Ronny Kohavi

English:

So we first did this. It was done in the UK for opening Hotmail, and then we moved it to MSN, so it would open search in new tab, and all the set of experiments were highly, highly beneficial. We published this. And I have to tell you, when I came to Airbnb, I talked to our joint friend Ricardo about this. And it was sort of done, it was very beneficial, and then it was semi forgotten, which is one of the things you learned about institutional memories. When you have winners, make sure to address them and remember them. So it was at Airbnb done for a long time before I joined that listings opened in a new tab, but other things that were designed in the future were not done. And I reintroduced this to the team, and we saw big improvements.

中文翻译:

我们最开始是在英国针对打开 Hotmail 做的这个实验，然后推广到了 MSN，让搜索结果在新标签页打开，整套实验都非常有益。我们发表了相关论文。我必须告诉你，当我来到 Airbnb 时，我和我们共同的朋友 Ricardo 聊过这件事。当时这个功能已经做了一部分，效果很好，但后来被半遗忘了，这就是所谓的“机构记忆”（institutional memory）问题。当你有了成功的案例，一定要记录并记住它们。在我加入 Airbnb 之前，房源详情页已经很久都是在新标签页打开了，但后来设计的一些新功能却没有遵循这一点。我重新向团队推荐了这个做法，我们看到了巨大的提升。

[00:10:35] Lenny

English:

Shout out to Ricardo, our mutual friend who helped make this conversation happen. There's this holy grail of experiments that I think people are always looking for of one hour of work and it creates this massive result. I imagine this is very rare, and don't expect this to happen. I guess in your experience, how often do you find one of these gold nuggets just lying around?

中文翻译:

向 Ricardo 致敬，他是我们共同的朋友，促成了这次对话。人们总是在寻找实验的“圣杯”：花一小时工作，产生巨大的结果。我想这非常罕见，不应该指望它发生。在你的经验中，这种“金矿”出现的频率有多高?

[00:10:57] Ronny Kohavi

English:

Yeah. So again, this is a topic that's near and dear to my heart. Everybody wants these amazing results, and I show them in chapter one in my book, multiple of these small efforts, huge gain.

中文翻译:

是的。这又是一个我非常关心的话题。每个人都想要这些惊人的结果，我在书的第一章展示了多个这种“小投入、大回报”的案例。

[00:11:13] Ronny Kohavi

English:

But as you said, they're very rare. I think most of the time, the winnings are made this inch by inch. And there's a graph that I show in my book, a real graph of how Bing ads has managed to improve the revenue per a thousand searches over time, and every month you can see a small improvement and a small improvement. Sometimes the degradation because of legal reasons or other things. There were some concern that we were not marking the ads properly. So you have to suddenly do something that you know is going to hurt revenue. But yes, I think most results are inch by inch. You improve small amounts, lots of them. I think that the best example that I can say is a couple of them that I can speak about.

中文翻译:

但正如你所说，它们非常罕见。我认为大多数时候，胜利是“寸进”的。我在书里展示了一张真实的图表，显示了 Bing 广告如何随着时间的推移提高每千次搜索的收入，每个月你都能看到一点点进步。有时会因为法律原因或其他事情出现倒退，比如有人担心我们没有正确标记广告，所以你必须突然做一些明知会损害收入的事情。但总的来说，大多数结果都是一点一滴积累的。你改进了一点点，但改进了很多处。我可以举几个我可以公开谈论的例子。

[00:12:00] Ronny Kohavi

English:

One is at Bing, the relevance team, hundreds of people all working to improve bing relevance. They have a metric, we'll talk about OEC, the overall evaluation criterion. But they have a metric that their goal is to improve it by 2% every year. It's a small amount, and that 2% you can see here's a 0.1, and here's a 0.15, here's a 0.2, and then they add up to around 2% every year, which is amazing.

中文翻译:

一个是 Bing 的相关性团队，数百人都在努力提高 Bing 的相关性。他们有一个指标——我们稍后会谈到 OEC（综合评价指标）——他们的目标是每年将该指标提高 2%。这是一个很小的数字，这 2% 是由 0.1%、0.15%、0.2% 这样一点点累加起来的，每年能达到 2% 已经非常了不起了。

[00:12:28] Ronny Kohavi

English:

Another example that I am allowed to speak about from Airbnb is the fact that we ran some 250 experiments in my tenure there in search relevance. And again, small improvements added up. So this became overall a 6% improvement to revenue. So when you think about 6%, it's a big number, but it came out not of one idea, but many, many smaller ideas that each gave you a small gain.

中文翻译:

另一个我可以谈论的 Airbnb 例子是，在我任职期间，我们在搜索相关性方面运行了大约 250 个实验。同样，微小的改进累积了起来，最终使收入整体提升了 6%。6% 是一个很大的数字，但它不是来自一个想法，而是来自许多许多微小的想法，每个想法都带来了一点点收益。

[00:13:00] Ronny Kohavi

English:

And in fact, again, there's another number I'm allowed to say. Of these experiments, 92% failed to improve the metric that we were trying to move. So only 8% of our ideas actually were successful at moving the key metrics.

中文翻译:

事实上，还有一个我可以公开的数字：在这些实验中，92% 的实验未能改善我们试图提升的指标。也就是说，只有 8% 的想法真正成功地推动了关键指标。

[00:13:17] Lenny

English:

There's so many threads I want to follow here, but let me follow this one right here. You just mentioned of 92% of experiments failed. Is that typical in your experience seeing experiments running a lot of companies? What should people expect when they're running experiments? What percentage should they expect to fail?

中文翻译:

这里有很多我想深入探讨的话题，但让我先接着这个说。你刚才提到 92% 的实验都失败了。根据你在许多公司观察实验的经验，这具有代表性吗？人们在运行实验时应该有什么样的预期？他们应该预期百分之多少的实验会失败？

[00:13:31] Ronny Kohavi

English:

Well, first of all, I published three different numbers for my career. So overall at Microsoft, about 66%, two thirds of ideas fail. And don't think the 66 is accurate. It's about two thirds. At Bing, which is a much more optimized domain after we've been optimizing it for a while, the failure rate was around 85%. So it's harder to improve something that you've been optimizing for a while. And then at Airbnb, this 92% number is the highest failure rate that I've observed.

中文翻译:

首先，我公布过我职业生涯中的三个不同数字。在微软，总体而言大约 66%（三分之二）的想法会失败。66% 只是个概数，大约就是三分之二。在 Bing，由于我们已经优化了很长时间，这是一个高度优化的领域，失败率约为 85%。优化已经很成熟的东西会更难。而在 Airbnb，92% 是我观察到的最高失败率。

[00:14:09] Ronny Kohavi

English:

Now I've quoted other sources. It's not that I worked at groups that were particularly bad, Booking, Google Ads, other companies published numbers that are around 80 to 90% failure rate of ideas. This is where it's important of experiments. It's important to realize that when you have a platform, it's easy to get this number. You look at how many experiments were run and how many of them launched. Not every experiment maps to an idea.

中文翻译:

我也引用过其他来源。并不是说我工作的团队特别差，Booking、Google Ads 等其他公司公布的数字也都在 80% 到 90% 的失败率左右。这就是实验的重要性。要意识到，当你有一个平台时，很容易得到这个数字：你看一下运行了多少实验，其中有多少最终发布了。并不是每个实验都对应一个想法。

[00:14:39] Ronny Kohavi

English:

So it's possible that when you have an idea, your first implementation, you start an experiment. Boom, it's egregiously bad, because you have a bug. In fact, 10% of experiments tend to be aborted on the first date. Those are usually not that the idea is bad, but that there is an implementation issue or something we haven't thought about, that forces an abort.

中文翻译:

有可能当你有一个想法时，你的第一次实现并开始实验，结果极其糟糕，因为有 Bug。事实上，10% 的实验往往在第一天就被终止了。这些通常不是因为想法不好，而是因为实现有问题，或者有些我们没想到的因素迫使实验终止。

[00:15:01] Ronny Kohavi

English:

You may iterate and pivot again. And ultimately, if you do two, or three, or four pivots or bug fixes, you may get to a successful launch. But those numbers of 80 to 92% failure rate are of experiments.

中文翻译:

你可能会迭代并再次调整。最终，如果你进行了两三次或四次调整或 Bug 修复，你可能会成功发布。但那 80% 到 92% 的失败率是指实验的失败率。

[00:15:17] Ronny Kohavi

English:

Very humbling. I know that every group that starts to run experiments, they always start off by thinking that somehow, they're different. And their success rate's going to be much, much higher, and they're all humbled.

中文翻译:

这非常让人清醒。我知道每个开始运行实验的团队，最初总觉得自己与众不同，觉得自己的成功率会高得多，但最后他们都会感到谦卑。

[00:15:29] Lenny

English:

You mentioned that you had this pattern of clicking a link and opening a new tab as a thing that just worked at a lot of different places.

中文翻译:

你提到“点击链接并在新标签页打开”这种模式在很多不同的地方都奏效。

[00:15:36] Ronny Kohavi

English:

Yeah.

中文翻译:

是的。

[00:15:37] Lenny

English:

Are there other versions of this? Do you do you collect a list of, "Here's things that often work when we want to move" there's some you could share. I don't know if you have a list in your head.

中文翻译:

还有其他类似的案例吗？你会收集一个清单吗，比如“当我们想要推动某项指标时，这些做法通常有效”？有没有一些可以分享的？我不知道你脑子里是否有一个清单。

[00:15:48] Ronny Kohavi

English:

I can give you two resources. One of them is a paper that we wrote called Rules of Thumb, and what we tried to do at that time at Microsoft was to just look at thousands of experiments that run and extract some patterns. And so that's one paper that we can then put in the notes.

中文翻译:

我可以给你两个资源。其中一个是我们在微软时写的一篇论文，叫《经验法则》（Rules of Thumb），当时我们尝试观察数千个运行的实验并提取一些模式。我们可以把这篇论文放在节目介绍里。

[00:16:07] Lenny

English:

Perfect.

中文翻译:

太好了。

[00:16:08] Ronny Kohavi

English:

But there's another more accurate, I would say, resource that's useful that I recommend to people. And it's a site called goodui.org, and goodui.org is exactly the site that tries to do what you're saying at scale.

中文翻译:

但还有一个更准确、更有用的资源，我推荐给大家。那是一个叫 goodui.org 的网站，它正是尝试大规模地做你刚才说的那件事。

[00:16:25] Ronny Kohavi

English:

So guy's name is Jacob [inaudible 00:16:28]. He asks people to send them results of experiments, and he puts them into patterns. There's probably 140 patterns I think at this point. And then for each pattern he says, "Well, who has that helped? How many times and by how much?" So you have an idea of this worked, three out of five times. And it was a huge win. In fact, you can find that open a new window in there.

中文翻译:

创办人叫 Jacob。他让人们把实验结果发给他，然后他把这些结果归纳成模式。目前大概有 140 种模式。对于每种模式，他会说明：“这帮助了谁？成功了多少次？提升了多少？”所以你会知道这个做法在五次中有三次奏效，而且是大获全胜。事实上，你可以在那里找到“在新窗口打开”这个模式。

[00:16:54] Lenny

English:

I feel like you feed that into ChatGPT, and you have basically a product manager creating a roadmap tool.

中文翻译:

我觉得如果你把这些输入 ChatGPT，基本上就得到了一个能自动创建路线图的“产品经理工具”。

[00:17:01] Ronny Kohavi

English:

In general, by the way, a lot of that is institutional memory, which is can you document things well enough so that the organization remembers the successes and failures, and learns from them?

中文翻译:

顺便说一下，总的来说，这很大程度上取决于机构记忆，即你是否能把事情记录得足够好，以便组织能够记住成功和失败，并从中学习？

[00:17:17] Ronny Kohavi

English:

I think one of the mistakes that some company makes is they launch a lot of experiments and never go back and summarize the learnings. So I've actually put a lot of effort in this idea of institutional learning, of doing the quarterly meeting of the most surprising experiments.

中文翻译:

我认为一些公司犯的错误之一是，他们启动了很多实验，却从未回头总结经验教训。所以我实际上在“机构学习”这个理念上投入了很多精力，比如举行“最令人惊讶实验”的季度会议。

[00:17:32] Ronny Kohavi

English:

By the way, surprising is another question that people often are not clear about. What is a surprising experiment? To me, a surprising experiment is one where the estimated result beforehand and the actual result differ by a lot. So that absolute value of the difference is large.

中文翻译:

顺便说一下，“令人惊讶”是另一个人们经常不清楚的问题。什么是令人惊讶的实验？对我来说，令人惊讶的实验是指预估结果与实际结果相差甚远的实验。也就是说，两者之差的绝对值很大。

[00:17:53] Ronny Kohavi

English:

Now you can expect something to be great and it's flat. Well, you learn something. But if you expect something to be small and it turns out to be great, like that ad title promotion, then you've learned a lot. Or conversely, if you expect that something will be small and it's very negative, you can learn a lot by understanding why this was so negative. And that's interesting.

中文翻译:

如果你预期某个改动会很棒，结果却毫无波动，那你学到了一些东西。但如果你预期影响很小，结果却非常巨大（比如那个广告标题提升），那你学到了很多。反之，如果你预期影响很小，结果却非常负面，你也可以通过理解为什么这么负面来学到很多。这很有趣。

[00:18:17] Ronny Kohavi

English:

So we focused not just on the winners, but also surprising losers, things that people thought would be a no-brainer to run. And then for some reason, it was very negative. And sometimes, it's that negative that gives you insight. Actually, I'm just coming up with one example that of that, that I should mention.

中文翻译:

所以我们不仅关注赢家，也关注“令人惊讶的输家”——那些人们认为理所当然会成功的改动，结果却非常负面。有时，正是这种负面结果能给你带来洞察。实际上，我刚想到一个例子，应该提一下。

[00:18:36] Ronny Kohavi

English:

We were running this experiment at Microsoft to improve the windows indexer, and the team was able to show on offline tests that it does much better at indexing, and they showed some relevance is higher, and all these good things. And then they ran it as an experiment. You know what happened? Surprising result. Indexing the relevance was actually high, but it killed a battery life.

中文翻译:

我们在微软运行过一个改进 Windows 索引器的实验。团队在离线测试中证明了它的索引效果好得多，相关性更高，总之各种优点。然后他们把它作为一个实验来运行。你知道发生了什么吗？令人惊讶的结果：索引相关性确实很高，但它耗尽了电池寿命。

[00:19:03] Ronny Kohavi

English:

So here's something that comes from left field that you didn't expect. It was consuming a lot more CPU on laptops. It was killing the laptops. And therefore, okay, we learned something. Let's document it. Let's remember this, so that we now take this other factor into account as we design the next iteration.

