Оглавление

Вступление

Получить числа легко; получить числа, которым можно *доверять*, сложно. Это практическое руководство от ведущих специалистов по экспериментам в Google, LinkedIn и Microsoft научит вас, как ускорить внедрение инноваций, используя *доверительные контролируемые онлайн-эксперименты* (trustworthy online controlled experiments), или, как их чаще называют, *A/B-тесты*. Основываясь на практическом опыте компаний, каждая из которых проводит более 20 000 контролируемых экспериментов в год, авторы делятся примерами, советами и предостережениями со студентами и профессионалами отрасли, приступающими к экспериментам, а также детально раскрывают сложные темы для опытных практиков, которые хотят усовершенствовать процесс принятия решений на основе экспериментальных данных в своей организации.

Прочитав эту книгу, вы научитесь:

- использовать научные методы для оценки гипотез с помощью контролируемых экспериментов;
- определять ключевые показатели и в идеале общий критерий оценки;
- проверять достоверность результатов и предупреждать экспериментаторов о нарушении предположений;
- выполнять быструю интерпретацию и итерации на основе полученных результатов;
- устанавливать ограничения для защиты ключевых бизнес-целей;
- создавать масштабируемые платформы, снижающие предельную стоимость экспериментов почти до нуля;
- избегать ловушек, таких как эффекты переноса, закон Тваймана, парадокс Симпсона и сетевые взаимодействия;
- решать на практике проблемы статистики, в том числе общие нарушения предположений.

Рон Кохави (Ron Kohavi) – вице-президент и технический сотрудник Airbnb. Он трудился над этой книгой, когда был членом технической команды и корпоративным вице-президентом в Microsoft. До этого он был директором по интеллектуальному анализу данных и персонализации в Amazon. Рон получил степень доктора информатики в Стэнфордском университете. Его статьи имеют более 40 000 цитирований, и три из них входят в 1000 наиболее цитируемых статей в области информатики.

Диана Тан (Diane Tang) – научный сотрудник Google, обладающий опытом в области крупномасштабного анализа данных и инфраструктуры, контролируемых онлайн-экспериментов и рекламных систем. Она получила

степень бакалавра гуманитарных наук в Гарварде и степень магистра/доктора в Стэнфорде, имеет патенты и публикации в области мобильных сетей, визуализации информации, методологии экспериментов, инфраструктуры данных, интеллектуального анализа данных и больших данных.

Я Сюй (Ya Xu) возглавляет отдел экспериментов и сбора данных в LinkedIn. Она опубликовала несколько работ по методике экспериментов и часто выступает на авторитетных конференциях и с лекциями в университетах. Ранее она работала в Microsoft и получила докторскую степень по статистике в Стэнфордском университете.

Предисловие

Если у нас есть данные, давайте рассмотрим их.
Если все, что у нас есть, – это мнения, делайте что я скажу.

– Джим Баркдейл,
бывший генеральный директор Netscape

Наша цель при написании этой книги – поделиться практическими уроками, извлеченными из многолетнего опыта проведения масштабных онлайн-экспериментов в Amazon и Microsoft (Рон Кохави), Google (Диана Тан), Microsoft и LinkedIn (Я Сюй). Хотя мы пишем эту книгу как частные лица, а не как представители Google, LinkedIn или Microsoft, мы опираемся на уроки и ошибки, с которыми столкнулись за годы нашей работы, и даем рекомендации как по программным платформам, так и по корпоративным аспектам использования контролируемых экспериментов для формирования культуры, основанной на данных, когда решения принимают исходя из достоверной информации, а не полагаясь на HiPPO (highest paid person's opinion, «мнение самой высокооплачиваемой персоны»). Мы считаем, что многие из этих уроков применимы в онлайн-среде, в больших или малых компаниях, или даже в командах и отделах внутри компании. Мы разделяем озабоченность по поводу необходимости оценки достоверности результатов экспериментов. Мы верим в скептицизм, выраженный законом Тваймана: любое значение, которое выглядит интересным или странным, обычно ошибочно; мы призываем читателей перепроверить результаты и провести тесты на достоверность, особенно если получены исключительно положительные результаты. Получить числа легко; получить числа, которым можно доверять, сложно!

Часть I предназначена для чтения всеми, независимо от уровня подготовки, и состоит из четырех глав:

- глава 1 представляет обзор преимуществ проведения контролируемых онлайн-экспериментов и вводит отраслевые термины;
- в главе 2 рассмотрен пример полного цикла организации и проведения эксперимента;
- в главе 3 описаны распространенные ошибки и способы повышения надежности экспериментов;
- в главе 4 рассказано, что нужно для создания экспериментальной платформы и масштабирования онлайн-экспериментов.

Части со II по V могут быть прочитаны по мере необходимости, но они написаны с упором на конкретную аудиторию. Часть II содержит пять глав,

посвященных основам, таким как показатели организации. Главы части II рекомендуются всем, особенно руководителям и менеджерам среднего звена. Часть III состоит из двух глав, в которых представлены методы, дополняющие контролируемые онлайн-эксперименты. Эти методы могут оказаться полезными для руководителей, специалистов по обработке данных, инженеров, аналитиков, менеджеров по продуктам и других специалистов, стремящихся эффективно расходовать время и усилия. Часть IV посвящена созданию платформы для экспериментов и предназначена для инженеров и разработчиков. Наконец, в части V рассматриваются вопросы расширенного анализа, и она предназначена для специалистов по данным.

Наш веб-сайт **https://experimentguide.com** дополняет эту книгу. Он содержит дополнительные материалы, исправления и предоставляет форум для открытого обсуждения. Все доходы от этой книги авторы намерены пожертвовать на благотворительность.

ЧАСТЬ I

Введение для всех

Глава 1

Введение и мотивация

Одно точное измерение стоит тысячи экспертных заключений.
– Адмирал Грейс Хоппер

В 2012 году сотрудник Bing, поисковой системы Microsoft, предложил изменить способ отображения заголовков объявлений. Идея заключалась в том, чтобы удлинить строку заголовка объявлений, объединив ее с текстом из первой строки под заголовком, как показано на рис. 1.1.

Никто не ожидал, что это самое простое изменение из сотен предложенных станет лучшей идеей для увеличения дохода в истории Bing!

Это предложение имело низкий приоритет и ждало своей очереди более шести месяцев, пока программист-разработчик не решил попробовать его, учитывая, насколько легко было внести изменение в код. Он внес изменение и начал проверять эффект на реальных пользователях, случайным образом показывая некоторым из них новый макет заголовка, а другим старый. При этом регистрировались взаимодействия пользователей с веб-сайтом, включая клики по рекламе и полученный от них доход. Это пример A/B-теста, простейшего типа контролируемого эксперимента, в котором сравниваются два варианта: A и B, или *контрольный* (control) и *тестовый* (treatment)[1].

Через несколько часов после запуска теста сработало предупреждение о слишком высоком уровне дохода, указывающее, что с экспериментом что-то не так. Новый макет заголовка приносил слишком много денег от рекламы. Такие предупреждения «слишком хорошо, чтобы быть правдой» очень полезны, поскольку они обычно указывают на серьезную ошибку, например случаи, когда выручка зарегистрирована дважды (двойной биллинг) или когда отображается только реклама, а остальная часть веб-страницы повреждена.

---

[1] Современная методика контролируемых экспериментов родилась в медицине и фармацевтике, и оттуда распространилась в другие отрасли. Поэтому в зарубежной литературе исторически принято обозначать экспериментальное воздействие и подопытную группу термином treatment (лечение, терапия). – *Прим. перев.*

Однако в этом эксперименте не было ошибки. Выручка Bing увеличилась на колоссальные 12 %, что в то время составляло более 100 млн долл. в год только в США, без заметного ущерба для ключевых показателей взаимодействия с пользователем. Эксперимент повторяли несколько раз в течение длительного периода.

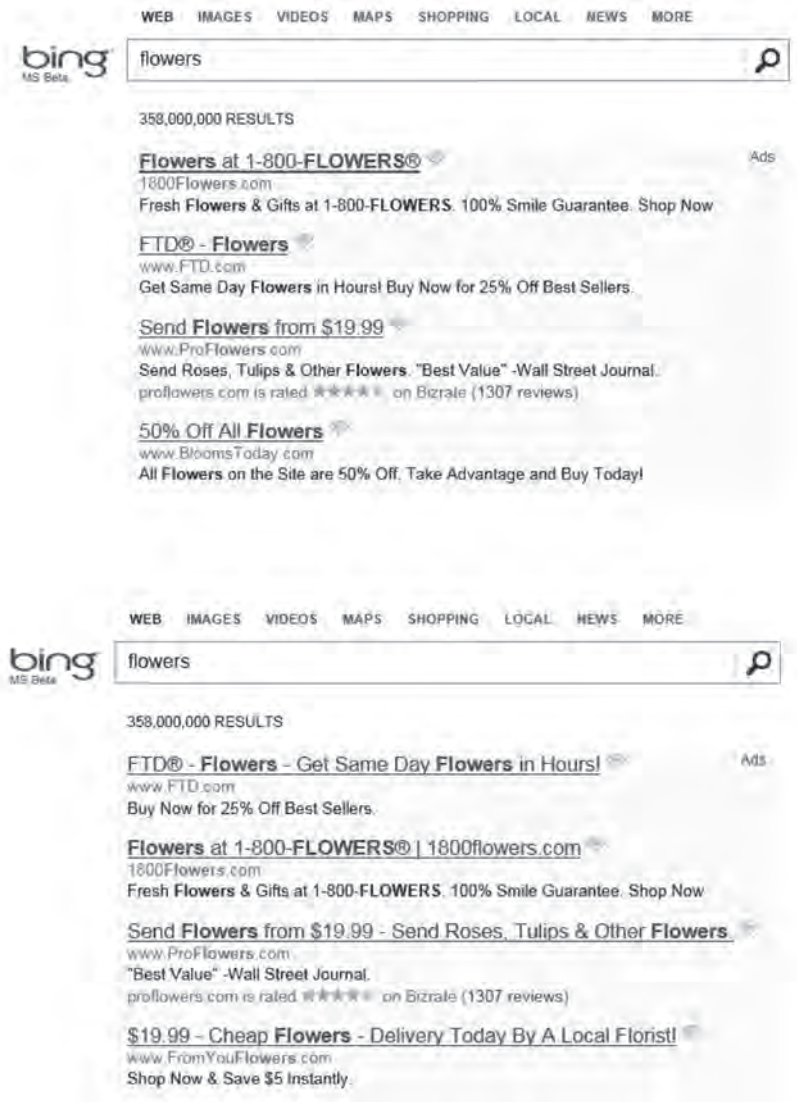


Рис. 1.1. Эксперимент, изменяющий способ отображения рекламы в Bing

Этот пример иллюстрирует несколько ключевых аспектов контролируемых онлайн-экспериментов:

- нам сложно понять ценность идеи. В данном случае простое изменение стоимостью более 100 млн долл. в год ждало своей очереди несколько месяцев;
- небольшие изменения могут оказать большое влияние. *Окупаемость инвестиций* (return on investment, ROI) в размере 100 млн долл. в год от нескольких дней работы одного программиста оказалась настолько экстремальной, насколько это вообще возможно;
- эксперименты с выраженным эффектом встречаются очень редко. Bing проводит более 10 000 экспериментов в год, но простые воздействия, приводящие к значительным улучшениям, появляются только раз в несколько лет;
- накладные расходы на проведение эксперимента должны быть небольшими. Инженеры Bing имели доступ к ExP, экспериментальной системе Microsoft, которая облегчила научную оценку идеи;
- *общий критерий оценки* (overall evaluation criterion, OEC) более подробно описанный в этой главе) должен быть сформулирован очень строго. В данном случае ключевым компонентом OEC была выручка, но одной лишь выручки недостаточно для формирования OEC. Бездумный подход может привести к заваливанию веб-сайта рекламой, что, как известно, отрицательно сказывается на опыте пользователей. Bing использует OEC, который соотносит доход с показателями взаимодействия с пользователем, включая количество сеансов на пользователя (т. е. когда пользователи отказываются от участия или, наоборот, увеличивают вовлеченность) и несколько других компонентов. Ключевым моментом нашего примера является то, что показатели взаимодействия с пользователем существенно не ухудшились, даже несмотря на то, что доход резко вырос.

В следующем разделе вводится терминология контролируемых экспериментов.

1.1. Терминология контролируемых онлайн-экспериментов

Контролируемые эксперименты имеют долгую и увлекательную историю, о которой мы уже рассказывали в статьях в интернете. Иногда их называют A/B-тестами, A/B/n-тестами (чтобы подчеркнуть множественность вариантов), *полевыми экспериментами* (field experiment), *рандомизированными контролируемыми экспериментами* (randomized controlled experiment), *сплит-тестами* (split test), *тестами с корзиной* (bucket test) и *пробными полетами* (flight test). В этой книге мы используем термины «контролируемые эксперименты» и «A/B-тесты» как синонимы, независимо от количества вариантов.

Контролируемые онлайн-эксперименты широко используются в таких компаниях, как Airbnb, Amazon, Booking.com, eBay, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yahoo!, Oath и Яндекс. Эти компании проводят от тысячи до десятков тысяч экспериментов каждый год иногда с участием миллионов пользователей и тестируют все подряд, включая изменения пользовательского интерфейса (UI), алгоритмы релевантности (поиск, реклама, персонализация, рекомендации и т. д.), задержку/быстродействие, системы управления контентом, системы поддержки клиентов и многое другое. Эксперименты проводятся по нескольким каналам: веб-сайтам, приложениям для компьютеров, мобильным приложениям и электронной почте.

В наиболее распространенных контролируемых онлайн-экспериментах пользователи случайным образом распределяются между вариантами на постоянной основе (пользователь получает один и тот же вариант за несколько посещений). В нашем первом примере от Bing контрольным вариантом было исходное отображение рекламы, а вариантом с воздействием – отображение рекламы с более длинными заголовками. Взаимодействие пользователей с веб-сайтом Bing было *измеримым*, т. е. отслеживалось и регистрировалось. По зарегистрированным данным вычисляются показатели, которые позволяют нам оценить разницу между вариантами для каждого показателя.

В простейших контролируемых экспериментах существует два варианта: контрольный (A) и тестовый (B), как показано на рис. 1.2.

Рис. 1.2. Простой контролируемый эксперимент: A/B-тест

Сейчас мы следуем терминологии Кохави и Лонгботтома (2017), а соответствующие термины из других областей будут представлены ниже. Вы найдете много других ресурсов по экспериментированию и A/B-тестированию в конце этой главы в разделе «Дополнительная литература».

Общий критерий оценки (OEC): количественный показатель цели эксперимента. Например, ваш OEC может быть выражен в днях на пользователя, указывая количество дней во время эксперимента, в течение которых пользователи были активны (т. е. они посетили сайт и предприняли какие-либо действия). Повышение этого показателя означает, что пользователи чаще посещают ваш сайт, что является отличным результатом. OEC должен быть измеримым в краткосрочной перспективе (продолжительность эксперимента), но при этом иметь причинное влияние на долгосрочные стратегические цели (см. раздел 1.7 и главу 7). В случае поисковой системы OEC может представлять собой комбинацию активности использования (например, количество сеансов на пользователя), релевантности (например, успешные сеансы, время до успеха) и дохода от рекламы (не все поисковые системы используют все перечисленные метрики или только эти метрики).

В статистике это часто называют *реакцией* (response) или *зависимой переменной* (dependent variable); другие синонимы – это *результат* (outcome), *оценка* (evaluation) и *функция пригодности* (fitness function). Эксперименты могут иметь несколько целей, и при анализе может применяться *сбалансированная система показателей* (balanced scorecard), хотя обычно рекомендуется использовать единственную метрику, возможно, представляющую собой взвешенную комбинацию таких целей.

Мы более подробно рассмотрим определение OEC для экспериментов в главе 7.

Параметр: контролируемая экспериментальная переменная, которая, как ожидается, влияет на OEC или другие интересующие показатели. Параметры иногда называют *факторами* или *переменными*. Параметрам присваивают *значения*, также называемые *уровнями*. В простых A/B-тестах обычно используют один параметр с двумя значениями. В онлайн-тестировании чаще используют конструкции с одной переменной и несколькими значениями (например, A/B/C/D). *Многовариантные тесты*, также называемые *мультивариантными* (multivariate test, MVT), оценивают вместе несколько параметров (переменных), таких как цвет и размер шрифта, что позволяет экспериментаторам находить глобальный оптимум при взаимодействии параметров (см. главу 4).