中文翻译:

这就是一个你完全没想到的意外情况。它在笔记本电脑上消耗了更多的 CPU，简直是在“谋杀”电脑。因此，好吧，我们学到了一些东西。让我们把它记录下来，记住这一点，这样我们在设计下一次迭代时就会考虑到这个因素。

[00:19:23] Lenny

English:

What advice do you have for people to actually remember these surprises? You said that a lot of it is institutional. What do you recommend people do so that they can actually remember this when people leave, say three years later?

中文翻译:

你对人们如何真正记住这些“惊喜”有什么建议？你说这很大程度上是机构性的。你建议人们怎么做，才能在人员离职（比如三年后）时，依然能记住这些教训？

[00:19:34] Ronny Kohavi

English:

Document it. We had a large deck internally of these successes and failures, and we encourage people to look at them. The other thing that's very beneficial is just to have your whole history of experiments and do some ability to search by keywords.

中文翻译:

记录下来。我们内部有一份关于这些成功和失败案例的大型幻灯片，我们鼓励大家去查看。另一件非常有益的事情是，保留完整的实验历史，并具备按关键词搜索的能力。

[00:19:52] Ronny Kohavi

English:

So I have an idea. Type a few keywords and see if from the thousands of experiments that ran... And by the way, these are very reasonable numbers. At Microsoft, just to let you know, when I left in 2019, we were on a rate of about 20 to 25,000 experiments every year. So every working, day we were starting something like 100 new treatments. Big numbers. So when you're running in a group like Bing, which is running thousands and thousands of experiments, you want to be able to ask, "Has anybody did an experiment on this or this or this?" And so that searching capability is in the platform.

中文翻译:

比如我有一个想法，输入几个关键词，看看在成千上万个运行过的实验中……顺便说一下，这些数字很真实。在微软，我 2019 年离开时，我们每年的实验量大约是 2 万到 2.5 万个。也就是说，每个工作日我们都会启动大约 100 个新的实验组。规模很大。所以当你在像 Bing 这样运行着成千上万个实验的团队中时，你会想问：“有没有人做过关于这个或那个的实验？”这种搜索能力就集成在平台中。

[00:20:32] Ronny Kohavi

English:

But more than that, I think just doing the quarterly meeting of the most successful... Most interesting, sorry, not just successful, most interesting experiments is very key. And that also helps the flywheel of experimentation.

中文翻译:

但除此之外，我认为定期举行季度会议，分享最成功的……抱歉，是最有趣的，不只是成功的，分享最有趣的实验是非常关键的。这也有助于推动实验的飞轮。

[00:20:45] Lenny

English:

It's a good segue to something I wanted to touch on, which is there's often, I guess a weariness of running too many experiments and being too data-driven, and the sense that experimentation just leads you to these micro optimizations, and you don't really innovate and do big things. What's your perspective on that? And then, can you be too experiment driven in your experience?

中文翻译:

这正好引出了我想谈的一个话题：人们往往会对运行太多实验、过于数据驱动感到厌倦，觉得实验只会导致微小的优化，而无法真正创新或做大事。你对此怎么看？根据你的经验，会存在“过度实验驱动”的情况吗？

[00:21:07] Ronny Kohavi

English:

I'm very clear that I'm a big fan of test everything, which is any code change that you make, any feature that you introduce has to be in some experiment. Because again, I've observed this surprising result that even small bug fixes, even small changes can sometimes have surprising unexpected impact.

中文翻译:

我非常明确，我是“测试一切”的坚定支持者。也就是说，你所做的任何代码更改、引入的任何功能，都必须放在实验中。因为我再次观察到了这种令人惊讶的结果：即使是微小的错误修复，甚至是微小的改动，有时也会产生令人惊讶、意想不到的影响。

[00:21:30] Ronny Kohavi

English:

And so I don't think it's possible to experiment too much. I think it is possible to focus on incremental changes because some people say, "Well, if we only tested 17 things around this," you have to think about, it's like in stock. You need a portfolio. You need some experiments that are incremental that move you in the direction that you know you're going to be successful over time if you just try enough. But some experiments, you have to allocate sometimes to these high risk, high reward ideas. We're going to try something that's most likely to fail, but if it does win, it's going to be a home run.

中文翻译:

所以我不认为实验会做得“过多”。但我认为确实存在“只关注渐进式改动”的可能。有些人会说：“如果我们只测试了 17 个相关的微小改动……”你必须像对待股票一样思考：你需要一个投资组合。你需要一些渐进式的实验，如果你尝试得足够多，随着时间的推移，它们会把你推向成功的方向。但你也必须为那些高风险、高回报的想法分配时间。我们要尝试一些极有可能失败的事情，但如果它赢了，那就是一个“全垒打”。

[00:22:14] Ronny Kohavi

English:

And so you have to allocate some efforts to that, and you have to be ready to understand and agree that most will fail. And I've amazing how many times I've seen people come up with new designs, or a radical new idea, and they believe in it, and that's okay. I'm just cautioning them all the time to say, "Hey, if you go for something big, try it out, but be ready to fail 80% of the time."

中文翻译:

所以你必须为此分配一些精力，并且必须准备好理解并接受大多数实验都会失败。令我惊讶的是，我见过无数次人们提出新设计或激进的新想法，他们深信不疑，这没关系。但我一直告诫他们：“嘿，如果你想做大的改动，去尝试吧，但要做好 80% 的时间都会失败的准备。”

[00:22:42] Ronny Kohavi

English:

And one true example, that again, I'm able to talk about because we put it in my book, is we were at Bing trying to change the landscape of search. And one of the ideas, the big ideas was we are going to integrate

with social. So we hooked into the Twitter fire hose feed and we hooked into Facebook, and we spent 100 person years on this idea.

中文翻译:

一个真实的例子——我能谈论它是因为我们把它写进了书里——我们在 Bing 尝试改变搜索的格局。其中一个大想法是我们要整合社交媒体。于是我们接入了 Twitter 的实时数据流（fire hose），接入了 Facebook，在这个想法上投入了 100 人/年的精力。

[00:23:14] Ronny Kohavi

English:

And it failed. You don't see it anymore. It existed for about a year and a half, and all the experiments were just negative to flat. And it was an attempt. It was fair to try it. I think it took us a little long to fail, to decide that this is a failure. But at least we had the data. We had hundreds of experiments that we tried. None of them were a breakthrough. And I remember mailing Qi Lu with some statistics showing that it's time to abort, it's time to fail on this. And he decided to continue more. And it's a million dollar question. Do you continue, and then maybe the breakthrough will come next month, or do you abort? And a few months later, we aborted.

中文翻译:

结果失败了。你现在再也看不到它了。它存在了大约一年半，所有的实验结果要么是负面的，要么是毫无波动的。这是一种尝试，尝试它是公平的。我认为我们承认失败、决定放弃的时间花得有点长了。但至少我们有数据。我们尝试了数百个实验，没有一个是突破性的。我记得给陆奇（Qi Lu）发邮件，用统计数据说明是时候终止了，是时候承认失败了。但他决定再继续一段时间。这是一个价值百万美元的问题：你是继续等待下个月可能出现的突破，还是现在终止？几个月后，我们终止了。

[00:24:07] Lenny

English:

That reminds me of at Netflix, they tried a social component that also failed. At Airbnb, early on there was a big social attempt to make, "Here's your friends that have stayed at these Airbnbs," completely had no impact. So maybe that's one of these learnings that we should document.

中文翻译:

这让我想起 Netflix 也尝试过社交组件，也失败了。在 Airbnb 早期，也有一个巨大的社交尝试，比如“这是住过这些房源的朋友”，结果完全没有影响。所以这也许就是我们应该记录下来的教训之一。

[00:24:21] Ronny Kohavi

English:

Yeah, this is hard. This is hard. But again, that's the value of experiments, which are this oracle that gives you the data. You may be excited about things. You may believe it's a good idea. But ultimately, the oracle is the controlled experiment. It tells you whether users are actually benefiting from it, whether you and the users, the company and the users.

中文翻译:

是的，这很难。但这就是实验的价值，它们就像是给你提供数据的“神谕”。你可能对某些事情感到兴奋，可能认为它是个好主意。但最终，受控实验才是神谕。它会告诉你用户是否真正从中受益，以及公司和用户是否双赢。

[00:24:48] Lenny

English:

There's obviously a bit of overhead and downside to running an experiment, setting all up, and analyzing the results. Is there anything that you ever don't think is worth A/B testing?

中文翻译:

运行实验、设置和分析结果显然会有一些开销和负面影响。有没有什么事情是你认为不值得进行 A/B 测试的?

[00:24:59] Ronny Kohavi

English:

First of all, there are some necessary ingredients to A/B testing. And I'll just say outright, not every domain is amenable to A/B testing. You can't A/B test mergers and acquisitions. It's something that happens once, you either acquire or you don't acquire.

中文翻译:

首先，A/B 测试有一些必要的先决条件。我直截了当地说，并非每个领域都适合 A/B 测试。你无法对并购（M&A）进行 A/B 测试。那是发生一次的事情，你要么收购，要么不收购。

[00:25:14] Ronny Kohavi

English:

So you do have to have some necessary ingredient. You need to have enough units, mostly users, in order for the statistics to work out. So if you're too small, it may be too early to A/B test. But what I find is that in software, it is so easy to run A/B testing and it is so easy to build a platform.

中文翻译:

所以你确实需要一些必要的条件。你需要有足够的样本单位（通常是用户），统计学才能奏效。所以如果你规模太小，进行 A/B 测试可能还为时过早。但我发现，在软件领域，运行 A/B 测试非常容易，构建平台也非常容易。

[00:25:39] Ronny Kohavi

English:

I don't say it's easy to build a platform. But once you build a platform, the incremental cost of running an experiment should approach zero. And we got to that at Microsoft, where after a while, the cost of running experiments was so low that nobody was questioning the idea that everything should be experimented with.

中文翻译:

我不是说构建平台很容易。但一旦你构建了平台，运行实验的边际成本应该趋近于零。我们在微软做到了这一点，一段时间后，运行实验的成本如此之低，以至于没有人再质疑“一切都应该进行实验”这个想法。

[00:25:59] Ronny Kohavi

English:

Now, I don't think we were there at Airbnb for example. The platform at Airbnb was much less mature, and required a lot more analysts in order to interpret the results and to find issues with it. So I do think there's this trade off. You're willing to invest in the platform. It is possible to get the marginal cost to be close to zero. But when you're not there, it's still expensive, and there may be reasons why not to run A/B tests.

中文翻译:

例如，我不认为我们在 Airbnb 达到了那个水平。Airbnb 的平台成熟度要低得多，需要更多的分析师来解释结果并发现问题。所以我认为这是一种权衡。如果你愿意投资平台，是有可能让边际成本接近于零的。但如果你还没达到那个阶段，实验仍然很昂贵，可能就有理由不运行 A/B 测试。

[00:26:28] Lenny

English:

You talked about how you may be too small to run A/B tests, and this is a constant question for startups is, when should we start running A/B tests? Do you have kind of a heuristic or a rule of thumb of, here's a time you should really start thinking about running an A/B test?

中文翻译:

你谈到规模太小可能无法运行 A/B 测试，这是初创公司经常问的问题：我们什么时候应该开始运行 A/B 测试？你有没有什么启发式的方法或经验法则，比如“到了这个时间点，你就真的该考虑运行 A/B 测试了”？

[00:26:42] Ronny Kohavi

English:

Yeah, a dollar question that everybody asks. So actually, we'll put this in the notes, but I gave a talk last year, what I called it is practical defaults. And one of the things I show there is that unless you have at least tens of thousands of users, the math, the statistics just don't work out for most of the metrics that you're interested in.

中文翻译:

是的，这是每个人都会问的关键问题。实际上，我们会把这个放在节目介绍里，但我去年做过一个演讲，我称之为“实用默认值”。我在那里展示的一点是，除非你至少有数万名用户，否则对于你感兴趣的大多数指标，数学和统计学根本无法奏效。

[00:27:05] Ronny Kohavi

English:

In fact, I gave an actual practical number of a retail site with some conversion rate, trying to detect changes that are at least 5% beneficial, which is something that startups should focus on. They shouldn't

focus on the 1%, they should focus on the 5 and 10%. Then you need something like 200,000 users.

中文翻译:

事实上，我给出了一个实际的数字：对于一个具有一定转化率的零售网站，如果你想检测出至少 5% 的收益提升（这是初创公司应该关注的，他们不该关注 1% 的提升，而应关注 5% 或 10%），那么你需要大约 20 万名用户。

[00:27:25] Ronny Kohavi

English:

So start experimenting when you're in the tens of thousands of users. You'll only be able to detect large effects. And then once you get to 200,000 users, then the magic starts happening. Then you can start testing a lot more. Then you have the ability to test everything and make sure that you're not degrading and getting value out of experimentation. So you ask for rule of thumb, 200,000 users, you're magical. Below that, start building the culture, start building the platform, start integrating. So that as you scale, you start to see the value.

中文翻译:

所以，当你拥有数万名用户时，就可以开始实验了，但你只能检测到巨大的影响。一旦你达到 20 万名用户，奇迹就开始发生了。那时你可以开始测试更多东西，有能力测试一切，确保你没有在退步，并从实验中获得价值。所以你问经验法则：20 万用户，你就进入了神奇阶段。在那之前，开始建立文化，开始构建平台，开始整合。这样随着规模的扩大，你就能开始看到价值。

[00:28:00] Lenny

English:

Love it. Coming back to this kind of concern people have of experimentation, keeps you from innovating and taking big bets, I know you have this framework overall evaluation criterion, and I think that helps with this. Can you talk a bit about that?

中文翻译:

太棒了。回到人们对实验的担忧——即实验会阻碍创新和进行重大博弈。我知道你有一个叫“综合评价指标”（OEC）的框架，我认为这有助于解决这个问题。你能谈谈吗?

[00:28:14] Ronny Kohavi

English:

The OEC or the overall evaluation criterion is something that I think many people that start to dabble in A/B testing miss. And the question is, what are you optimizing for? And it's a much harder question that people think because it's very easy to say we're going to optimize for money, revenue. But that's the wrong question, because you can do a lot of bad things that will improve revenue. So there has to be some countervailing metric that tells you, how do I improve revenue without hurting the user experience?