Вариант: тестируемый пользовательский опыт, обычно путем присвоения значений параметрам. В простом A/B-тесте A и B – это два варианта, обычно называемые контрольным и тестовым. В некоторой литературе под вариантом понимают только тест; однако мы рассматриваем контроль как особый вариант: существующую версию, с которой будет выполняться сравнение. Например, в случае обнаружения ошибки в эксперименте вы должны прервать эксперимент и убедиться, что всем пользователям назначен (возвращен) контрольный вариант.

Рандомизатор: к *объектам эксперимента* (например, пользователям или страницам) применяется псевдослучайный процесс (например, хеширование) для сопоставления их с вариантами. Правильная рандомизация важна для обеспечения статистического сходства популяций, отнесенных к различным вариантам, что позволяет с высокой вероятностью определять причинные эффекты. Вы должны сопоставлять объекты с вариантами постоянным и независимым образом (т. е. если объектом рандомизации является пользователь, он должен постоянно видеть один и тот же опыт, а назначение пользователя определенному варианту не должно ничего говорить вам о назначении другого пользователя другому варианту). Этот подход применяется очень часто, и мы тоже настоятельно рекомендуем использовать пользователей в качестве объекта рандомизации при проведении контролируемых экспериментов для онлайн-аудитории. Некоторые экспериментальные проекты выбирают рандомизацию по страницам, сеансам или пользовательским дням (т. е. эксперимент остается неизменным для пользователя в течение каждого 24-часового окна, определенного сервером); см. главу 14 для получения дополнительной информации.

Правильная рандомизация имеет решающее значение! Если планом эксперимента предусмотрен равный процент пользователей для каждого варианта, то у каждого пользователя должны быть равные шансы быть назначенным каждому варианту. Не относитесь к рандомизации легкомысленно. Приведенные ниже примеры демонстрируют сложность и важность правильной рандомизации.

- Корпорации RAND в 1940-х годах потребовались случайные числа для методов Монте-Карло, поэтому они создали книгу из миллиона случайных цифр, сгенерированных с помощью импульсной машины. Однако из-за особенностей оборудования исходная таблица, как выяснилось, имела значительные смещения, и цифры пришлось рандомизировать заново в следующем издании книги (RAND 1955).
- Контролируемые эксперименты изначально использовались в медицине. Управление по делам ветеранов США провело эксперимент по применению стрептомицина для лечения туберкулеза, но испытания не увенчались успехом, поскольку врачи внесли свои предубеждения и повлияли на процесс отбора. Аналогичные испытания в Великобритании были проведены с использованием слепых протоколов и оказались успешными, что стало переломным моментом в методике контролируемых испытаний препаратов.

На присвоение варианта не должен влиять ни один фактор. Пользователи (объекты) не могут быть распределены каким-либо «обдуманным» способом. Важно отметить, что случайность означает не «спонтанный или незапланированный отбор, а преднамеренный выбор, основанный на вероятностях» (Mosteller, Gilbert and McPeek, 1983). В одной из своих работ Сенн (Senn, 2012) обсуждает мифы рандомизации.

1.2. Зачем нужны эксперименты? Корреляции, причинно-следственная связь и доверительность

Допустим, вы работаете в компании, распространяющей контент по подписке, такой как Netflix, где X % пользователей уходят (закрывают подписку) каждый месяц. Вы решаете ввести новую функцию интерфейса и видите, что коэффициент оттока пользователей, использующих эту функцию, составляет 0,5 X %, т. е. половину. У вас может возникнуть соблазн заявить о причинной связи; дескать, новая функция сокращает отток пользователей наполовину. Отсюда следует вывод, что, если мы сделаем эту функцию более заметной и заставим пользователей чаще ее использовать, количество подписок вырастет. Неправильно! На основе имеющихся данных нельзя сделать вывод о том, как влияет на отток пользователей эта функция, и возможны разные варианты – увеличение, уменьшение или неизменность.

Пример, демонстрирующий это заблуждение, можно найти в Microsoft Office 365, другом сервисе по подписке. Пользователи Office 365, которые видят сообщения об ошибках и сталкиваются со сбоями, демонстрируют более низкий коэффициент оттока, но это не означает, что Office 365 должен показывать больше сообщений об ошибках или что компания Microsoft должна снижать качество кода, вызывая больше сбоев. Оказывается, все три события вызваны одним фактором: использованием. Активные пользователи продукта видят больше сообщений об ошибках, сталкиваются с бо̀льшим количеством сбоев и имеют более низкий процент оттока. Корреляция не подразумевает причинно-следственной связи, и неверная трактовка этих наблюдений приводит к ошибочным решениям.

В 1995 году Гаятт и коллеги представили *иерархию доказательств* как способ оценки рекомендаций в медицинской литературе. Впоследствии Гринхал расширила этот метод в своих исследованиях практики доказательной медицины (1997, 2014). На рис. 1.3 показана простая иерархия доказательств в переводе на нашу терминологию. Рандомизированные контролируемые эксперименты – золотой стандарт для установления причинно-следственной связи. Систематические обзоры, т. е. метаанализ контролируемых экспериментов, предоставляют дополнительные доказательства и возможности обобщения.

Рис. 1.3. Простая иерархия доказательств для оценки качества методики исследования (Greenhalgh, 2014)

Также применяются более сложные модели, такие как *уровни доказательности* (levels of evidence) Оксфордского центра доказательной медицины (2009).

Платформы для экспериментов, используемые нашими компаниями, позволяют экспериментаторам в Google, LinkedIn и Microsoft проводить десятки тысяч контролируемых онлайн-экспериментов в год с высокой степенью доверия к результатам. Мы считаем, что контролируемые онлайн-эксперименты:

- являются лучшим научным способом выявить причинно-следственную связь с высокой степенью достоверности;
- позволяют обнаруживать небольшие изменения, которые труднее обнаружить другими методами, например изменения во времени (чувствительность);
- позволяют обнаруживать неожиданные изменения. Этот момент часто недооценивают, но многие эксперименты демонстрируют неожиданное влияние на другие показатели, будь то снижение быстродействия, увеличение количества сбоев/ошибок или каннибализация[1] кликов от других функций.

Основное внимание в этой книге уделяется выявлению потенциальных ошибок в экспериментах и предложению методов, повышающих доверие к результатам. Контролируемые онлайн-эксперименты предоставляют беспрецедентную возможность автоматически собирать надежные данные в широком масштабе, хорошо рандомизировать объекты и избегать или об-

[1] Перетягивание кликов на себя, зачастую без видимой пользы. – *Прим. перев.*

наруживать подводные камни (см. главу 11). Мы рекомендуем использовать другие, менее надежные методы, в том числе наблюдательные исследования, лишь в случае невозможности проведения контролируемых онлайн-экспериментов.

1.3. Необходимые ингредиенты для проведения эффективных контролируемых экспериментов

Не каждое решение может быть подкреплено научной строгостью контролируемого эксперимента. Например, вы не сможете провести контролируемый эксперимент по корпоративным слияниям и поглощениям, так как невозможно одновременно реализовать слияние/поглощение и его противоположность (т. е. отсутствие такого события). Далее мы рассмотрим необходимые технические компоненты для проведения *полезных* контролируемых экспериментов, а затем обсудим организационные постулаты. В главе 4 мы рассматриваем модель зрелости экспериментирования. Итак, для проведения контролируемых экспериментов нужны следующие компоненты и условия.

1. Экспериментальные объекты (например, пользователи), которые могут быть отнесены к разным вариантам без перекрестного взаимного влияния (или с небольшим влиянием); например, пользователи в тестовой группе не влияют на пользователей в контрольной группе (см. главу 22).
2. Достаточное количество экспериментальных объектов (например, пользователей). Чтобы контролируемые эксперименты были полезными, мы рекомендуем использовать тысячи экспериментальных объектов: чем больше число, тем тоньше эффекты, которые можно обнаружить. Хорошая новость заключается в том, что даже небольшие стартапы в области программного обеспечения обычно быстро набирают достаточное количество пользователей и могут начать проводить контролируемые эксперименты, поначалу стремясь к выраженным эффектам. По мере роста бизнеса возрастает значимость мелких эффектов (например, крупные веб-сайты должны иметь возможность обнаруживать небольшие изменения ключевых показателей, влияющих на пользовательский опыт и доли процента изменения дохода), а чувствительность улучшается с ростом базы пользователей.
3. Ключевые показатели, в идеале OEC, сформулированы и могут быть измерены на практике. Если реальные цели слишком сложно измерить, важно договориться о суррогатных целях (см. главу 7). Вы должны иметь возможность собирать надежные данные, в идеале – дешево и масштабно. В экспериментах с программным обеспечением обычно легко удается регистрировать системные события и действия пользователя (см. главу 13).

4. Простота внесения изменений. Программное обеспечение обычно легче изменить, чем оборудование; но даже в программном обеспечении некоторые области требуют определенного уровня гарантии качества. Изменения в рекомендательном алгоритме легко внести и оценить; изменения программного обеспечения в системах управления полетом самолетов требуют совершенно иного процесса утверждения Федеральным управлением гражданской авиации (FAA). Серверное программное обеспечение намного проще изменить, чем клиентское (см. главу 12), поэтому обращение к серверным службам из клиентских приложений становится все более распространенным подходом, что позволяет выполнять обновления и изменения служб быстрее и с помощью контролируемых экспериментов.

Большинство нетривиальных онлайн-сервисов соответствует или может соответствовать необходимым условиям для запуска гибкого процесса разработки, основанного на контролируемых экспериментах. Многие приложения, работающие по принципу обращения к серверным службам, также могут относительно легко подстроиться под эти требования. Томке (Thomke) писал, что организации ощущают максимальную выгоду от экспериментов, когда они используются в сочетании с «инновационной системой». Такой инновационной системой является гибкая разработка программного обеспечения.

Когда контролируемые эксперименты невозможны, можно провести моделирование и использовать другие экспериментальные методы (см. главу 10). Ключевая мысль здесь заключается в том, что если есть возможность проводить контролируемые эксперименты, то лучше их и выбрать, поскольку они предоставляют наиболее надежный и чувствительный механизм для оценки эффекта от изменений.

1.4. Постулаты

В 2013 году Кохави сформулировал три основных постулата для организаций, которые хотят проводить контролируемые онлайн-эксперименты.

1. Организация хочет принимать решения на основе данных и официально оформила OEC.
2. Организация готова инвестировать в инфраструктуру и тесты, чтобы проводить контролируемые эксперименты и гарантировать достоверность результатов.
3. Организация осознает, что плохо умеет оценивать значимость идей.

***Постулат 1**: Организация хочет принимать решения на основе данных и официально оформила OEC*

Вы редко услышите, как кто-то во главе организации говорит, что они не хотят ориентироваться на данные (заметным исключением являлась Apple под руководством Стива Джобса, где Кен Сегалл заявил, что «мы не тести-

ровали ни одной рекламы для печати, телевидения, рекламных щитов, интернета, розничной торговли и т. д.»). Но измерение дополнительных выгод для пользователей от новых функций требует затрат, и объективные измерения обычно показывают, что прогресс не такой радужный, как первоначально предполагалось. Многие организации не будут тратить ресурсы, необходимые для определения и измерения прогресса. Часто бывает проще (и политически выгоднее) составить план, выполнить его и объявить об успехе с ключевым показателем «процент выполнения плана», игнорируя, оказывает ли новая функция хоть какое-то положительное влияние на ключевые показатели компании.

Чтобы стать компанией, опирающейся на данные, необходимо определить OEC, легко измеряемый за относительно короткие промежутки времени (например, от одной до двух недель). В крупных организациях может быть несколько OEC или несколько ключевых показателей, которые используются совместно с уточнениями для разных областей. Сложнее всего найти показатели, измеримые за короткий период, но при этом достаточно чувствительные, чтобы показать различия, и позволяющие прогнозировать долгосрочные цели. Например, показатель «прибыль» – не лучший вариант OEC, поскольку сиюминутные действия (например, повышение цен) могут увеличить краткосрочную прибыль, но и навредить ей в долгосрочной перспективе. И наоборот, *длительность жизненного цикла клиента* (customer lifetime value, CLV) – это стратегически мощный OEC. Трудно переоценить важность разработки хорошего OEC, который подходит именно вашей организации (см. главу 6).

Чтобы не складывалось впечатление, будто решения в компании зависят от единственного источника данных (например, контролируемого эксперимента), часто используют термины-эвфемизмы, такие как «основанный на больших данных» или «информационно обоснованный». В этой книге мы используем подобные термины как синонимы. Решение и в самом деле следует принимать с использованием множества источников данных, включая контролируемые эксперименты, опросы, оценки затрат на сопровождение нового кода и т. д. Организация, основанная на данных или информационных потоках, собирает соответствующие данные для принятия решения и формирования обоснованного мнения высокопоставленных лиц, а не полагается только на их интуицию.

***Постулат 2**: Организация готова инвестировать в инфраструктуру и тесты, чтобы проводить контролируемые эксперименты и гарантировать достоверность результатов*

В области онлайн-программирования (веб-сайты, мобильные, настольные приложения и сервисы) необходимые условия для контролируемых экспериментов могут быть выполнены с помощью разработки программного обеспечения (см. раздел 1.3): можно надежно рандомизировать пользователей,

есть возможность собирать телеметрию и относительно легко вносить изменения в программное обеспечение, например внедрять новые функции (см. главу 4). Даже на относительно небольших веб-сайтах наберется достаточно пользователей для проведения необходимых статистических тестов.

Контролируемые эксперименты особенно полезны в сочетании с гибкой разработкой программного обеспечения (Мартин, 2008, К. С. Рубин, 2012), процессом развития клиентов (Бланк, 2005) и *минимально жизнеспособными продуктами* (minimum viable product, MVP) – это понятие предложил Эрик Рис в книге «Экономичный стартап» (Ries, 2011).

В других областях может быть трудно или невозможно надежно проводить контролируемые эксперименты. Некоторые воздействия, необходимые для контролируемых экспериментов в области медицины, могут быть неэтичными или незаконными. Аппаратные устройства могут иметь длительные сроки изготовления и модификации, поэтому контролируемые эксперименты с пользователями редко проводятся на новых аппаратных устройствах (например, новых мобильных телефонах). В этих ситуациях могут потребоваться дополнительные методы (см. главу 10), применяемые, когда невозможно провести контролируемые эксперименты.

Предполагая, что вы можете проводить контролируемые эксперименты, важно обеспечить их надежность. При проведении онлайн-экспериментов получить числа легко; получить числа, которым можно доверять, сложно. Глава 3 посвящена получению достоверных результатов.

***Постулат 3**: Организация осознает, что плохо умеет оценивать значимость идей*

Новые функции появляются потому, что команды считают их полезными, однако во многих областях большинство идей не помогают улучшить ключевые показатели. Только треть идей, протестированных в Microsoft, пошла на пользу показателям, на которых была направлена разработка. Еще труднее добиться успеха в хорошо оптимизированных областях, таких как Bing и Google, где показатель успешности некоторых показателей составляет около 10–20 %.

Фарид Мосават (Farid Mosavat), директор по продуктам и жизненному циклу Slack, написал в Twitter, что с учетом всего опыта Slack только около 30 % экспериментов по монетизации показывают положительные результаты: «Если вы работаете в команде экспериментаторов, привыкните к тому, что в лучшем случае 70 % вашей работы пропадет напрасно, и в соответствии с этим выстраивайте свои процессы».

Авинаш Кошик (Avinash Kaushik) написал в своем учебнике по экспериментам и тестированию что «в 80 % случаев мы ошибаемся в оценке того, чего хочет клиент». Майк Моран (Mike Moran) писал, что Netflix считает 90 % того, что они тестируют, заведомо ошибочным. Регис Хадиарис (Regis Hadiaris) из Quicken Loans писал, что «за те пять лет, что я проводил тесты, я примерно

так же правильно угадывал результаты, как бейсболист высшей лиги бьет по мячу. Естественно – я занимаюсь этим пять лет и могу только "угадывать" результат теста примерно в 33 % случаев!» Дэн МакКинли (Dan McKinley) из Etsy написал: «Почти все тесты терпят неудачу», а в отношении новых функций он написал: «Было унизительно осознавать, насколько редко они преуспевают с первой попытки. Я сильно подозреваю, что этот опыт универсален, но он не получил всеобщего признания или принятия». Наконец, Колин Макфарланд (Colin McFarland) в своей в книге *Experiment!* написал: «Независимо от того, насколько вы уверены, что это несложно, сколько исследований вы провели или сколько конкурентов занимаются этим, экспериментальные идеи терпят неудачу намного чаще, чем вы думаете».