中文翻译:

OEC，即综合评价指标，我认为许多刚开始接触 A/B 测试的人都会忽略它。问题在于：你在优化什么？这是一个比人们想象中难得多的问题，因为说“我们要优化金钱、收入”非常容易。但那是错误的问题，因为你可以

做很多坏事来增加收入。所以必须有一些制衡指标（countervailing metric）来告诉你：如何在不损害用户体验的情况下增加收入？

[00:28:53] Ronny Kohavi

English:

So let's take a good example with search. You can put more ads on a page and you will make more money. There's no doubt about it. You will make more money in the short term. The question is, what happens to the user experience, and how is that going to impact you in the long term?

中文翻译:

让我们以搜索为例。你可以在页面上放更多广告，你肯定会赚更多钱。毫无疑问，短期内你会赚更多。问题是，用户体验会发生什么变化，这在长期内会如何影响你？

[00:29:13] Ronny Kohavi

English:

So we've run those experiments, and we were able to map out this number of ads causes this much increase to churn, this number of ads causes this much increase to the time that users take to find a successful result. And we came up with an OEC that is based on all these metrics that allows you to say, "Okay, I'm willing to take this additional money if I'm not hurting the user experience by more than this much." So there's a trade-off there.

中文翻译:

所以我们运行了这些实验，并能够绘制出：广告数量增加会导致流失率增加多少，会导致用户找到成功结果所需的时间增加多少。我们提出了一个基于所有这些指标的 OEC，让你能够说：“如果我对用户体验的损害不超过这个程度，我愿意赚这笔额外的钱。”所以这里有一个权衡。

[00:29:41] Ronny Kohavi

English:

One of the nice ways to phrase this, as a constraint optimization problem. I want you to increase revenue, but I'm going to give you a fixed amount of average real estate that you can use. So for one query, you can have zero ads. For another query, you can have three ads. For a third query, you can have wider, bigger ads. I'm just going to count the pixels that you take, the vertical pixels. And I will give you some budget. And if you can under the same budget make more money, you're good to go.

中文翻译:

一种很好的表达方式是将其视为“约束优化问题”。我希望你增加收入，但我会给你一个固定的平均页面空间（real estate）预算。对于一个查询，你可以放零个广告；对于另一个，你可以放三个；对于第三个，你可以放更宽、更大的广告。我只计算你占用的像素，即垂直像素。我会给你一定的预算，如果你能在相同的预算下赚更多的钱，那你就成功了。

[00:30:16] Ronny Kohavi

English:

So that to me turns the problem from a badly defined, let's just make more money. Any page can start plastering more ads and make more money short term, but that's not the goal. The goal is long-term growth and revenue. Then you need to insert these other criteria, and what am I doing to the user experience? One way around it is to put this constraint. Another one is just to have these other metrics. Again, something that we did, to look at the user experience. How long does it take the user to reach a successful click? What percentage of sessions are successful? These are key metrics that were part of the overall evaluation criteria, that we've used.

中文翻译:

所以对我来说，这把问题从一个定义模糊的“让我们赚更多钱”转变成了一个更科学的问题。任何页面都可以通过贴满广告在短期内赚更多钱，但那不是目标。目标是长期的增长和收入。那么你就需要加入其他标准：我对用户体验造成了什么影响？一种方法是设置约束条件，另一种是引入其他指标。这也是我们做过的事情，比如观察用户体验：用户需要多长时间才能完成一次成功的点击？成功会话的百分比是多少？这些都是我们使用的综合评价指标中的关键指标。

[00:30:55] Ronny Kohavi

English:

I can give you another example by the way, from the hotel industry or Airbnb that we both worked at. You can say, "I want to improve conversion rate," but you can be smarter about it and say, "It's not just enough to convert a user to buy or to pay for a listing. I want them to be happy with it several months down the road when they actually stay there."

中文翻译:

顺便说一下，我可以给你另一个来自酒店行业或我们都工作过的 Airbnb 的例子。你可以说“我想提高转化率”，但你可以更聪明一点，说：“仅仅让用户转化、购买或支付房源是不够的。我希望他们在几个月后真正入住时感到满意。”

[00:31:19] Ronny Kohavi

English:

So that could be part of your OEC to say, "What is the rating that they will give to that listing when they actually stay there?" And that causes an interesting problem, because you don't have this data now. You're going to have it three months from now when they actually stay. So you have to build the training set that allows you to make a prediction about whether this user, whether Lenny is going to be happy at this cheap place. Or whether no, I should offer him something more expensive, because Lenny likes to stay at nicer places where the water actually is hot and comes out of the faucet.

中文翻译:

所以这可以成为你 OEC 的一部分，即：“当他们真正入住时，会给该房源打多少分？”这引发了一个有趣的问题，因为你现在没有这个数据，你要等三个月后他们入住时才有。所以你必须构建一个训练集，让你能够预测这个用户（比如 Lenny）在这个便宜的地方是否会开心。或者不，我应该给他推荐更贵的地方，因为 Lenny 喜欢住在更好的地方，那里水龙头里流出的水真的是热的。

[00:31:52] Lenny

English:

That is true. Okay, so it sounds like the core to this approach is basically have a drag metric that makes sure you're not hurting something that's really important to the business, and then being very clear on what's the long-term metric we care most about.

中文翻译:

确实如此。好，所以听起来这种方法的核心基本上是设置一个“阻力指标”（drag metric），确保你没有损害对业务非常重要的东西，然后非常明确我们最关心的长期指标是什么。

[00:32:05] Ronny Kohavi

English:

To me, the key word is lifetime value, which is you have to define the OEC such that it is causally predictive of the lifetime value of the user. And that's what causes you to think about things properly, which is, am I doing something that just helps me short term, or am I doing something that will help me in the long term? Once you put that model of lifetime value, people say, "Okay, what about retention rates? We can measure that. What about the time to achieve a task? We can measure that." And those are these countervailing metrics that make the OEC useful.

中文翻译:

对我来说，关键词是“终身价值”（LTV）。你必须定义 OEC，使其能够因果性地预测用户的终身价值。这会促使你正确地思考问题：我做的事情是仅仅在短期内有帮助，还是在长期内有帮助？一旦你建立了终身价值模型，人们就会说：“好吧，那留存率呢？我们可以测量。完成任务的时间呢？我们也可以测量。”这些就是让 OEC 变得有用的制衡指标。

[00:32:43] Lenny

English:

And to understand these longer term metrics, what I'm hearing is use models, and forecast, and predictions, or would you suggest sometimes use a long-term holdout or some other approach? What do you find is the best way to see these long term?

中文翻译:

为了理解这些长期指标，我听到的是使用模型、预测和预估。或者你会建议有时使用长期对照组（long-term holdout）或其他方法吗？你认为观察这些长期影响的最佳方式是什么？

[00:32:57] Ronny Kohavi

English:

Yeah, so there's two ways that I like to think about it. One is you can run long-term experiments for the goal of learning something. So I mentioned that at Bing, we did run these experiments where we increased the ads and decreased the ads, so that we will understand what happens to key metrics.

中文翻译:

是的，我喜欢从两个方面来思考。一是你可以为了学习某些东西而运行长期实验。正如我提到的，在 Bing，我们确实运行过增加广告和减少广告的实验，以便了解关键指标会发生什么变化。

[00:33:16] Ronny Kohavi

English:

The other thing is you can just build models that use some of our background knowledge or use some data science to look at historical... I'll give you another good example of this. When I came to Amazon, one of the teams that reported to me was the email team that it was not the transactional emails when you buy something, you get an email. But it was the team that sent these recommendations. "Here's a book by an author that you bought. Here's a product that we recommend." And the question is, how do we give credit to that team?

中文翻译:

另一件事是你可以构建模型，利用我们的背景知识或数据科学来查看历史数据……我再给你举一个很好的例子。当我来到亚马逊时，向我汇报的团队之一是邮件团队。这不是指你买东西后收到的交易邮件，而是发送推荐信息的团队，比如“这是你买过书的作者出的新书”、“这是我们推荐的产品”。问题是，我们如何给这个团队记功（归因）？

[00:33:49] Ronny Kohavi

English:

And the initial version was, whenever a user comes from the email and purchases something on Amazon, we're going to give that email credit. Well, it turned out this had no countervailing metric. The more emails you send, the more money you're going to credit the team. And so that led to spam. Literally a really interesting problem. The team just ramped up the number of emails that they were sending out, and claimed to make more money, and their fitness function improved.

中文翻译:

最初的版本是：每当用户通过邮件来到亚马逊并购买东西时，我们就给那封邮件记功。结果发现，这没有制衡指标。你发送的邮件越多，记在团队名下的钱就越多。这导致了垃圾邮件泛滥。这真的是一个非常有趣的问题。团队只是不断增加发送邮件的数量，并声称赚了更多钱，他们的适应度函数（fitness function）确实提高了。

[00:34:20] Ronny Kohavi

English:

So then we backed up and then we said, "Okay, we can either phrase this as a constraint satisfaction problem. You're allowed to send user an email every X days or," which is what we ended up doing is, "Let's model the cost of spamming the users."

中文翻译:

于是我们退后一步说：“好吧，我们要么将其表述为一个约束满足问题——比如允许每 X 天给用户发一封邮件；要么——这也是我们最终做的——让我们对‘骚扰用户’的成本进行建模。”

[00:34:37] Ronny Kohavi

English:

What's that cost? Well, when they unsubscribe, we can't mail them. So we did some data science study on the side and we said, "What is the value that we're losing from an unsubscribe?" And we came up with a

number, it was a few dollars. But the point was, now we have this countervailing metric. We say, "Here's the money that we generate from the emails. Here's the money that we're losing on long-term value. What's the trade-off?" And then when we started to incorporate those formula, more than half the campaigns that were being sent were negative.

中文翻译:

成本是什么？当用户退订时，我们就不能再给他们发邮件了。所以我们做了一些侧面的数据科学研究，问：“一个退订让我们损失了多少价值？”我们得出了一个数字，大概是几美元。但关键在于，现在我们有了这个制衡指标。我们说：“这是邮件产生的收入，这是我们损失的长期价值。权衡是什么？”当我们开始引入这些公式时，发现超过一半的邮件营销活动其实是负收益的。

[00:35:14] Ronny Kohavi

English:

So it was a huge insight at Amazon about how to send the right campaigns. And this is what I like about these discoveries. This fact that we integrated the unsubscribe led us to a new feature to say, "Well, let's not lose their future lifetime value through email. When they unsubscribe, let's offer them by default to unsubscribe from this campaign."

中文翻译:

这在亚马逊是一个巨大的洞察，关于如何发送正确的营销活动。这就是我喜欢这些发现的地方。我们将退订因素整合进来的事实引导我们开发了一个新功能：“好吧，让我们不要因为邮件而失去他们未来的终身价值。当他们退订时，让我们默认提供‘仅退订此系列活动’的选项。”

[00:35:41] Ronny Kohavi

English:

So when you get an email, there's a new book by the author. The default to unsubscribe would be unsubscribe me from author emails. And so now, the negative, the countervailing metric is much smaller. And so again, this was a breakthrough in our ability to send more emails, and understand based on what users were unsubscribing from, which ones are really beneficial.

中文翻译:

所以当你收到一封关于某作者新书的邮件时，默认的退订选项是“退订该作者的邮件”。这样一来，负面的制衡指标就小得多。这又是我们在发送更多邮件能力上的一个突破，并且能根据用户退订的内容，理解哪些邮件是真正有益的。

[00:36:06] Lenny

English:

I love the surprising results.

中文翻译:

我喜欢这些令人惊讶的结果。

[00:36:08] Ronny Kohavi

English:

We all love them. This is the humbling reality. And people talk about the fact that A/B testing sometimes leads you to incremental... I actually think that many of these small insights lead to fundamental insights about which areas to go, some strategies we should take, some things we should develop. Helps a lot.

中文翻译:

我们都喜欢。这就是让人清醒的现实。人们常说 A/B 测试有时会导致你陷入渐进式改进……但我实际上认为，许多这些微小的洞察会引导你产生关于发展方向、应采取的策略以及应开发的功能的根本性洞察。非常有帮助。

[00:36:31] Lenny

English:

This makes me think about how every time I've done a full redesign of a product, I don't think ever, has it ever been a positive result. And then the team always ends up having to claw back what they just hurt and try to figure out what they messed up. Is that your experience too?

中文翻译:

这让我想起，每次我做产品的全面重新设计（redesign）时，我印象中从来没有过正向的结果。最后团队总是不得不收回那些造成损害的改动，并试图弄清楚到底搞砸了什么。这也是你的经验吗？

[00:36:47] Ronny Kohavi

English:

Absolutely, yeah. In fact, I've published some of these in LinkedIn posts showing a large set of big launches and redesigns that dramatically failed, and it happens very often. So the right way to do this is to say, "Yes, we want to do a redesign, but let's do it in steps and test on the way and adjust," so you don't need to take 17 new changes, that many of them are going to fail. Start to move incrementally in a direction that you believe is beneficial. Adjust on the way.

中文翻译:

绝对是这样。事实上，我在 LinkedIn 上发过一些帖子，展示了一大堆惨遭失败的重大发布和重新设计案例，这种情况非常频繁。所以正确的方法应该是：“是的，我们想做重新设计，但让我们分步进行，边走边测试并调整。”这样你就不需要一次性做 17 个新改动，因为其中很多都会失败。朝着你认为有益的方向逐步推进，并在途中进行调整。

[00:37:24] Lenny

English:

The worst part of those experiences I find is it took three to six months to build it. And by the time it's launched, it's just like, "We're not going to unlaunch this. Everyone's been working in this direction. All the new features are assuming this is going to work," and you're basically stuck.

中文翻译:

我发现这些经历中最糟糕的部分是，花了三到六个月来构建它。等到发布时，大家就会觉得：“我们不能撤销发布。每个人都一直朝着这个方向努力，所有新功能都假设这个设计会奏效。”然后你就基本上被困住了。

[00:37:41] Ronny Kohavi

English:

I mean, this is a sunk cost fallacy. We invested so many years in it. Let's launch this, even though it's bad for the user. No, that's terrible. Yeah. Yeah. So this is the other advantage of recognizing this humble reality that most ideas fail. If you believe in that statistics that I published, then doing 17 changes together is more likely to be negative. Do them in smaller increments, learn from, it's called OFAT one-factor-at-a-time. Do one factor, learn from it, and adjust. Of the 17, maybe you have four good ideas. Those are the ones that will launch and be positive.