Такая печальная статистика присуща не каждой области экспериментов, но большинство из тех, кто проводил контролируемые эксперименты на веб-сайтах и приложениях, ориентированных на клиентов, испытали на себе эту унизительную реальность: *мы не умеем правильно оценивать значимость идей*.

1.5. Постепенные улучшения

На практике улучшения ключевых показателей достигаются за счет множества небольших изменений: от 0,1 до 2 %. Многие эксперименты влияют только на определенный сегмент пользователей, поэтому вы должны разделить влияние 5 % улучшения на 10 % ваших пользователей, что приведет к гораздо меньшему влиянию (например, 0,5 %, если подопытная популяция аналогична остальной популяции пользователей); более подробно мы обсудим этот вопрос в главе 3. Как говорит герой Аль Пачино в фильме «Каждое воскресенье», «… к победе идут дюйм за дюймом».

Пример рекламы Google

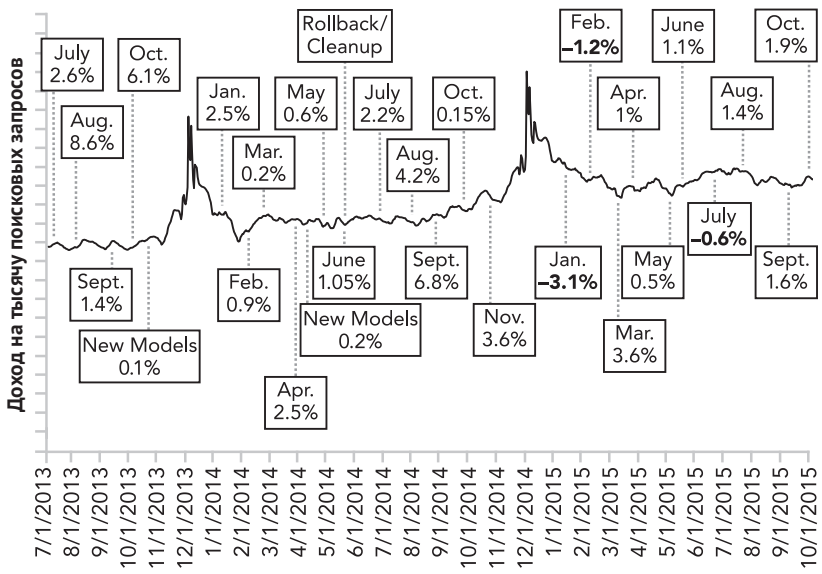
В 2011 году, после более чем года разработки и дополнительных экспериментов, Google запустил улучшенный механизм ранжирования рекламы. Инженеры разработали и экспериментально проверили новые и улучшенные модели измерения показателя качества рекламы в рамках существующего механизма ранжирования рекламы, а также изменения самого аукциона рекламы. Они провели сотни контролируемых экспериментов и выполнили несколько итераций; некоторые на всех рынках, а некоторые – в долгосрочной перспективе на определенных рынках, чтобы лучше понять влияние на рекламодателей. Это значительное изменение бэкенда – подкрепленное контролируемыми экспериментами – в конечном итоге подтвердило, что планирование нескольких изменений и их последующее объединение улучшило взаимодействие с пользователем за счет предоставления более качественной рекламы и улучшило опыт рекламодателей, получивших более низкую среднюю цену за более качественную рекламу.

Пример Bing Relevance

Команда подразделения Relevance в Bing состоит из нескольких сотен человек, которым поручено ежегодно улучшать один показатель OEC на 2 %. Это сумма эффектов воздействия (т. е. дельта OEC) во всех контролируемых экспериментах, проведенных на пользователях в течение года, при условии, что они являются аддитивными. Поскольку команда проводит тысячи экспериментов и некоторые из них могут оказаться положительными случайно, баллы в размере 2 % присваиваются только на основе репликации эксперимента: как только реализация идеи покажется успешной, возможно, после нескольких итераций и уточнений, выполняется *удостоверяющий эксперимент* (sertification experiment). Положительный эффект этого эксперимента подтверждает достижение целевого балла 2 %. В последнее время планируется снизить порог эффекта для повышения точности воздействия.

Пример рекламной службы Bing

Команда рекламной службы Bing стабильно увеличивала доход на 15–25 % в год, но большинство улучшений происходило постепенно. Каждый месяц команда формировала «пакеты» – результаты многих экспериментов, как показано на рис. 1.4. Большинство улучшений были небольшими, некоторые ежемесячные пакеты даже оказывались отрицательными из-за нехватки места под рекламу или требований законодательства.



(*) Числа изменены по очевидным причинам

Рис. 1.4. Доход от рекламы Bing во времени (ось Y показывает рост примерно на 20 % в год). Конкретные числа не важны

Любопытно и полезно наблюдать сезонные всплески примерно в декабре, когда покупательное намерение пользователей резко возрастает, поэтому рекламное пространство увеличивается и доход на тысячу поисковых запросов становится больше.

1.6. Примеры интересных контролируемых онлайн-экспериментов

Интересны эксперименты, в которых абсолютная разница между ожидаемым и фактическим результатом велика. Если вы предполагали, что что-то должно произойти, и это произошло, это вас мало чему научит. Если вы думали, что что-то должно произойти, а этого не произошло, значит, вы узнали кое-что важное. И если вы думали, что произойдет что-то незначительное, а результаты являются большим сюрпризом и ведут к прорыву, вы узнали что-то очень ценное.

Примеры Bing в начале этой главы и в этом разделе рассказывают истории необычного успеха с удивительными, весьма полезными результатами. Попытка Bing интегрироваться с социальными сетями, такими как Facebook и Twitter, является примером ожидания сильного эффекта и его отсутствия – попытки были прекращены после того, как многочисленные эксперименты не принесли никакой пользы в течение двух лет.

Хотя устойчивый прогресс – это вопрос непрерывных экспериментов и множества небольших улучшений, дальше мы приведем несколько примеров неожиданно большой отдачи, демонстрирующих, насколько плохо мы оцениваем полезность идей.

Пример пользовательского интерфейса: 41 оттенок синего

Мелкие дизайнерские решения могут давать значительный эффект, что неизменно демонстрируют и Google, и Microsoft. Google протестировал 41 градацию синего цвета[1] на страницах результатов поиска, разочаровав тогдашних лидеров визуального дизайна. Тем не менее корректировки цветовой схемы, внесенные Google, в конечном итоге оказали положительное влияние на вовлечение пользователей (обратите внимание, что Google не сообщает о результатах отдельных промежуточных изменений) и привели к тесному сотрудничеству между дизайнерами и разработчиками экспериментов. Доработки цветовой схемы Bing от Microsoft аналогичным образом показали, что пользователи стали более успешны в решении своих задач, период вовлеченности стал длиннее, а монетизация возросла до суммы более 10 млн долл. США в год.

Это отличные примеры крошечных изменений, оказывающих огромное

[1] Цвет выбран не случайно. Оттенки синего цвета, в отличие от других цветов, хорошо различает даже большинство дальтоников. – *Прим. перев.*

влияние, но, если учесть, что уже опробован широкий спектр цветов, маловероятно, что игра с цветами в последующих экспериментах приведет к ощутимым улучшениям.

Предложение в нужное время

В 2004 году Amazon разместил на главной странице рекламный блок кредитной карты. Он был очень прибыльным, но имел слишком низкий *коэффициент отклика* (click-through rate, CTR). Команда провела эксперимент, переместив блок на страницу корзины покупок, которую пользователь видит после добавления товара, и снабдила блок простыми вычислениями, наглядно показывающими экономию, которую получит пользователь (рис. 1.5).

Поскольку пользователи, добавляющие товар в корзину, имеют четкое намерение совершить покупку, это предложение отображается в нужное время. Контролируемый эксперимент показал, что это простое изменение увеличило годовую прибыль Amazon на десятки миллионов долларов.