中文翻译:

这就是“沉没成本谬误”（sunk cost fallacy）。“我们投入了这么多时间，即使对用户不好，也发布吧。”不，那太可怕了。所以，承认“大多数想法都会失败”这个谦卑现实的另一个好处是：如果你相信我公布的统计数据，那么同时做 17 个改动更有可能产生负面影响。应该以更小的增量来做，从中学习，这被称为 OFAT（一次一个因素）。做一个因素，从中学习，然后调整。在那 17 个想法中，也许你有 4 个好主意，这些才是最终发布并产生正向影响的想法。

[00:38:22] Lenny

English:

I generally agree with that, and always try to avoid a big redesign, but it's hard to avoid them completely. There's often team members that are really passionate like, "We just need to rethink this whole experience. We're not going to incrementally get there." Have you found anything effective in helping people either see this perspective, or just making a larger bet more successful?

中文翻译:

我通常同意这一点，并总是试图避免重大的重新设计，但很难完全避免。经常会有团队成员充满激情地说：“我们只需要重新思考整个体验，靠渐进式改进是达不到目标的。”你有没有发现什么有效的方法，能帮助人们理解这个观点，或者让重大的博弈变得更成功?

[00:38:42] Ronny Kohavi

English:

By the way, I'm not opposed to large redesigns. I try to give the team the data to say, "Look, here are lots of examples where big redesigns fail." Try to decompose your redesign if you can't decompose it to one factor at a time, to a small set of factors at a time. And learn from these smaller changes what works and what doesn't.

中文翻译:

顺便说一下，我不反对重大的重新设计。我尝试给团队提供数据，告诉他们：“看，这里有很多重大重新设计失败的例子。”尝试分解你的重新设计，如果不能分解到一次一个因素，那就分解到一次一小部分因素。从这些较小的改动中学习哪些有效，哪些无效。

[00:39:08] Ronny Kohavi

English:

Now, it's also possible to do a complete redesign. Just, as you said yourself, be ready to fail. I mean, do you really want to work on something for six months or a year, and then run the A/B test, and realize that you've hurt revenues or other key metrics by several percentage points? And a data-driven organization will not allow you to launch. What are you going to write in your annual review?

中文翻译:

当然，进行彻底的重新设计也是可能的。只是正如你自己所说，要做好失败的准备。我的意思是，你真的想花六个月或一年时间做某件事，然后运行 A/B 测试，结果发现你损害了收入或其他关键指标好几个百分点吗？一个数据驱动的组织是不会允许你发布的。那你年终总结怎么写？

[00:39:33] Lenny

English:

But nobody ever thinks it's going to fail. They think, "No, we got this. We've talked to so many people."

中文翻译:

但没人会觉得自己会失败。他们会想：“不，我们能行，我们已经和很多人聊过了。”

[00:39:38] Ronny Kohavi

English:

But I think organizations that start to run experiments are humbled early on from the smaller changes. Right? You're right. I'll tell you a funny story. When I came from Amazon to Microsoft, I joined the group, and for one reason or another, that group disbanded a month after I joined.

中文翻译:

但我认为，开始运行实验的组织在早期就会从那些微小的改动中学会谦卑。你说得对。我给你讲个有趣的故事。当我从亚马逊来到微软时，我加入了一个小组，但由于某种原因，那个小组在我加入一个月后就解散了。

[00:39:57] Ronny Kohavi

English:

And so people came to me and said, "Look, you just joined the company. You're at partner level. You figure out how you can help Microsoft." And I said, "I'm going to build an experimentation platform," because nobody at Microsoft is running experiments. And more than 50% of ideas in Amazon that we tried failed. And the classical response was, "We have better PMs here."

中文翻译:

于是人们来找我说：“看，你刚加入公司，你是合伙人级别，你想想怎么能帮到微软。”我说：“我要构建一个实验平台。”因为当时微软没人运行实验。而在亚马逊，我们尝试的想法有超过 50% 都失败了。当时最经典的反应是：“我们这里的 PM（产品经理）更优秀。”

[00:40:26] Ronny Kohavi

English:

Right? There was this complete denial that it's possible that 50% of ideas that Microsoft is implementing, in a three-year development cycle by the way. This is how long it took Office to release. It was a classical every three years we release.

中文翻译:

对吧？当时人们完全否认微软实施的想法有 50% 可能会失败——顺便说一下，当时是三年的开发周期。Office 过去就是每三年发布一次。

[00:40:42] Ronny Kohavi

English:

And the data came about showing that Bing was the first to truly implement experimentation at scale. And we shared with the rest of the companies the surprising results. And so when Office was... And this was credit to Qi Lu and Satya Nadella, they were ones that says, "Ronny, you try to get Office to run experiments. We'll give you the air support." And it was hard, but we did it. It took a while, but Office started to run experiments, and they realized that many of their ideas are failing.

中文翻译:

后来数据证明，Bing 是第一个真正大规模实施实验的部门。我们向公司其他部门分享了那些令人惊讶的结果。所以当 Office 部门……这要归功于陆奇和萨提亚·纳德拉（Satya Nadella），他们说：“Ronny，你去试着让 Office 运行实验，我们会给你空中支援。”这很难，但我们做到了。花了一段时间，但 Office 开始运行实验后，他们意识到自己的许多想法确实在失败。

[00:41:20] Lenny

English:

You said that there's a site of a failed redesigns. Is that in your book or is that a site that you can point people to, to help build this case?

中文翻译:

你提到有一个关于失败的重新设计的网站。是在你的书里，还是你可以指给人们看的一个网站，用来帮助建立这种案例？

[00:41:29] Ronny Kohavi

English:

I teach this in my class, but I think I've posted this on LinkedIn and answered some questions. I'm happy to put that in the notes.

中文翻译:

我在课堂上讲过这个，我也在 LinkedIn 上发过帖子并回答过一些问题。我很乐意把它放在节目介绍里。

[00:41:36] Lenny

English:

Okay, cool. We'll put that in the show notes. Because I think that's the kind of data that often helps convince a team, "Maybe we shouldn't rethink this entire onboarding flow from scratch. Maybe we should iterate towards and learn as we go."

中文翻译:

好的，太棒了。我们会把它放在节目介绍里。因为我认为这种数据通常有助于说服团队：“也许我们不应该从头开始重新思考整个新手引导流程。也许我们应该逐步迭代并边走边学。”

[00:41:48] Lenny

English:

This episode is brought to you by Eppo. Eppo is a next generation A/B testing platform built by Airbnb alums for modern growth teams. Companies like DraftKings, Zapier, ClickUp, Twitch, and Cameo rely on Eppo to power their experiments.

中文翻译:

本期节目由 Eppo 赞助。Eppo 是由 Airbnb 校友为现代增长团队构建的下一代 A/B 测试平台。DraftKings、Zapier、ClickUp、Twitch 和 Cameo 等公司都依赖 Eppo 来驱动他们的实验。

[00:42:02] Lenny

English:

Wherever you work, running experiments is increasingly essential, but there are no commercial tools that integrate with a modern growth team stack. This leads to wasted time building internal tools or trying to run your own experiments through a clunky marketing tool.

中文翻译:

无论你在哪里工作，运行实验都变得越来越重要，但目前还没有商业工具能与现代增长团队的技术栈完美整合。这导致人们浪费时间构建内部工具，或者尝试通过笨重的营销工具运行实验。

[00:42:15] Lenny

English:

When I was at Airbnb, one of the things that I loved most about working there was our experimentation platform, where I was able to slice and dice data by device types, country, user stage.

中文翻译:

当我在 Airbnb 时，我最喜欢的工作内容之一就是我们的实验平台，我可以在那里按设备类型、国家、用户阶段对数据进行切片分析。

[00:42:25] Lenny

English:

Eppo does all that and more, delivering results quickly, avoiding annoying prolonged analytic cycles, and helping you easily get to the root cause of any issue you discover. Eppo lets you go beyond basic click through metrics, and instead use your North Star metrics like activation, retention, subscription and

payments. Eppo supports tests on the front end, on the back end, email marketing, even machine learning clients. Check out Eppo at geteppo.com, that's geteppo.com, and 10X your experiment velocity.

中文翻译:

Eppo 具备所有这些功能甚至更多，它能快速交付结果，避免烦人的漫长分析周期，并帮助你轻松找到发现的任何问题的根本原因。Eppo 让你超越基本的点击率指标，转而使用激活、留存、订阅和支付等北极星指标。Eppo 支持前端、后端、邮件营销甚至机器学习客户端的测试。访问 geteppo.com 查看 Eppo，让你的实验速度提升 10 倍。

[00:42:55] Lenny

English:

Is it ever worth just going, "Let's just rethink this whole thing and just give it a shot," to break out of a local minima or local maxima essentially?

中文翻译:

为了突破所谓的“局部最小值”或“局部最大值”，是否值得直接说“让我们重新思考整件事并试一试”？

[00:43:03] Ronny Kohavi

English:

Yeah. So I think what you said is fair. I do want to allocate some percentage of resources to big bets. As you said, we've been optimizing this thing to hell. Could we completely redesign it? It's a very valid idea. You may be able to break out of a local minima. What I'm telling you is 80% of the time, you will fail. So be ready for that. What people usually expect is, "My redesign is going to work." No, you're most likely going to fail, but if you do succeed, it's a breakthrough.

中文翻译:

是的。我认为你说的很公平。我确实想为重大博弈分配一定比例的资源。正如你所说，我们已经把这个东西优化到极致了，能不能彻底重新设计它？这是一个非常合理的想法。你可能会突破局部最小值。但我告诉你的是，80% 的时间你会失败。所以要做好准备。人们通常预期的是：“我的重新设计会奏效。”不，你极有可能失败，但如果你真的成功了，那就是一个突破。

[00:43:35] Lenny

English:

I like this 80% rule of thumb. Is that just a simple way of thinking about it? 80% of your-

中文翻译:

我喜欢这个 80% 经验法则。这只是一个简单的思考方式吗？你 80% 的——

[00:43:39] Ronny Kohavi

English:

That's my rule of thumb. And I've heard people say it's 70% or 80%. But it's in that area where I think when you talk about how much to invest in the known versus the high risk, high reward, that's usually the

right percentage that most organizations end up doing this allocation, right? You interviewed Shreyas. I think he mentioned that Google is like 70% the searching ads, and it's 20% for some of the apps and new stuff, and then it's the 10% for infrastructure.

中文翻译:

这是我的经验法则。我听人说是 70% 或 80%。但在讨论投入已知领域与高风险高回报领域的比例时，这通常是大多数组织最终采取的分配比例，对吧？你采访过 Shreyas，我想他提到过 Google 大约是 70% 投入搜索广告，20% 投入应用和新事物，10% 投入基础设施。

[00:44:16] Lenny

English:

And I think the most important point there is if you're not running an experiment, 70% of stuff you're shipping is hurting your business.

中文翻译:

我认为最重要的一点是：如果你不运行实验，你发布的 70% 的东西可能正在损害你的业务。

[00:44:23] Ronny Kohavi

English:

Well, it's not hurting, it's flat too negative. Some of them are flat. And by the way, flat to me, if something is not Statsig, that's a no ship, because you've just introduced more code. There is a maintenance overhead to shipping your stuff. I've heard people say, "Look, we already spent all this time. The team will be demotivated if we don't ship it." And I'm, "No, that's wrong guys. Let's make sure that we understand that shipping this project has no value, is complicating the code base. Maintenance costs will go up." You don't ship on flat, unless it's a legal requirement. When legal comes along and says, "You have to do X or Y," you have to ship on flat or even negative. And that's understandable.

中文翻译:

嗯，不一定是损害，而是“毫无波动”到“负面”。有些是毫无波动的。顺便说一下，对我来说，如果某个改动没有达到统计显著（Statsig），那就不能发布，因为你只是引入了更多代码。发布东西是有维护开销的。我听人说：“看，我们已经花了这么多时间，如果不发布，团队会失去动力。”我会说：“不，伙计们，那是错的。我们要明白，发布这个没有价值的项目只会让代码库变得复杂，维护成本会上升。”除非是法律要求，否则不要发布毫无波动的改动。当法务部门说“你必须做 X 或 Y”时，即使是毫无波动甚至负面的改动你也得发布，那是可以理解的。

[00:45:08] Ronny Kohavi

English:

But again, I think that's something that a lot of people make the mistake of saying, "Legal told us we have to do this, therefore we're going to take the hits." No, legal gave you a framework that you have to work under. Try three different things, and ship the one that hurts the least.

中文翻译:

但同样，我认为很多人犯的错误是说：“法务告诉我们必须这样做，所以我们要承受损失。”不，法务只是给了你一个必须遵守的框架。尝试三种不同的方案，然后发布损害最小的那一个。

[00:45:25] Lenny

English:

That reminds me when Airbnb launched the rebrand, even that they ran as an experiment with the entire homepage redesigned, the new logo, and all that. And I think there was a long-term holdout even, and I think it was positive in the end from what I remember.

中文翻译:

这让我想起 Airbnb 发布品牌重塑（rebrand）时，甚至连那个也是作为一个实验运行的，包括整个首页重新设计、新 Logo 等等。我记得甚至还有一个长期对照组，最后结果是正向的。

[00:45:41] Lenny

English:

Speaking of Airbnb, I want to chat about Airbnb briefly. I know you're limited in what you can share, but it's interesting that Airbnb seems to be moving in this other direction where it's becoming a lot more top-down, Brian vision oriented. And Brian's even talked about how he's less motivated to run experiments. He doesn't want to run as many experiments as they used to. Things are going well, and so it's hard to argue with the success potentially. You worked there for many years. You ran the search team essentially. I guess, what was your experience like there? And then roughly, what's your sense of how things are going, where it's going?

中文翻译:

说到 Airbnb，我想简短聊聊。我知道你能分享的内容有限，但有趣的是，Airbnb 似乎正朝着另一个方向发展，变得更加自上而下，以 Brian（切斯基）的愿景为导向。Brian 甚至谈到他运行实验的动力减弱了，他不想像以前那样运行那么多实验。目前情况进展顺利，所以很难反驳这种成功。你在那里工作了很多年，领导搜索团队。你在那里的经历是怎样的？你觉得现在情况如何，未来会走向何方？

[00:46:15] Ronny Kohavi

English:

Well as you know, I'm restricted from talking about Airbnb. I will say a few things that I am allowed to say. One is in my team in search relevance, everything was A/B tested. So while Brian can focus on some of the design aspects, the people who are actually doing the neural networks and the search, everything was A/B tested to hell. So nothing was launching without an A/B test. We had targets around improving certain metrics, and everything was done A/B test.

中文翻译:

如你所知，我被限制谈论 Airbnb。我会说几点我被允许说的。一是在我的搜索相关性团队中，所有东西都要经过 A/B 测试。所以虽然 Brian 可以专注于某些设计方面，但真正做神经网络和搜索的人，所有东西都经过了极其严格的 A/B 测试。没有任何东西是在没有 A/B 测试的情况下发布的。我们有改进某些指标的目标，一切都通过 A/B 测试完成。

[00:46:50] Ronny Kohavi

English:

Now other teams, some did, some did not. I will say that when you say things are going well, I think we don't know the counterfactual. I believe that had Airbnb kept people like Greg Greeley, which was pushing for a lot more data driven, and had Airbnb run more experiments, it would've been in a better state than today. But it's the counterfactual. We don't know.