Рис. 1.5. Предложение Amazon по кредитной карте с экономией от общей стоимости корзины

Персональные рекомендации

Грег Линден (Greg Linden) из Amazon разработал алгоритм для отображения персонализированных рекомендаций на основе товаров в корзине покупателя. Когда вы добавляете товар, появляются рекомендации; если вы добавите еще один товар, появятся новые рекомендации. Линден отмечает, что, хотя алгоритм выглядел многообещающе, «старший вице-президент по маркетингу был категорически против», утверждая, что это отвлечет людей от покупок. Грегу «строго запретили работать над этим дальше». Тем не менее он провел контролируемый эксперимент, и «нововведение выиграло с таким огромным отрывом, что отказ от него можно было бы компенсировать только значительными изменениями в Amazon. Поэтому рекомендации на основе корзины немедленно запустили в работу». Теперь рекомендации на основе корзины используют многие сайты.

Скорость имеет большое значение

В 2012 году инженер-разработчик Microsoft Bing внес изменения в технологию разработки кода JavaScript, что значительно сократило HTML-код,

отправляемый клиентам, и привело к повышению быстродействия. Контролируемый эксперимент показал удивительное улучшение ключевых показателей. Разработчики провели дополнительный эксперимент, чтобы оценить влияние быстродействия сервера. Результат показал, что повышение быстродействия сервера также значительно улучшает ключевые пользовательские метрики, такие как процент успеха и время до успеха, и каждое улучшение производительности на 10 мс (1/30 длительности моргания глаза) полностью окупает годовые затраты на оклад разработчика.

К 2015 году, когда производительность Bing значительно улучшилась, встал вопрос о том, имеет ли значение дальнейшее повышение быстродействия, если сервер и так возвращает результаты менее чем за секунду на 95-м процентиле (т. е. для 95 % запросов). Команда Bing провела дополнительное исследование, показавшее, что ключевые показатели пользователей по-прежнему значительно улучшаются. Хотя относительное влияние на доход несколько снизилось, выручка Bing за это время выросла настолько, что каждая миллисекунда повышения быстродействия значила больше, чем в прошлом; доход от сокращения времени отклика всего на четыре миллисекунды перекрывал годовой оклад инженера (см. главу 5 с подробным описанием этого эксперимента и анализом важности быстродействия)!

Эксперименты с быстродействием проводились в нескольких компаниях, и результаты показали, насколько важен этот параметр. В Amazon эксперимент с замедлением на 100 мс снизил продажи на 1 %. Совместный доклад инженеров Bing и Google продемонстрировал значительное влияние быстродействия на ключевые показатели, включая отдельные запросы, доход, количество кликов, удовлетворенность и время до клика.

Борьба с вредоносными программами

Реклама – прибыльный бизнес, и устанавливаемые пользователями бесплатные приложения часто содержат вредоносный код, который засоряет страницы рекламой. На рис. 1.6 показано, как страница Bing выглядела для пользователя с вредоносным ПО. Обратите внимание, что на страницу было добавлено несколько рекламных объявлений (выделены красным).

Мало того, что вредоносный код удалял рекламу Bing, лишая Microsoft дохода, он отображал нерелевантную рекламу низкого качества, что создавало неудобства для пользователей, которые, возможно, не понимали, почему они видят так много рекламы. Microsoft провела управляемый эксперимент, направленный на 3,8 млн пользователей, потенциально пострадавших от вредоносного кода, в котором основные процедуры, изменяющие DOM (document object model, объектная модель документа), были переопределены, чтобы разрешить только ограниченные модификации из надежных источников. Результаты показали улучшение всех ключевых показателей Bing, включая количество сеансов на пользователя. Следовательно, пользователи стали посещать сайт чаще или уходить реже. Кроме того, пользователи стали более успешными в поиске, быстрее переходили по по-

лезным ссылкам, а годовой доход увеличился на несколько миллионов долларов. Вдобавок время загрузки страницы – ключевой показатель быстродействия, который мы обсуждали ранее, – для очищенных от вредоносного кода страниц улучшилось на сотни миллисекунд.

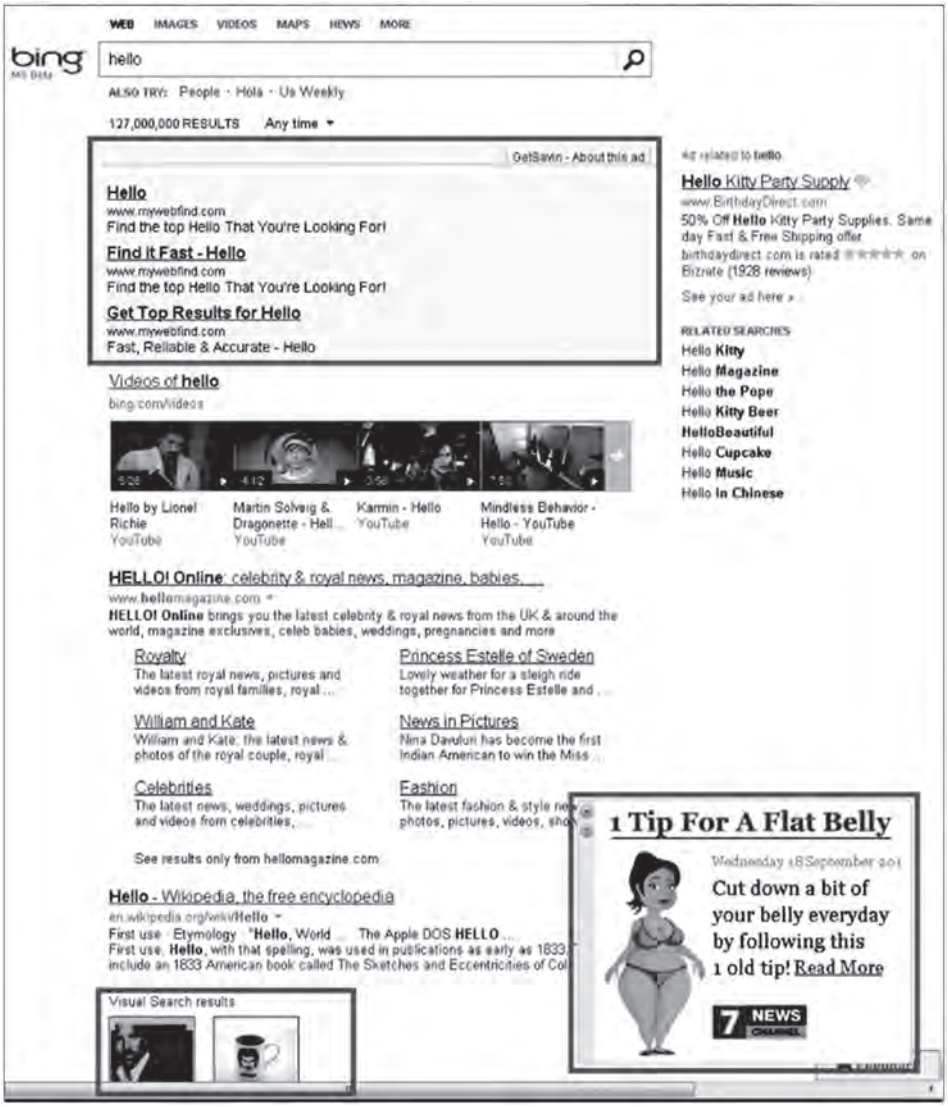


Рис. 1.6. Страница Bing, когда у пользователя есть вредоносное ПО, показывает несколько посторонних объявлений

Изменения в бэкенде

Изменения в бэкенд-алгоритмах часто упускаются из виду как область приложения контролируемых экспериментов, но они могут дать существенные результаты. Мы можем видеть это как на примере команд Google, LinkedIn и Microsoft, работающих над множеством небольших изменений, как мы говорили выше, так и в следующем примере с участием Amazon.

Еще в 2004 году существовал хороший алгоритм рекомендаций, основанный на двух наборах. Сначала рекомендация Amazon звучала как «Люди, которые купили товар X, купили товар Y», но затем ее обобщили до вида «Люди, которые ***просмотрели*** товар X, купили товар Y» и «Люди, которые просмотрели товар X, ***просмотрели*** товар Y». Было предложено использовать тот же алгоритм для рекомендации «Люди, которые ***искали*** X, купили товар Y». Сторонники обновления алгоритма привели примеры неполных поисковых запросов, таких как «24», который большинство людей ассоциируют с телешоу с Кифером Сазерлендом в главной роли. Поиск Amazon дает плохие результаты (слева на рис. 1.7), такие как компакт-диски с 24 итальянскими песнями, одежда для 24-месячных малышей, 24-дюймовая вешалка для полотенец и т. д. Новый алгоритм дал первоклассные результаты (справа на рис. 1.7), вернув в поисковую выдачу DVD-диски с шоу и соответствующими книгами, на основе того, что люди *действительно* купили после поиска «24». Одной из слабых сторон алгоритма было то, что всплывали некоторые элементы, вообще не имеющие отношения к поисковому запросу; однако Amazon провела контролируемый эксперимент, и, даже несмотря на недостатки, это изменение увеличило общий доход Amazon на 3 % – сотни миллионов долларов.

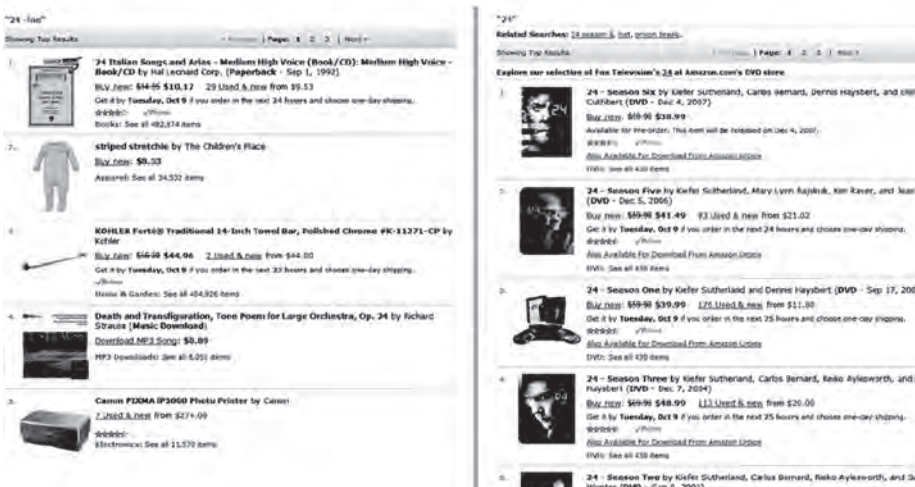


Рис. 1.7. Поиск Amazon по запросу «24» без рекомендательного алгоритма (слева) и с применением алгоритма (справа)

1.7. Стратегия, тактика и их связь с экспериментами

Мы твердо убеждены, что, если имеются необходимые условия для проведения контролируемых онлайн-экспериментов, они должны использоваться для принятия организационных решений на всех уровнях, от стратегии до тактики.

Стратегия компании и контролируемые эксперименты синергичны. Дэвид Коллис (David Collis) из Lean Strategy писал, что «эффективная стратегия не подавляет предпринимательское поведение, а поощряет его, определяя границы, в которых должны иметь место инновации и эксперименты». Он дает определение процесса *бережливой стратегии* (lean strategy), которая уберегает от крайностей как жесткого планирования, так и безудержного экспериментирования.

Качественно проведенные эксперименты с правильно подобранными показателями дополняют бизнес-стратегию, разработку продукта и повышают операционную эффективность, приводя организацию к управлению данными, а не интуицией. Инкапсулируя стратегию в OEC, контролируемые эксперименты могут обеспечить отличную обратную связь для стратегии. Достаточно ли хорошо повышают OEC идеи, оцениваемые с помощью экспериментов? С другой стороны, неожиданные результаты экспериментов могут пролить свет на альтернативные стратегические возможности, ведущие к смене вектора развития. Решения по разработке продукта важны для стратегии развития, и тестирование нескольких вариантов разработки обеспечивает разработчикам полезную обратную связь. Наконец, многие тактические изменения могут улучшить *операционную эффективность*, которую Портер определяет как «выполнение аналогичных действий лучше, чем их выполняют конкуренты».

Далее мы рассмотрим два ключевых сценария.

***Сценарий 1**. У вас есть бизнес-стратегия, и у вас есть продукт с достаточным количеством пользователей для экспериментов*

В этом сценарии эксперименты могут помочь подняться на холм локального оптимума на основе вашей текущей стратегии и продукта:

- эксперименты помогают выявить области с высокой рентабельностью инвестиций: те, которые улучшают OEC больше всего в пересчете на затраты. Тестирование различных областей с помощью MVP помогает быстрее исследовать обширный набор областей до выделения значительных ресурсов;
- эксперименты могут подсказать направление оптимизации, которое не очевидно для разработчиков, но может иметь большое значение для пользователя (например, цвет, интервал, быстродействие);
- эксперименты помогают непрерывно и постепенно улучшать дизайн сайта, вместо того чтобы вводить радикальное изменение сайта, которое подвергает пользователей шоку отказа от привычки (пользова-

тели привыкли к старому дизайну и его функциям) и обычно терпит неудачу – не только в достижении намеченных целей, но даже не достигает паритета со старым сайтом по ключевым показателям;

- эксперименты могут иметь решающее значение для оптимизации внутренних алгоритмов и инфраструктуры, например алгоритмов рекомендаций и ранжирования.

Наличие стратегии критически важно для проведения экспериментов: именно она определяет выбор OEC. Далее контролируемые эксперименты помогают ускорить инновации, побуждая команды оптимизировать и улучшать OEC. Как правило, мы наблюдаем неправильное использование экспериментов именно там, где не потрудились выбрать OEC должным образом. Выбранные метрики должны соответствовать ключевым характеристикам и не допускать неоднозначности (см. главу 7).

В наших компаниях у нас есть не только команды, которые сосредоточены на правильном проведении экспериментов, но также есть команды, ориентированные на показатели: выбор показателей, проверка показателей и развитие показателей с течением времени. Эволюция показателей будет происходить как по мере развития вашей стратегии с течением времени, так и по мере того, как вы узнаете больше об ограничениях существующих показателей, например о том, что показатель CTR слишком неоднозначен и его необходимо доработать. Группы, отвечающие за показатели, также работают над определением того, какие показатели, измеримые в краткосрочной перспективе, определяют долгосрочные цели, поскольку эксперименты обычно проводятся в относительно короткие сроки. Хаузер и Кац писали, что «фирма должна определить показатели, на которые команда может повлиять сегодня, но в конечном итоге влияющие на долгосрочные цели фирмы» (см. главу 7).

Привязка стратегии к OEC также создает *стратегическую целостность* (strategic integrity). Синофски и Иенсити (Sinofski, Iansiti) подчеркивают, что «стратегическая целостность – это не создание блестящей стратегии или наличие идеальной организации: это обеспечение реализации правильных стратегий организацией, которая обладает согласованностью и знает, как их реализовать. Речь идет о согласовании директив, направленных сверху вниз, с задачами, направленными снизу вверх». OEC – идеальный механизм, чтобы прояснить стратегию и согласовать желаемые функции со стратегией.

В конце концов, без хорошего OEC вы тратите ресурсы впустую – представьте эксперименты по улучшению питания или освещения на тонущем круизном лайнере. Значение показателя безопасности пассажиров в OEC для подобных экспериментов должно быть чрезвычайно высоким – на самом деле настолько высоким, чтобы вообще не трогать безопасность. Это может быть определено либо с помощью большого веса показателя в OEC, либо, что то же самое, с использованием безопасности пассажиров в качестве *граничного показателя* (см. главу 21). В программном обеспечении ана-

логом безопасности пассажиров круизного лайнера служат сбои программного обеспечения: если доработка увеличивает количество сбоев продукта, опыт считается настолько плохим, что другие факторы в сравнении с ним отступают на второй план.