中文翻译:

至于其他团队，有的做了，有的没做。我想说的是，当你谈到“情况进展顺利”时，我认为我们不知道“反事实”（counterfactual）结果。我相信，如果 Airbnb 留住了像 Greg Greeley 这样推动数据驱动的人，如果 Airbnb 运行了更多实验，它现在的状态会比今天更好。但那是反事实，我们无从得知。

[00:47:14] Lenny

English:

That's a really interesting perspective. Airbnb's such an interesting natural experiment of a way of doing things differently. There's de-emphasizing experiments, and also, they turned off paid ads during Covid. And I don't know where it is now, but it feels like it's become a much smaller part of the growth strategy. Who knows if they've ramped it up to back to where it's today, but I think it's going to be a really interesting case study looking back five, 10 years from now.

中文翻译:

这是一个非常有趣的视角。Airbnb 本身就是一个非常有趣的“自然实验”，展示了不同的做事方式。他们不再强调实验，而且在疫情期间关闭了付费广告。我不知道现在情况如何，但感觉这已经成为增长战略中很小的一部分。谁知道他们现在是否已经恢复到了以前的水平，但我认为五到十年后回过头来看，这将是一个非常有趣的案例研究。

[00:47:38] Ronny Kohavi

English:

It's a one-off experiment where it's hard to assign value to some of the things that Airbnb is doing. I personally believe it could have been a lot bigger and a lot more successful if it had run more controlled experiments. But I can't speak about some of those that I ran and that showed that some of the things that were initially untested were actually negative and could be better.

中文翻译:

这是一个孤立的实验，很难给 Airbnb 正在做的一些事情分配价值。我个人认为，如果它运行了更多受控实验，它本可以规模更大、更成功。但我不能谈论我运行过的一些实验，那些实验表明，一些最初未经测试的东西实际上是负面的，本可以做得更好。

[00:48:04] Lenny

English:

All right. Mysterious. One more question. Airbnb, you were there during Covid, which was quite a wild time for Airbnb. We had Sanchan on the podcast talking about all the craziness that went on when travel basically stopped, and there was a sense that Airbnb was done, and travel's not going to happen for years and years. What's your take on experimentation in that world where you have to really move fast, make crazy decisions, and make big decisions? What was it like during that time?

中文翻译:

好吧，真神秘。还有一个问题。疫情期间你在 Airbnb，那对 Airbnb 来说是一个非常疯狂的时期。Sanchan 曾在播客上谈到当旅游业基本停滞时发生的各种疯狂事情，当时有一种感觉是 Airbnb 完蛋了，旅游业多年都不会恢复。在那种必须快速行动、做出疯狂决定和重大决定的环境下，你对实验有什么看法？那段时间是什么样的？

[00:48:34] Ronny Kohavi

English:

So I think actually in a state like that, it's even more important to run A/B tests, right? Because what you want to be able to see is if we're making this change, is it actually helping in the current environment? There's this idea of external generalizability. Is it going to work out now during Covid? Is it going to generalize later on? These are things that you can really answer with the controlled experiments, and sometimes it means that you might have to replicate them six months down when Covid say is not as impactful as it is.

中文翻译:

所以我认为，在那种状态下，运行 A/B 测试实际上更加重要，对吧？因为你想看到的是，如果我们做出这个改变，它在当前环境下是否真的有帮助？这里有一个“外部普遍性”（external generalizability）的概念：它在疫情期间有效吗？以后能推广吗？这些是你真正可以通过受控实验来回答的问题，有时这意味着你可能需要在六个月后（当疫情影响减弱时）重复实验。

[00:49:11] Ronny Kohavi

English:

Saying that you have to make decisions quickly, to me, I'll point you to the success rate. If in peace time you're wrong two thirds to 80% of the time, why would you be subtly right in wartime, in Covid time?

中文翻译:

至于说必须快速做决定，我会让你看看成功率。如果在和平时期你 2/3 到 80% 的时间都是错的，为什么在战争时期、在疫情期间你就会突然变正确了呢？

[00:49:26] Ronny Kohavi

English:

So I don't believe in the idea that because bookings went down materially, the company should suddenly not be data driven and do things differently. I think if Airbnb stayed the course, did nothing, the revenue would've gone up in the same way.

中文翻译:

所以我不相信因为预订量大幅下降，公司就应该突然不再以数据驱动，转而采取不同的做法。我认为如果 Airbnb 坚持原有路线，什么都不做，收入也会以同样的方式回升。

[00:49:49] Lenny

English:

Fascinating.

中文翻译:

很有意思。

[00:49:49] Ronny Kohavi

English:

In fact, if you look at one investment, one big investment that was done at the time was online experiences, and the initial data wasn't very promising. And I think today, it's a footnote.

中文翻译:

事实上，如果你看当时的一项重大投资——“在线体验”（online experiences），最初的数据并不乐观。我想在今天，它只是一个注脚。

[00:50:01] Lenny

English:

Yeah. Another case study for the history books, Airbnb experiences. I want to shift a little bit and talk about your book, which you mentioned a couple times. It's called Trustworthy Online Controlled Experiments, and I think it's basically the book on A/B testing. Let me ask you, what surprised you most about writing this book, and putting it out, and the reaction to it?

中文翻译:

是的，又一个可以载入史册的案例研究——Airbnb 体验。我想转换一下话题，聊聊你的书，你刚才提到过几次。书名叫《值得信赖的在线对照实验》（Trustworthy Online Controlled Experiments），我认为它基本上就是 A/B 测试领域的权威著作。我想问你，写这本书、出版它以及人们的反应，最让你惊讶的是什么？

[00:50:24] Ronny Kohavi

English:

I was pleasantly surprised that it sold more than what we thought, more than what Cambridge predicted. So when first we were approached by Cambridge after a tutorial that we did to write a book, I was like, "I don't know, this is too small of a niche area."

中文翻译:

我感到惊喜的是，它的销量超过了我们的预期，也超过了剑桥大学出版社的预测。最初我们在做完一个教程后，剑桥出版社找我们写书，我当时想：“我不知道，这领域是不是太小众了。”

[00:50:47] Ronny Kohavi

English:

And they were saying, "So you'll be able to sell a few thousand copies and help the world." And I found my co-authors, which are great. And we wrote a book that we thought is not statistically oriented, has fewer formulas than you normally see, and focuses on the practical aspects and on trust, which is the key.

中文翻译:

他们说：“这样你就能卖出几千本，并帮助这个世界。”我找到了很棒的合著者。我们写了一本我们认为不完全以统计学为导向的书，公式比通常看到的要少，重点放在实践层面和“信任”上，而信任正是关键。

[00:51:10] Ronny Kohavi

English:

The book, as I said, was more successful. It sold over 20,000 copies in English. It was translated to Chinese, Korean, Japanese, and Russian. And so it's great to see that we help the world become more data-driven with experimentation, and I'm happy because of that. I was pleasantly surprised.

中文翻译:

正如我所说，这本书非常成功。英文版卖出了 2 万多册，还被翻译成了中文、韩文、日文和俄文。看到我们通过实验科学帮助世界变得更加数据驱动，我感到非常高兴。我感到惊喜。

[00:51:31] Ronny Kohavi

English:

By the way, all proceeds from the book are donated to charity. So if I'm pitching the book here, there is no financial gain for me from having more copies sold. I think we made that decision, which was a good decision. All proceeds go with the charity.

中文翻译:

顺便说一下，这本书的所有收益都捐给了慈善机构。所以如果我在这里推销这本书，多卖出一本对我个人也没有经济利益。我认为我们做了一个很好的决定，所有收益都归慈善机构。

[00:51:47] Lenny

English:

Amazing. I didn't know that. We'll link to the book in the show notes. Trust is in the title. You just mentioned how important trust is to experimentation. A lot of people talk about, "How do I run experiments faster?" You focus a lot on trust. Why is trust so important in running experiments?

中文翻译:

太棒了，我以前不知道。我们会在节目介绍里放上书的链接。书名里就有“信任”。你刚才提到信任对实验有多重要。很多人在谈论“如何更快地运行实验”，而你非常关注信任。为什么信任在运行实验中如此重要？

[00:52:03] Ronny Kohavi

English:

So to me, the experimentation platform is the safety net, and it's an oracle. So it serves really two purposes. The safety net means that if you launch something bad, you should be able to abort quickly, right? Safe deployments, safe velocity. There's some names for this. But this is one key value that the platform can give you.

中文翻译:

对我来说，实验平台既是“安全网”，也是“神谕”。它有两个主要用途。安全网意味着如果你发布了糟糕的东西，你应该能够快速终止，对吧？安全部署、安全速度，这些都是它的代名词。这是平台能提供的一个核心价值。

[00:52:25] Ronny Kohavi

English:

The other one, which is the more standard one, is at the end of the two-week experiment, we will tell you what happened to your key metric and to many of the other surrogate, and debugging, and guardrail metrics. Trust builds up, it's easy to lose.

中文翻译:

另一个用途更标准：在为期两周的实验结束时，我们会告诉你关键指标发生了什么变化，以及许多其他代理指标、调试指标和护栏指标的情况。信任是逐渐建立的，但也极易失去。

[00:52:43] Ronny Kohavi

English:

And so to me, it is very important that when you present this and say, "This is science, this is a controlled experiment, this is the resolve," you better believe that this is trustworthy.

中文翻译:

所以对我来说，当你展示结果并说“这是科学，这是受控实验，这是结论”时，你最好确信它是值得信赖的。

[00:52:57] Ronny Kohavi

English:

And so I focus on that a lot. I think it allowed us to gain the organizational trust that this is really... And the nice thing is when we built all this checks to make sure that the experiment is correct, if there were something wrong with it, we would stop and say, "Hey, something is wrong with the experiment."

中文翻译:

所以我非常关注这一点。我认为这让我们获得了组织的信任。好的一点是，当我们建立所有这些检查以确保实验正确时，如果实验出了问题，我们会停下来并说：“嘿，实验出错了。”

[00:53:17] Ronny Kohavi

English:

And I think that's something that some of the early implementations in other places did not do, and it was a big mistake. I've mentioned this in my book so I can mention this here.

中文翻译:

我认为其他地方早期的一些实施方案没有做到这一点，这是一个巨大的错误。我在书里提到过，所以在这里也可以提一下。

[00:53:28] Ronny Kohavi

English:

Optimizely in its early days were very statistically naive. They sort of said, "Hey, we're real time. We can compute your P values in real time," and then you can stop an experiment when the P value is statistically significant. That is a big mistake. That inflates your, what's called type one error or the false positive rate materially. So if you think you've got a 5% type one error, or you aim for that P value less than 0.05, using real time P value monitoring to optimize the offer, you would probably have a 30% error rate.

中文翻译:

Optimizely 在早期阶段在统计学上非常幼稚。他们大概是说：“嘿，我们是实时的，我们可以实时计算你的 P 值。”然后你就可以在 P 值达到统计显著时停止实验。这是一个巨大的错误。这会大幅增加所谓的“第一类错误”（type one error）或假阳性率。如果你认为你的第一类错误率是 5%，或者你的目标是 P 值小于 0.05，但你使用实时 P 值监测来优化方案，你的错误率可能会达到 30%。

[00:54:06] Ronny Kohavi

English:

So what this led is that people that started using Optimizely thought that the platform was telling them they were very successful. But when they actually started to see, "Well it told us this is positive revenue, but I don't see this over time. By now, we should have made double the money."

中文翻译:

这导致的结果是，开始使用 Optimizely 的人以为平台告诉他们非常成功。但当他们实际观察时会发现：“平台告诉我们收入是正向的，但随着时间的推移我并没看到增长。到现在为止，我们本该赚到两倍的钱了。”

[00:54:23] Ronny Kohavi

English:

So their questions started to come up around the trust in the platform. There's a very famous post that somebody wrote about how, "Optimizely almost got me fired," by a person who basically said, "Look, I came to the org. I said, 'We have all these successes.' But then I said, 'Something is wrong.'"

中文翻译:

于是他们开始质疑平台的信任度。有一篇非常著名的帖子叫《Optimizely 差点让我被解雇》，作者基本上是说：“看，我来到这个组织，我说‘我们取得了所有这些成功’，但后来我发现‘出问题了’。”

[00:54:40] Ronny Kohavi

English:

And he tells of how he ran an A/A test when there is no difference between the A and the B. And Optimizely told him that it was statistically significant too many times. Optimizely learned. Optimizely, several people pointed, I pointed this out in my Amazon review of the book that the authors wrote early on. I said, "Hey, you're not doing the statistics correctly."

中文翻译:

他讲述了他是如何运行 A/A 测试（即 A 组和 B 组没有区别）的，而 Optimizely 却太多次告诉他结果是统计显著的。Optimizely 后来吸取了教训。好几个人指出了这一点，我也在亚马逊上对作者早期写的书的评论中指出了这一点。我说：“嘿，你们的统计方法不对。”

[00:55:05] Ronny Kohavi

English:

Ramesh Johari at Stanford pointed this out, became a consultant to the company, and then they fixed it. But to me, that's a very good example of how to lose trust. They lost a lot of trust in the market. They lost all this trust because they built something that had very much inflated error rate.

中文翻译:

斯坦福大学的 Ramesh Johari 指出了这一点，并成为了该公司的顾问，然后他们修复了问题。但对我来说，这是一个如何失去信任的绝佳例子。他们在市场上失去了很多信任，就是因为他们构建了一个错误率极高的系统。

[00:55:26] Lenny

English:

That is pretty scary to think about you've been running all these experiments, and they weren't actually telling you accurate results. What are signs that what you're doing may not be valid if you're starting to run experiments? And then how do you avoid having that situation? What kind of tips can you share for people trying to run experiments?

中文翻译:

想到一直在运行这些实验，结果却不准确，确实挺可怕的。如果你开始运行实验，有哪些迹象表明你所做的可能无效？如何避免这种情况？你能为尝试运行实验的人分享哪些建议？

[00:55:47] Ronny Kohavi

English:

There's a whole chapter of that in my book, but I'll say one of the things that is the most common occurrence by far, which is a sample ratio mismatch. Now, what is a sample ratio mismatch?