Определение контрольных показателей для экспериментов важно для определения того, что организация не желает менять, поскольку стратегия, по словам Портера, также «требует от вас идти на компромисс в конкурентной борьбе – выбирать, чего не следует делать». Злополучный рейс 401 авиакомпании Eastern Air Lines потерпел крушение в 1972 году, потому что экипаж сосредоточил внимание на перегоревшей индикаторной лампочке шасси и не заметил, что случайно отключил автопилот; высота (ключевой ограничительный показатель) постепенно уменьшалась, и самолет разбился в окрестностях города Эверглейдс во Флориде, в результате чего погиб 101 человек. Повышение операционной эффективности может обеспечить компании долгосрочное преимущество, как отметили Портер в статье «Японские компании редко имеют стратегии» (1996) и Вариан в статье о методике Кайдзен (2007).

Сценарий 2. У вас есть продукт, у вас есть стратегия, но результаты говорят о том, что вам нужно подумать о смене направления

В первом сценарии контролируемые эксперименты – отличный инструмент для восхождения на холм. Если вы думаете о многомерном пространстве идей с OEC как о «высоте», которую нужно оптимизировать, то вы, возможно, делаете шаги к вершине. Но иногда, основываясь на внутренних данных о скорости изменений или внешних данных о темпах роста или других контрольных показателях, вам нужно подумать о смене направления: стремлении к другому месту в пространстве решений, которое может располагаться на более высоком холме, или изменении стратегии и OEC (а значит, и ландшафта местности).

В общем, мы рекомендуем всегда иметь портфель идей: в большинстве случаев следует вкладывать средства в попытки оптимизации «рядом» с текущим местоположением, но время от времени следует пробовать несколько радикальных идей, чтобы увидеть, приведет ли смена направления к подножию более высокого холма. Наш опыт показывает, что большинство резких перемен терпят неудачу (например, глубокий редизайн сайта), однако существует определенный компромисс между риском и вознаграждением: редкие успехи могут привести к большим вознаграждениям, перекрывающим многие неудачи.

Когда вы проверяете радикальные идеи, методика проведения и оценка результатов эксперимента имеют свои особенности. В частности, необходимо учитывать:

- *продолжительность экспериментов*. Например, при тестировании значительного изменения дизайна пользовательского интерфейса экспериментальные изменения, измеренные в краткосрочной перспективе, могут зависеть от эффектов новизны или неприятия изменений. Пря-

мое сравнение контрольной и подопытной группы может не показать истинный долгосрочный эффект. На двустороннем рынке тестирование изменения, если оно недостаточно велико, может не повлиять на рынок. Хорошая аналогия – кубик льда в очень холодной комнате: небольшое повышение температуры ближе к комнатной может быть незаметным, но как только вы превысите точку плавления (например, 0 °C), кубик льда растает. В этих сценариях могут потребоваться более длительные и масштабные эксперименты или альтернативные схемы, такие как эксперименты на уровне страны, использованные в приведенном выше примере качества рекламы Google (см. также главу 23);

- *количество протестированных идей*. Вам может понадобиться много разных экспериментов, потому что каждый эксперимент проверяет только определенную тактику, которая является составной частью общей стратегии. Единственный эксперимент, неспособный улучшить OEC, может быть связан с плохой конкретной тактикой, что не обязательно указывает на плохую общую стратегию. Эксперименты, по сути своей, проверяют конкретные гипотезы, тогда как охват стратегии намного шире. Тем не менее контролируемые эксперименты помогают уточнить стратегию или показать ее неэффективность и стимулировать смену направления. Если различные тактики, оцененные с помощью контролируемых экспериментов, раз за разом терпят неудачу, возможно, пора вспомнить высказывание Уинстона Черчилля: «Какой бы красивой ни была стратегия, вам следует время от времени смотреть на результаты». Около двух лет подряд Bing придерживался стратегии интеграции с социальными сетями, в частности с Facebook и Twitter, добавив на страницу третью панель с результатами социального поиска. После того как на стратегию было потрачено более 25 млн долл. без значительного влияния на ключевые показатели, от нее пришлось отказаться. Иногда очень трудно отказаться от уже сделанной большой ставки, но экономическая теория говорит нам, что неудачные ставки – это невозвратные затраты, и мы должны принимать дальновидные решения на основе доступных данных, которые собираются по мере проведения новых экспериментов.

Эрик Райс использует термин «достигнутый провал» для компаний, которые успешно, добросовестно и неукоснительно выполняют план, который, как выясняется впоследствии, был заведомо провальным. Вместо этого он дает совет:

> «Методология бережливого стартапа рассматривает действия стартапа как эксперименты, проверяющие его стратегию, чтобы понять, какие части стратегии блестящие, а какие провальные. Настоящий эксперимент следует научному подходу. Он начинается с четкой гипотезы, которая предсказывает, что должно произойти. Затем он эмпирически проверяет эту гипотезу».

Из-за того, что проведение экспериментов для оценки стратегии требует времени и трудностей, некоторые авторы, например Синофски и Янсити, пишут:

> «… процесс разработки продукта чреват риском и неопределенностью. Это два очень разных понятия. Мы не можем уменьшить неопределенность – вы не знаете, чего вы не знаете».

Мы не согласны: возможность проводить контролируемые эксперименты позволяет значительно снизить неопределенность, тестируя минимально жизнеспособный продукт, получая данные и применяя их. Тем не менее не у всех есть в запасе несколько лет, чтобы потратить их на тестирование новой стратегии, и в этом случае вам, возможно, придется принимать решения в условиях неопределенности.

Есть одна полезная концепция, о которой следует помнить, – это *ожидаемая ценность информации* (expected value of information, EVI), предложенная Дугласом Хаббардом. Она показывает, как дополнительная информация может помочь вам в принятии решений. Возможность проводить контролируемые эксперименты позволяет значительно уменьшить неопределенность, пробуя минимально жизнеспособный продукт, собирая новые данные и повторяя эксперименты.

1.8. Дополнительное чтение

Есть несколько книг, непосредственно посвященных онлайн-экспериментам и A/B-тестированию (Siroker and Koomen 2013, Goward 2012, Schrage 2014, McFarland 2012, King et al.2017). В большинстве своем книги предлагают отличные мотивирующие истории, но мало говорят о статистике. Последняя книга Георгия Георгиева включает исчерпывающие объяснения математической статистики.

Литература, относящаяся к контролируемым экспериментам, обширна (Mason et al.1989, Box et al.2005, Keppel, Saufley and Tokunaga 1992, Rossi, Lipsey and Freeman 2004, Imbens and Rubin 2015, Pearl 2009, Angrist and Pischke 2014, Gerber and Green 2012).

Существует несколько учебных пособий по проведению контролируемых экспериментов в сети (Peterson 2004, 76–78, Eisenberg 2005, 283–286, Chatham, Temkin and Amato 2004, Eisenberg 2005, Eisenberg 2004; Peterson 2005, 248–253, Tyler and Ledford 2006, 213–219, Sterne 2002, 116–119, Kaushik 2006).

Многорукий бандит (multi-armed bandit) – это тип эксперимента, в котором распределение трафика эксперимента может динамически обновляться по мере продвижения эксперимента. Например, мы можем по-новому оценивать эксперимент каждый час, чтобы увидеть, как работает каждый из вариантов, и регулировать долю трафика, которую получает каждый ва-

риант. Вариант, который выглядит хорошо, получает больше трафика, а вариант, который неэффективен, получает меньше.

Эксперименты, основанные на многоруких бандитах, обычно более эффективны, чем «классические» A/B-эксперименты, потому что они постепенно перемещают трафик в сторону выигрышных вариантов, а не дожидаются окончания эксперимента. Несмотря на то что существует широкий спектр проблем, для решения которых они подходят, некоторые основные ограничения заключаются в том, что целью оценки должен быть один OEC (например, можно просто сформулировать компромисс между несколькими показателями) и что OEC должен быть достаточно хорошо измерим между перераспределениями, например рейтинг кликов по сравнению с сеансами. Также может возникнуть потенциальная предвзятость, возникающая из-за того, что пользователи, показавшие плохой результат, неравномерно перераспределяются среди выигрышных вариантов.

В декабре 2018 года трое соавторов этой книги организовали Первый практический саммит контролируемых онлайн-экспериментов. Тринадцать организаций, в том числе Airbnb, Amazon, Booking.com, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Яндекс и Стэнфордский университет, прислали в общей сложности 34 эксперта, которые представили обзор и обсудили проблемы, возникшие в ходе секционных заседаний. Читателям, интересующимся проблемами контролируемых экспериментов, будет полезно ознакомиться с материалами этого саммита.