中文翻译:

我的书里有一整章讲这个，但我会说目前为止最常见的一种情况，就是“样本比例不匹配”（Sample Ratio Mismatch，简称 SRM）。什么是样本比例不匹配？

[00:56:00] Ronny Kohavi

English:

If you design the experiment to send 50% of users to control and 50% of users to treatment, it's supposed to be a random number, or a hash function. If you get something off from 50%, it's a red flag.

中文翻译:

如果你设计的实验是将 50% 的用户分配到对照组，50% 分配到实验组，这本该是一个随机数或哈希函数的结果。如果你得到的结果偏离了 50%，那就是一个警示信号。

[00:56:15] Ronny Kohavi

English:

So let's take a real example. Let's say you're running an experiment, and it's large, it's got a million users, and you've got 50.2. So people say, "Well, I don't know. It's not going to be exactly the same as 50.2. Reasonable or not?" Well, there's a formula that you can plug in. I have a spreadsheet available for those that are interested, and you can tell, here's how many users are in control. Here's how many users have in treatment. My design was 50/50, and it tells you the probability that this could have happened by chance.

中文翻译:

举个真实的例子。假设你正在运行一个大型实验，有一百万用户，你得到的比例是 50.2%。人们会说：“嗯，我不知道，它不可能正好是 50%，50.2% 合理吗？”其实有一个公式可以套用。我为感兴趣的人提供了一个电子表格，你可以输入：对照组有多少用户，实验组有多少用户，我的设计是 50/50。它会告诉你这种情况偶然发生的概率。

[00:56:45] Ronny Kohavi

English:

Now in a case like this, you plug in the numbers, it might tell you that this should happen one in half a million experiments. Well, unless you've run half a million experiment, very unlikely that you would get a 50.2 versus 49.8 split. And therefore, something is wrong with the experiment.

中文翻译:

在这种情况下，你输入数字，它可能会告诉你这种情况在 50 万次实验中才会发生一次。那么，除非你已经运行了 50 万次实验，否则你极不可能得到 50.2% 对 49.8% 的分配。因此，实验肯定出问题了。

[00:57:06] Ronny Kohavi

English:

I remember when we implemented this check, we were surprised to see how many experiments suffered from this. Right? And there's a paper that was published, 2018, where we share that at Microsoft, even though we'd be running experiments for a while, is around 8% of experiments that suffered from the sample ratio mismatch.

中文翻译:

我记得当我们实施这项检查时，我们惊讶地发现竟然有这么多实验存在这个问题。2018 年发表的一篇论文显示，在微软，尽管我们已经运行实验很长时间了，仍有大约 8% 的实验存在样本比例不匹配的问题。

[00:57:29] Ronny Kohavi

English:

And it's a big number. I think about this. You're running 20,000 experiments a year. So many of them, 8% of them are invalid. And somebody has to go down and understand, what happened here? We know that

we can't trust the results, but why?

中文翻译:

这是一个很大的数字。想想看，你每年运行 2 万个实验，其中 8% 是无效的。必须有人去深入了解到底发生了什么。我们知道不能信任这些结果，但为什么？

[00:57:44] Ronny Kohavi

English:

So over time, you begin to understand there's something wrong with the data pipeline. There's something that happens with bots. Bots are a very common factor for causing sample ratio mismatch. So that paper that was published by my team talks about how to diagnose sample ratio mismatches.

中文翻译:

随着时间的推移，你开始明白可能是数据流水线出了问题，或者是机器人（bots）的影响。机器人是导致样本比例不匹配的一个非常常见的因素。我团队发表的那篇论文讨论了如何诊断样本比例不匹配。

[00:58:06] Ronny Kohavi

English:

In the last probably year and a half, it was amazing to see all these third party companies implement sample ratio mismatches, and all of them were reporting, "Oh my god, 6%, 8%, 10%." So it's sometimes fun to go back and say, how many of your results in the past were invalid before you had this sample ratio mismatched test?

中文翻译:

在过去的约一年半里，看到所有这些第三方公司都开始实施样本比例不匹配检查，真是令人惊叹，而且他们都在报告：“天哪，6%、8%、10%。”所以有时回头想想挺有意思的：在你们拥有这项测试之前，过去有多少结果是无效的？

[00:58:32] Lenny

English:

Yeah, that's frightening. Is the most common reason this happens is you're assigning users in the wrong place in your code?

中文翻译:

是的，这很吓人。发生这种情况最常见的原因是在代码中分配用户的位置不对吗？

[00:58:40] Ronny Kohavi

English:

So when you say most common, I think the most common is bots. Somehow, they hit the controller, the treatment in different proportions. Because you change the website, the bot may fail to parse the page, and try to hit it more often. And that's a classical example. Another one is just the data pipeline.

中文翻译:

当你问最常见的原因时，我认为最常见的是机器人。不知何故，它们以不同的比例访问对照组或实验组。因为你更改了网站，机器人可能无法解析页面，从而尝试更频繁地访问。这是一个经典的例子。另一个原因是数据流水线。

[00:58:58] Ronny Kohavi

English:

We've had cases where we were trying to remove bad traffic under certain conditions, and it was skewed because of the control and treatment. I've seen people that start an experiment in the middle of the site on some page, but they don't realize that some campaign is pushing people from the side.

中文翻译:

我们遇到过这样的情况：我们试图在某些条件下移除异常流量，但由于对照组和实验组的差异，结果产生了偏差。我也见过有人在网站中间的某个页面开始实验，但他们没意识到某个营销活动正从侧面把人引流过来。

[00:59:13] Ronny Kohavi

English:

So there's multiple reasons. It is surprising how often this happens. And I'll tell you a funny story, which is when we first added this test to the platform, we just put a banner say, "You have a sample ratio mismatch. Do not trust these results." And we noticed that people ignored it. They were starting to present results that had this banner.

中文翻译:

所以原因有很多。这种情况发生的频率之高令人惊讶。我给你讲个有趣的故事：当我们第一次在平台中加入这项测试时，我们只是放了一个横幅，写着：“你存在样本比例不匹配，请勿信任这些结果。”结果我们发现人们竟然忽略了它，他们开始展示带有这个横幅的结果。

[00:59:37] Ronny Kohavi

English:

And so we blanked out the scorecard. We put a big red, "Can't see this result. You have a sample ratio mismatch. Click to expose the results." And why we do we need that okay? We need that okay button because you want to be able to debug the reasons, and sometimes the metrics help you understand why you have a sample ratio mismatch.

中文翻译:

于是我们清空了计分卡。我们放了一个巨大的红色提示：“无法查看此结果。你存在样本比例不匹配。点击以显示结果。”为什么我们需要那个确认按钮？因为你需要能够调试原因，有时指标能帮你理解为什么会出现样本比例不匹配。

[01:00:00] Ronny Kohavi

English:

So we blanked out the scorecard, we had this button, and then we started to see that people pressed the button and still presented the results of experiments with sample ratio mismatch. And so we ended up

with an amazing compromise, which is every number in the scorecard was highlighted with a red line, so that if you took a screenshot, other people could tell you how to sample ratio mismatch.

中文翻译:

我们清空了计分卡，加了按钮，结果我们发现人们按下按钮后，依然在展示带有样本比例不匹配的实验结果。所以我们最终达成了一个惊人的妥协：计分卡中的每一个数字都用红线标出，这样如果你截图，别人一眼就能看出你存在样本比例不匹配。

[01:00:24] Lenny

English:

Freaking product managers.

中文翻译:

这些产品经理真是的。

[01:00:26] Ronny Kohavi

English:

This is intuition. People just say, "Well, my [inaudible 01:00:30] was small, therefore I can still present the results." People want to see success. I mean, this is a natural bias, and then we have to be very conscientious and fight that bias and say when something looks too good to be true, investigate.

中文翻译:

这就是直觉。人们会说：“嗯，我的偏差很小，所以我还是可以展示结果。”人们想看到成功。这是一种天生的偏见，我们必须非常自觉地对抗这种偏见，并说：当某件事看起来好得不真实时，去调查它。

[01:00:45] Lenny

English:

Which is a great segue to something you mentioned briefly, something called Twyman's law. Yeah. Can you talk about that?

中文翻译:

这正好引出了你简短提到过的“特怀曼法则”（Twyman's law）。你能谈谈吗？

[01:00:51] Ronny Kohavi

English:

Yeah. So Twyman's law, the general statement is if any figure that looks interesting or different is usually wrong. It was first said by this person in the UK who worked in radio media, but I'm a big fan of it. And my main claim to people is if the result looks too good to be true, your normal movement of an experiment is under 1% and you suddenly have a 10% movement, hold the celebratory dinner. It was just your first reaction, right? Let's take everybody to a fancy dinner, because we just improved revenue by millions of dollars. Hold that dinner, investigate, see, because there's a large probability that something is wrong

with the result. And I will say that nine out of 10, when we call it Twyman's law, it is the case that we find some flaw in the experiment.

中文翻译:

是的。特怀曼法则的一般表述是：任何看起来有趣或与众不同的数据通常都是错误的。它最初是由英国一位从事无线电媒体工作的人提出的，但我非常推崇它。我对人们的主要主张是：如果结果看起来好得不真实——比如你平时的实验波动都在 1% 以下，突然出现了一个 10% 的波动——先别急着开庆祝晚宴。那是你的第一反应，对吧？“带大家去吃顿大餐，因为我们刚把收入提高了数百万美元。”先别吃，去调查，去核实，因为极大概率是结果出错了。我会说，十次中有九次，当我们引用特怀曼法则时，确实能发现实验中的某种缺陷。

[01:01:45] Ronny Kohavi

English:

Now there are obviously outliers. That first experiment that I shared where we promoted that made long titles, that was successful. But that was replicated multiple times, and double and triple checked, and everything was good about it. Many other results that were so big turn out to be false. So I'm a big fan of Twyman's law. There's a deck, I could also give this in the note, where I shared some real examples of Twyman's law.

中文翻译:

当然也有例外。我分享的第一个实验——提升标题长度的那个——就是成功的。但那是经过多次重复实验、双重甚至三重检查后确认没问题的。许多其他巨大的结果最终都被证明是错误的。所以我非常推崇特怀曼法则。我有一份幻灯片，也可以放在节目介绍里，里面分享了一些特怀曼法则的真实案例。

[01:02:14] Lenny

English:

Amazing. I want to talk about rolling this out of companies and things that you run into that fail. But before I get to that, I'd love for you to explain P value. I know that people kind of misunderstand it, and this might be a good time to just help people understand, what is it actually telling you, P value of say 0.05?

中文翻译:

太棒了。我想谈谈在公司推广这些方法以及你遇到的失败案例。但在那之前，我想请你解释一下 P 值（P value）。我知道人们对它有些误解，现在也许是帮助大家理解的好时机：比如 P 值为 0.05 到底告诉了我们什么？

[01:02:30] Ronny Kohavi

English:

I don't know if this is the right forum for explaining P values, because the definition of a P value is simple. What it hides is very complicated. So I'll say one thing, which is many people assign one minus P value as the probability that your treatment is better than control. So you ran an experiment, you got a P value of 0.02. They think there's a 98% probability that the treatment is better than the control. That is wrong. So rather than defining P values, I want to caution everybody that the most common interpretation is incorrect.

中文翻译:

我不知道这里是不是解释 P 值的合适场合，因为 P 值的定义很简单，但它背后隐藏的东西非常复杂。我会说一点：许多人将“1 减去 P 值”视为实验组优于对照组的概率。比如你运行了一个实验，P 值是 0.02，他们就认为有 98% 的概率实验组优于对照组。那是错误的。所以与其定义 P 值，我更想提醒大家：最常见的解释是不正确的。

[01:03:08] Ronny Kohavi

English:

P value assumes, it's a conditional probability or an assumed probability. It assumes that the null hypothesis is true. And we're computing the probability that the data we're seeing matches the hypothesis, this null hypothesis.

中文翻译:

P 值假设——它是一个条件概率或假设概率。它假设“原假设”（null hypothesis）为真。我们计算的是：我们看到的数据与这个原假设相匹配的概率。

[01:03:27] Ronny Kohavi

English:

In order to get the probability that most people want, we need to apply Bayes' rules and invert the probability from the probability of the data given the hypothesis, to the probability of the hypothesis given the data. For that, we need an additional number, which is the probability, the prior probability that the hypothesis that you're testing is successful or not.

中文翻译:

为了得到大多数人想要的那个概率，我们需要应用贝叶斯法则，将“给定假设下出现该数据的概率”反转为“给定数据下该假设成立的概率”。为此，我们需要一个额外的数字，即你正在测试的假设成功的“先验概率”（prior probability）。

[01:03:49] Ronny Kohavi

English:

That's an unknown. What we do is we can take historical data and say, "Look, people fail two thirds of the time or 80% of the time." And we can apply that number and compute that. We've done that in a paper that I will give in the notes, so that you can assess the number that you really want, what's called a false positive risk.

中文翻译:

这是一个未知数。我们能做的是参考历史数据，比如：“看，人们 2/3 或 80% 的时间都会失败。”我们可以应用这个数字来计算。我们在节目介绍里的一篇论文中做过这个计算，这样你就可以评估你真正想要的那个数字，即所谓的“假阳性风险”（false positive risk）。

[01:04:10] Ronny Kohavi

English:

So I think that's something for people to internalize, that what you really want to look at is this false positive risk, which tends to be much, much higher than the 5% that people think, right? So I think the classical example in the Airbnb where the failure rate was very, very high, is that when you get a statistically significant result, let me actually pull the note so that I know the actual number. If you're at Airbnb, or Airbnb search where the success rate is only 8%, if you get a statistically significant result with a P value less than 0.05, there is a 26% chance that this is a false positive result. It's not 5%, it's 26%.

中文翻译:

所以我认为人们需要内化这一点：你真正应该关注的是假阳性风险，它往往比人们认为的 5% 要高得多。以 Airbnb 为例，那里的失败率非常高。在 Airbnb 搜索团队，成功率只有 8%，如果你得到一个 P 值小于 0.05 的统计显著结果——让我查一下笔记看准确数字——实际上有 26% 的概率这是一个假阳性结果。不是 5%，而是 26%。

[01:04:54] Ronny Kohavi

English:

So that's the number that you should have in your mind. And that's why when I worked at Airbnb, one of the things we did is we said, "Okay, if you're less than 0.05, but above 0.01, rerun, replicate." When you replicate, you can combine the two experiments, and get a combined P value using something called Fisher's method or Stouffer's method, and that gives you the joint probability. And that's usually much, much lower. So if you get two 0.5's or something like that, then the probability that you've got them is much, much lower.

中文翻译:

所以这才是你应该记在脑子里的数字。这就是为什么我在 Airbnb 工作时，我们采取的一种做法是：“好吧，如果 P 值小于 0.05 但大于 0.01，那就重新运行，重复实验。”当你重复实验时，你可以合并两个实验，并使用费舍尔方法（Fisher's method）或斯托弗方法（Stouffer's method）获得合并后的 P 值，这会给你联合概率。这通常会低得多。如果你得到两个 0.05 左右的结果，那么它们同时发生的概率就会低得多。

[01:05:26] Lenny

English:

Wow, I've never heard it described that way. It makes me think about how even data scientists in our teams are always just like, "This isn't perfect. We're not 100% sure this experiment is positive." But on balance, if we're launching positive experiments, we're probably doing good things. It's okay if sometimes we're wrong.

中文翻译:

哇，我从来没听过这种描述。这让我想起，即使是我们团队的数据科学家也总是说：“这并不完美，我们不能 100% 确定这个实验是正向的。”但总的来说，如果我们发布的是正向实验，我们可能就是在做正确的事。偶尔出错也没关系。

[01:05:42] Ronny Kohavi

English:

By the way, it's true. On balance, you're probably better than 50/50, but people don't appreciate how much that 26% that I mentioned is high. And the reason that I want to be sure is that I think it leads to this

idea of the learning, the institutional knowledge. What you want to be able to say is share with the org's success. And so you want to be really sure that you're successful. So by lowering the P value, by forcing teams to work with the P value maybe below 0.01 and do replication on higher, then you can be much more successful, and the false positive rate will be much, much lower.

中文翻译:

顺便说一下，确实如此。总的来说，你可能比 50/50 的概率要好，但人们没有意识到我提到的 26% 有多高。我之所以想要确定，是因为我认为这关系到学习和机构知识。你希望能够向组织分享成功经验，所以你必须非常确定你是成功的。因此，通过降低 P 值要求（比如强迫团队使用低于 0.01 的 P 值），或者对较高的 P 值进行重复实验，你可以变得更成功，假阳性率也会低得多。

[01:06:20] Lenny

English:

Fascinating. And also shows the value of keeping track of what percentage your experiments are failing historically at the company or within that specific product. Say someone listening wants to start running experiments, say they have tens of thousands of users at this point. What would be the first couple steps you'd recommend?

中文翻译:

很有意思。这也展示了追踪公司或特定产品历史上实验失败比例的价值。假设听众中有人想开始运行实验，且目前拥有数万名用户。你会推荐最开始的几个步骤是什么？

[01:06:38] Ronny Kohavi

English:

Well, so if they have somebody in the org that has previously been involved with a experiment, that's a good way to consult internally. I think the key decision is whether you want to build or buy. There's a whole series of eight sessions that I posted on LinkedIn where I invited guest speakers to talk about this problem. So if people are interested, they can look at what the vendors say and what agency said about build versus buy question. And it's usually not a zero one, it's usually both. You build some and you buy some, and it's a question of do you build 10% or do you build in 90%?

中文翻译:

如果组织里有人以前参与过实验，那是内部咨询的好办法。我认为关键决策是“自建还是购买”。我在 LinkedIn 上发布了由八个环节组成的系列内容，邀请了嘉宾来讨论这个问题。如果大家感兴趣，可以看看供应商和代理机构对“自建 vs 购买”问题的看法。这通常不是非黑即白的，通常是两者结合。你自建一部分，购买一部分，问题在于你是自建 10% 还是自建 90%?

[01:07:17] Ronny Kohavi

English:

I think for people starting, the third party products that are available today are pretty good. This wasn't the case when I started working. So when I started running experiments at Amazon, we were building the platform because nothing existed. Same at Microsoft. I think today, there's enough vendors that provide good experimentation platforms that are trustworthy, that I would say not a good way to consider using one of those.

中文翻译:

我认为对于刚起步的人来说，现在的第三方产品已经相当不错了。我刚开始工作时可不是这样。我在亚马逊开始运行实验时，我们必须自建平台，因为当时什么都没有。在微软也是如此。我认为今天已经有足够的供应商提供值得信赖的优秀实验平台，我会说考虑使用其中之一是个不错的选择。

[01:07:44] Lenny

English:

Say you're at a company where there's resistance to experimentation and A/B testing, whether it's a startup or a bigger company. What have you found works in helping shift that culture, and how long does that usually take, especially at a larger company?

中文翻译:

假设你在一家对实验和 A/B 测试有抵触情绪的公司，无论是初创公司还是大公司。你发现哪些方法能有效转变这种文化？这通常需要多长时间，特别是在大公司？

[01:07:57] Ronny Kohavi

English:

My general experience is with Microsoft, where we went with this beach head of Bing. We were running a few experiments and then we were asked to focus on Bing, and we scaled experimentation and built a platform at scale at Bing.

中文翻译:

我的主要经验来自微软，当时我们以 Bing 作为“滩头阵地”。我们先运行了几个实验，然后被要求专注于 Bing，我们在 Bing 大规模推广了实验并构建了大规模平台。

[01:08:13] Ronny Kohavi

English:

Once Bing was successful and we were able to share all these surprising results, I think that many, many more people in the company were amenable. It was also the case that helped a lot that, there's a usual cross pollination. People from Bing move out to other groups, and that helped these other groups say, "Hey, there's a better way to build software."

中文翻译:

一旦 Bing 取得了成功，并且我们能够分享所有这些令人惊讶的结果，我认为公司里就有越来越多的人愿意接受了。此外，人才的跨部门流动（cross pollination）也起到了很大作用。Bing 的人去了其他小组，这帮助其他小组意识到：“嘿，有一种更好的构建软件的方法。”

[01:08:34] Ronny Kohavi

English:

So I think if you're starting out, find a place, find a team where experimentation is easy to run. And by that, I mean they're launching often, right? Don't go with the team that launches every six months, or Office used to launch every three years. Go with the team that launches frequently. They're running on

sprints, they launch every week or two. Sometimes they launch daily. I mean, Bing used to launch multiple times a day.

中文翻译:

所以如果你刚起步，找一个容易运行实验的地方或团队。我的意思是，找那些发布频繁的团队。不要去找每六个月发布一次的团队，或者像以前每三年发布一次的 Office 团队。找那些频繁发布的团队，他们按冲刺（sprint）运行，每周或每两周发布一次，有时甚至每天发布。Bing 以前每天发布好几次。

[01:08:59] Ronny Kohavi

English:

And then make sure that you understand the question of the OEC. Is it clear what they're optimizing for? There are some groups where you can come up with a good OEC. Some groups are harder.

中文翻译:

然后确保你理解 OEC 的问题。他们优化的目标明确吗？有些小组你可以提出很好的 OEC，有些小组则比较难。

[01:09:11] Ronny Kohavi

English:

I remember one funny example was the microsoft.com website, which this is not MSN, this is microsoft.com, has multiple different constituencies that are trying to determine this is a support site, and this is the ability to sell software through this site, and warn you about safety and updates. It has so many goals. I remember when the team said, "We want to run experiments," and I brought the group in and some of the managers and I said, "Do you know what you're optimizing for?"

中文翻译:

我记得一个有趣的例子是 microsoft.com 网站（不是 MSN，是微软官网）。它有多个不同的利益相关方，有的认为这是支持网站，有的认为这是卖软件的网站，有的认为这是提醒安全和更新的网站。它有太多目标。我记得当团队说“我们想运行实验”时，我把小组和一些经理召集起来问：“你们知道自己在优化什么吗？”

[01:09:47] Ronny Kohavi

English:

It was very funny because they surprised me. They said, "Hey Ronny, we read some of your papers. We know there's this term called OEC. We decided the time on site is our OEC." And I said, "Wait a minute. Some of your main goals is support site. Is people spending more time on the support site a good thing or a bad thing?" And then half the room thought that more time is better, and half the room thought that more time is worse. So an OEC is bad if directionally, you can't agree on it.

中文翻译:

非常有趣，因为他们让我大吃一惊。他们说：“嘿 Ronny，我们读过你的一些论文，我们知道 OEC 这个词。我们决定把‘页面停留时间’作为我们的 OEC。”我说：“等一下。你们的主要目标之一是提供支持。人们在支持页面停留更长时间是好事还是坏事？”结果房间里一半的人认为时间越长越好，另一半人认为时间越长越糟。所以，如果大家在方向上无法达成一致，那么这个 OEC 就是糟糕的。

[01:10:18] Lenny

English:

That's a great tip. Along these same lines, I know you're a big fan of platforms and building a platform to run experiments, versus just one-off experiments. Can you just talk briefly about that to give people a sense of where they probably should be going with their experimentation approach?

中文翻译:

这是一个很好的建议。沿着这个思路，我知道你是平台的坚定支持者，主张构建平台来运行实验，而不是只做一次性实验。你能简要谈谈吗，让大家了解他们的实验方法应该朝什么方向发展？

[01:10:32] Ronny Kohavi

English:

Yeah, so I think the motivation is to bring the marginal cost of experiments down to zero. So the more you self-service, go to a website, set up your experiment, define your targets, define the metrics that you want, right? People don't appreciate that the number of metrics starts to grow really fast if you're doing things right. At Bing, you could define 10,000 metrics that you wanted to be in your scorecard. Big numbers.

中文翻译:

是的，我认为动力在于将实验的边际成本降至零。所以你应该更多地采用自助服务：去网站设置实验、定义目标、定义你想要的指标。如果方法正确，指标的数量会增长得非常快。在 Bing，你可以在计分卡中定义 1 万个指标。规模非常大。

[01:11:02] Ronny Kohavi

English:

So it was so big, and people said it's computationally inefficient. We broke them into templates so that if you were launching a UI experiment, you would get this set of 2,000. If you were doing a revenue experiment, you would get this set of 2,000.

中文翻译:

因为规模太大，人们说计算效率太低。于是我们把它们分成模板：如果你发布的是 UI 实验，你会得到这 2000 个指标；如果你做的是收入实验，你会得到另外 2000 个。

[01:11:15] Ronny Kohavi

English:

So the point was build a platform that can quickly allow you to set up and run an experiment, and then analyze it. I think one of the things that I will say at Airbnb is the analysis was relatively weak, and so lots of data scientists were hired to be able to compensate for the fact that the platform didn't do enough.

中文翻译:

所以重点是构建一个能让你快速设置、运行并分析实验的平台。我想说的是，在 Airbnb，分析功能相对较弱，所以雇佣了大量数据科学家来弥补平台功能的不足。

[01:11:36] Ronny Kohavi

English:

And this happens in other organizations too, where there's this trade-off. If you're building a good platform, invest in it so that more and more automation will allow people to look at the analysis, without the need to involve a data scientist.

中文翻译:

这在其他组织中也会发生，这是一种权衡。如果你在构建一个好的平台，就投入其中，让越来越多的自动化功能允许人们查看分析结果，而无需数据科学家的参与。

[01:11:53] Ronny Kohavi

English:

We published a paper. Again, I'll give it in the notes with this nice matrix of six axes, and how you move from crawl, to walk, to run, to fly, and what you need to build on those six axes. So one of the things that I do sometimes when I consult is I go into the org and say, "Where do you think you are on these six axes?" And that should be the guidance for what are the things you need to do next.

中文翻译:

我们发表过一篇论文，我也会放在节目介绍里。里面有一个包含六个维度的矩阵，描述了你如何从“爬行”到“行走”，再到“奔跑”和“飞行”，以及在这些维度上你需要构建什么。所以我做咨询时经常做的一件事就是问：“你认为自己在这六个维度上处于什么位置？”这应该成为你下一步行动的指南。

[01:12:21] Lenny

English:

This is going to be the most epic show notes episode we've had yet. Maybe a last question. We talked about how important trust is to running experiments, and how even though people talk about speed, trust ends up being most important. Still, I want to ask you about speed. Is there anything you recommend for helping people run experiments faster and get results more quickly that they can implement?

中文翻译:

这绝对会是节目介绍最丰富的一期。最后一个问题。我们谈到了信任对运行实验有多重要，以及尽管人们在谈论速度，但信任最终才是最重要的。不过，我还是想问问关于速度的问题。你有什么建议能帮助人们更快地运行实验并更快地获得可实施的结果吗？

[01:12:40] Ronny Kohavi

English:

Yeah, so I'll say a couple of things. One is if your platform is good, then when the experiment finishes, you should have a scorecard soon after. Maybe takes a day, but it shouldn't be that you have to wait a week for the data scientist. To me, this is the number one way to speed up things.

中文翻译:

是的，我会说两点。一是如果你的平台足够好，那么当实验结束时，你应该很快就能得到计分卡。也许需要一天，但不应该让你等数据科学家一个星期。对我来说，这是加速的第一要务。

[01:13:00] Ronny Kohavi

English:

Now, in terms of using the data efficiently, there are mechanisms out there under the title of variance reduction that help you reduce the variance of metrics so that you need less users, so that you can get results faster. Some examples that you might think about are capping metrics. So if your revenue metric is very skewed, maybe you say, "Well, if somebody purchased over $1,000, let's make that $1,000." At Airbnb, one of the key metrics for example, is nights booked.

中文翻译:

其次，在高效使用数据方面，有一些被称为“方差缩减”（variance reduction）的机制，可以帮助你减少指标的方差，从而需要更少的用户，更快地获得结果。你可以考虑的一些例子包括“指标封顶”（capping metrics）。如果你的收入指标非常偏态，也许你可以说：“如果有人购买超过 1000 美元，我们就按 1000 美元计算。”例如在 Airbnb，一个关键指标是预订间夜数。

[01:13:30] Ronny Kohavi

English:

Well, it turns out that some people book tens of nights. They're like an agency or something, hundreds of nights. You may say, "Okay, let's just cap this. It's unlikely that people book more than 30 days in a given month." So that various reduction technique will allow you to get statistically significant results faster.

中文翻译:

结果发现，有些人会预订几十晚，他们可能像中介之类的，甚至预订几百晚。你可以说：“好吧，让我们给这个指标封顶。人们在一个月内预订超过 30 天是不太可能的。”这种方差缩减技术将允许你更快地获得统计显著的结果。

[01:13:53] Ronny Kohavi

English:

And a third technique is called cupid, which is an article that we published. Again, I can give it in the notes, which uses the pre-experiment data to adjust the result. And we can show that you get the result as unbiased, but with lower variance, and hence, it requires fewer users.

中文翻译:

第三种技术叫做 CUPED，这是我们发表过的一篇文章。我也可以把它放在节目介绍里。它利用实验前的数据来调整结果。我们可以证明，你得到的结果是无偏的，但方差更低，因此需要的用户更少。

[01:14:11] Lenny

English:

Ronny, is there anything else you want to share before we get to our very exciting lightning round?

中文翻译:

Ronny，在进入精彩的闪电轮提问之前，你还有什么想分享的吗？

[01:14:15] Ronny Kohavi

English:

No, I think we've asked a lot of good questions. Hope people enjoy this.

中文翻译:

没有了，我认为我们讨论了很多好问题。希望大家喜欢。

[01:14:20] Lenny

English:

I know they will.

中文翻译:

我知道他们会的。

[01:14:21] Ronny Kohavi

English:

Lightning round.

中文翻译:

闪电轮开始。

[01:14:22] Lenny

English:

Lightning round. Here we go. I'm just going to roll right into it. What are two or three books that you've recommended most to other people?

中文翻译:

闪电轮，我们开始。我直接切入正题。你向别人推荐最多的两三本书是什么？

[01:14:29] Ronny Kohavi

English:

There's a fun book called Calling Bullshit, which despite the name, which is a little extreme, I think, for a title, it actually has a lot of amazing insights that I love. And it sort of embodies, in my opinion, a lot of the Twyman's law showing that things that are too extreme, your bullshit meter should go up and say, "Hey, I don't believe that." So that's my number one recommendation.

中文翻译:

有一本有趣的书叫《拆穿胡说八道》（Calling Bullshit），尽管书名听起来有点极端，但它确实有很多我喜欢的惊人见解。在我看来，它体现了很多特怀曼法则的精神，表明当事情太极端时，你的“胡说八道探测器”就该

响了，并说：“嘿，我不相信那个。”这是我的首选推荐。

[01:14:57] Ronny Kohavi

English:

There's a slightly older book that I love called Hard Facts, Dangerous Half-Truths And Total Nonsense by the Stanford professors from the Graduate School of Business. Very interesting to see many of the things that we grew up with as well understood turn out to have no justification.

中文翻译:

还有一本我喜欢的稍旧一点的书，叫《事实、半真半假的谎言与胡言乱语》（Hard Facts, Dangerous Half-Truths And Total Nonsense），由斯坦福商学院的教授编写。看到许多我们从小就认为理所当然的事情其实毫无根据，这非常有趣。

[01:15:21] Ronny Kohavi

English:

So a stranger book, which I love, sort of on the verge of psychology, it's called Mistakes Were Made (But Not by Me), about all the fallacies that we fall into, and the humbling results from that.

中文翻译:

还有一本更奇特的书，我也很喜欢，有点偏向心理学，叫《错不在我》（Mistakes Were Made (But Not by Me)），讲的是我们容易陷入的所有谬误，以及由此产生的让人清醒的结果。

[01:15:37] Lenny

English:

The titles of these are hilarious, and there's a common theme across all these books. Next question, what is a favorite recent movie or TV show?

中文翻译:

这些书名都很有趣，而且都有一个共同的主题。下一个问题，最近最喜欢的电影或电视节目是什么？

[01:15:47] Ronny Kohavi

English:

So I recently saw a short series called Chernobyl, the disaster. I thought it was amazingly well done. Highly recommended it, based on true events. As usual, there's some freedom for the artistic movie. It was kind of interesting at the end, they say, "This woman in the movie wasn't really a woman. It was a bunch of 30 data scientists." Not data scientists, 30 scientists that in real life, presented all the data to the leadership of what to do.

中文翻译:

我最近看了一部短剧叫《切尔诺贝利》（Chernobyl）。我觉得拍得非常好，强烈推荐，它是根据真实事件改编的。像往常一样，电影艺术会有一些虚构成分。结尾很有趣，他们说：“电影里的那个女性角色在现实中并不

是一个人，而是 30 名科学家。”不是数据科学家，是 30 名科学家，他们在现实中向领导层提交了关于该怎么做的所有数据。

[01:16:22] Lenny

English:

I remember that. Fun fact, I was born in Odessa, Ukraine, which was not so far from Chernobyl. And I remember my dad told me he had to go to work. They called him into work that day to clean some stuff off the trees. I think ash from the explosion or something. It was far away where I don't think we were exposed, but we were in the vicinity. That's pretty scary. My wife, every time something's wrong with me, she's like, "That must be a Chernobyl thing." Okay, next question. Favorite interview question you like to ask people when you're interviewing them?

中文翻译:

我记得那个。有趣的是，我出生在乌克兰的敖德萨，离切尔诺贝利不远。我记得我爸告诉我他那天必须去上班，他们叫他去清理树上的一些东西，我想是爆炸产生的灰烬之类的。虽然离得很远，我不认为我们受到了辐射，但我们确实在附近。挺吓人的。我妻子每次看我哪里不舒服，都会说：“那一定是切尔诺贝利后遗症。”好，下一个问题。面试别人时你最喜欢问的问题是什么？

[01:16:56] Ronny Kohavi

English:

So it depends on the interview, but when I do a technical interview, which I do less of, but one question that I love that it's amazing how many people it throws away for languages like C++, is tell me what the static qualifier does. And for multiple, you can do it for a variable, you can do it for function. And it is just amazing that I would say more than 50% of people that interview for engineering job cannot get this, and get it awfully wrong.

中文翻译:

这取决于面试类型。当我做技术面试时（虽然现在做得少了），我喜欢问一个关于 C++ 等语言的问题，惊讶的是这个问题能刷掉很多人：告诉我 `static` 限定符的作用是什么？它可以用于变量，也可以用于函数。令人惊讶的是，我想说超过 50% 面试工程岗位的人都答不上来，或者错得离谱。

[01:17:31] Lenny

English:

Definitely the most technical answer to this question yet.

中文翻译:

这绝对是目前为止最技术性的回答。

[01:17:34] Ronny Kohavi

English:

Very technical, yeah.

中文翻译:

非常技术化，是的。

[01:17:34] Ronny Kohavi

English:

I love it.

中文翻译:

我喜欢。

[01:17:36] Lenny

English:

What's a favorite recent product you've discovered that you love?

中文翻译:

你最近发现并喜爱的产品是什么？

[01:17:39] Ronny Kohavi

English:

Blink cameras. So a Blink camera is this small camera. You stick in two AA batteries, and it lasts for about six months. They claim up to two years. My experience is usually about six months. But it was just amazing to me how you can throw these things around in the yard and see things that you would never know otherwise. Some animals that go by. We had a skunk that we couldn't figure out how he was entering, so I threw five cameras out and I saw where he came in.

中文翻译:

Blink 摄像头。Blink 摄像头是一种很小的摄像头，装两节五号电池就能用大约六个月。他们声称能用两年，但我的经验通常是六个月。让我惊讶的是，你可以把这些东西随处放在院子里，看到你平时根本不会知道的事情。比如路过的动物。我们曾有一只臭鼬，我们搞不清楚它是怎么进来的，于是我放了五个摄像头，最后看到了它进来的地方。

[01:18:18] Lenny

English:

Where'd he come in?

中文翻译:

它从哪儿进来的？

[01:18:19] Ronny Kohavi

English:

He came in under a hole on the fence that was about this high. I have a video of this thing just squishing underneath. We never would've assumed that it came from there, from the neighbor. But yeah, these

things have just changed. And when you're away on a trip, it's always nice to be able to say, "I can see my house. Everything's okay." At one point, we had a false alarm, and the cops came in and had this amazing video of how they're entering the house and pulling the guns out.

中文翻译:

它从篱笆下面一个大概这么高的洞钻进来的。我有一段视频，拍到它就那样挤了过去。我们从来没想过它会从邻居那边的那个地方进来。但是，是的，这些东西改变了生活。当你外出旅行时，能看到“我能看到我的房子，一切都好”总是很棒。有一次我们遇到了误报，警察进来了，我拍到了他们进入房子并拔出枪的惊人视频。

[01:18:56] Lenny

English:

You got to share that on TikTok. That's good content. Wow. Okay. Blink cameras. We'll set those up in my house asap.

中文翻译:

你得把那个发到 TikTok 上，那是很好的素材。哇，好，Blink 摄像头，我也要尽快在家里装上。

[01:19:04] Ronny Kohavi

English:

Yes.

中文翻译:

是的。

[01:19:06] Lenny

English:

What is something relatively minor you've changed in the way your teams develop product, that has had a big impact on their ability to execute?

中文翻译:

在团队开发产品的方式上，你做过哪些相对较小但对执行能力产生重大影响的改变？

[01:19:14] Ronny Kohavi

English:

I think this is something that I learned at Amazon, which is a structured narrative. So Amazon has some variance of this, which sometimes go by the name of a six pager or something. But when I was at Amazon, I still remember that email from Jeff, which is, "No more PowerPoint. I'm going to force you to write a narrative."

中文翻译:

我想这是我在亚马逊学到的东西，即“结构化叙述”（structured narrative）。亚马逊有几种不同的形式，有时被称为“六页纸报告”（six pager）之类的。我在亚马逊时，还记得杰夫（贝佐斯）发的那封邮件：“不再使用 PowerPoint，我要强迫你们写叙述性文档。”

[01:19:34] Ronny Kohavi

English:

I took that to heart. And many of the features that the team presented instead of a PowerPoint, you start off with a structured document that tells you what you need, the questions you need to answer for your idea. And then we review them as a team.

中文翻译:

我牢记在心。团队展示的许多功能不再使用 PowerPoint，而是从一份结构化文档开始，说明你的需求以及你的想法需要回答的问题。然后我们作为一个团队进行评审。

[01:19:51] Ronny Kohavi

English:

And Amazon, these were paper-based. Now it's all based on Word or Google Docs where people comment, and I think the impact of that was amazing. I think the ability to give people honest feedback and have them appreciate, and have it stay after the meeting in these notes on the document, just amazing.

中文翻译:

在亚马逊，这些最初是纸质的。现在都是基于 Word 或 Google Docs，人们可以在上面评论。我认为这种影响是惊人的。能够给人们诚实的反馈，让他们理解，并让这些反馈在会议结束后依然留在文档的注释中，这太棒了。

[01:20:13] Lenny

English:

Final question, have you ever run an A/B test on your life, either your dating life, your family, your kids? And if so, what did you try?

中文翻译:

最后一个问题，你有没有在生活中运行过 A/B 测试？比如约会生活、家庭或孩子？如果有，你尝试了什么？

[01:20:21] Ronny Kohavi

English:

So there aren't enough units. Remember I said you need 10,000 of something to run true A/B tests? I will say a couple of things. One is I try to emphasize to my family, and friends, and everybody, this idea called the hierarchy of evidence. When you read something, there's a hierarchy of trust levels. If something is anecdotal, don't trust it. If there was an experiment, it was observational. Give it some bit of trust. As you get more up and up to a natural experiment, and controlled experiments, and multiple controlled

experiments, your trust levels should go up. So I think that that's a very important thing that a lot of people miss when they see something in the news is, where does it come from?

中文翻译:

样本量不够。记得我说过需要 1 万个样本才能运行真正的 A/B 测试吗？我会说两点。一是我尝试向我的家人、朋友和所有人强调“证据层级”（hierarchy of evidence）这个概念。当你读到某些东西时，信任程度是有层级的。如果是轶事传闻，不要相信；如果是观察性实验，给予一定的信任；随着层级上升到自然实验、受控实验以及多个受控实验，你的信任程度应该随之提高。我认为很多人在新闻中看到某些东西时会忽略这一点：它到底来自哪里？

[01:21:06] Ronny Kohavi

English:

I have a talk that I've shared of all these observational studies that people made that were published. And then somehow, a control experiment was run later on and proved that it was directionally incorrect. So I think there's a lot to learn about this idea of the hierarchy of evidence, and share it with our family, and kids, and friends. I think there's a book that's based on this. It's like How to Read a Book.

中文翻译:

我分享过一个演讲，列举了所有那些已发表的观察性研究，结果后来运行的受控实验证明它们在方向上是错误的。所以我认为关于“证据层级”有很多值得学习的地方，并可以分享给家人、孩子和朋友。我想有一本书就是基于这个理念的，类似于《如何阅读一本书》。

[01:21:34] Lenny

English:

Well, Ronny, the experiment of us recording a podcast I think is 100% positive P value 0.0. Thank you so much for being here.

中文翻译:

好了，Ronny，我认为我们录制这期播客的实验是 100% 正向的，P 值为 0.0。非常感谢你的到来。

[01:21:44] Ronny Kohavi

English:

Thank you so much for inviting me and for great questions.

中文翻译:

非常感谢你的邀请和这些精彩的问题。

[01:21:47] Lenny

English:

Amazing. I appreciate that. Two final questions. Where can folks find you online if they want to reach out, and is there anything that listeners can do for you?

中文翻译:

太棒了，我很感激。最后两个问题。如果大家想联系你，可以在哪里找到你？听众能为你做些什么吗？

[01:21:55] Ronny Kohavi

English:

Finding me online is easy. It's LinkedIn. And what can people do for me? Understand the idea of control experiments as a mechanism to make the right data-driven decisions. Use science. Learn more by reading my book if you want. Again, all proceeds go to charity. And if you want to learn more, there's a class that I teach every quarter on Maven. We'll put in the notes how to find it, and some discount for people who managed to stay all the way to the end of this podcast.

中文翻译:

在网上找我很简单，就是 LinkedIn。大家能为我做些什么？理解受控实验作为做出正确数据驱动决策的机制。运用科学。如果愿意，可以通过阅读我的书了解更多，再次强调，所有收益都捐给慈善机构。如果你想进一步学习，我在 Maven 上每季度都有一门课。我们会在节目介绍里说明如何找到它，并为坚持听到播客最后的人提供一些折扣。

[01:22:31] Lenny

English:

Yeah, that's awesome. We'll include that at the top so people don't miss it, so there's going to be a code to get a discount on your course. Ronny, thank you again so much for being here. This was amazing.

中文翻译:

太好了。我们会把它放在开头，这样大家就不会错过了，会有一个课程折扣码。Ronny，再次感谢你的到来，这太精彩了。

[01:22:39] Ronny Kohavi

English:

Thank you so much.

中文翻译:

非常感谢。

[01:22:40] Lenny

English:

Bye everyone. Thank you so much for listening. If you found this valuable, you can subscribe to the show on Apple Podcasts, Spotify, or your favorite podcast app. Also, please consider giving us a rating or leaving a review, as that really helps other listeners find the podcast. You can find all past episodes or learn more about the show at lennyspodcast.com. See you in the next episode.

中文翻译:

大家再见。非常感谢收听。如果你觉得这期节目有价值，可以在 Apple Podcasts、Spotify 或你喜欢的播客应用上订阅。此外，请考虑给我们评分或留下评论，这能帮助其他听众发现这个播客。你可以在

lennyspodcast.com 找到所有往期节目或了解更多信息。下期节目见。