

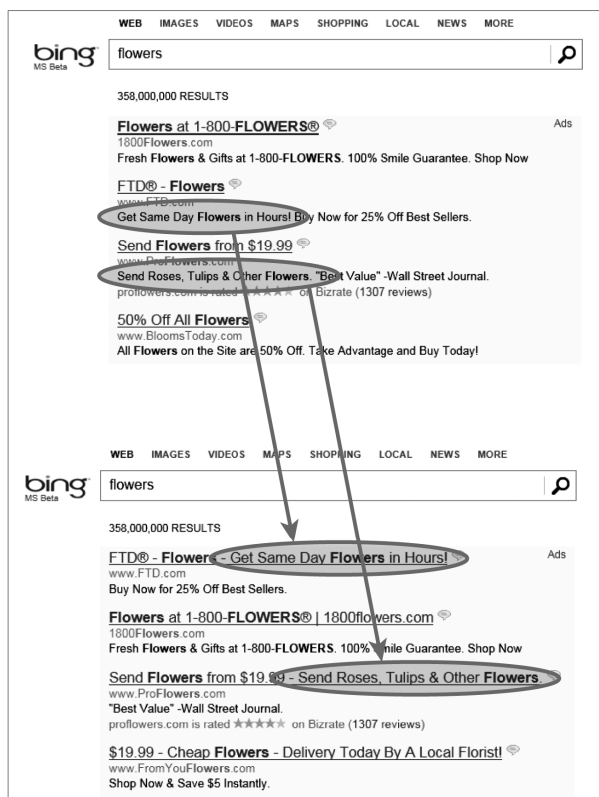
Rozdział 1.

Wstęp i motywacja

Jeden dokładny pomiar jest wart więcej niż tysiąc eksperckich opinii.

— *Admirał Grace Hopper*

W 2012 r. pewien pracownik zespołu ds. silnika wyszukiwarki Microsoft Bing zaproponował, aby zmienić sposób wyświetlania nagłówków reklam (Kohavi i Thomke, 2017). Pomysł polegał na tym, żeby wydłużyć linię tytułową reklamy, łącząc ją z pierwszym wersem tekstu pod tytułem (zobacz rysunek 1.1).



Rysunek 1.1. Eksperyment zmieniający sposób wyświetlania reklam w wyszukiwarce Bing

Nikt się nie spodziewał, że ta prosta modyfikacja wśród tysięcy innych sugestii będzie pomysłem generującym największe zyski w historii Binga!

Wspomniana funkcjonalność otrzymała niski priorytet i czekała w odmętach rejestru produktu przeszło pół roku, aż pewien programista sięgnął po to zadanie, zachęcony tym, jak łatwo było tę zmianę wprowadzić w kodzie. Zaimplementował koncepcję i zaczął oceniać jej skuteczność w testach z prawdziwymi użytkownikami, pokazując niektórym nowy układ strony, a innym stary. Interakcje użytkowników były nagrywane, łącznie z kliknięciami reklam i wygenerowanymi w ten sposób dochodami. To właśnie przykład testu A/B — najprostszej wersji kontrolowanego eksperymentu porównującego dwie opcje: A i B, czyli *grupę eksperymentalną i grupę kontrolną*.

Po kilku godzinach od rozpoczęcia testu uruchomił się alarm zbyt wysokiego dochodu, ostrzegający, że coś poszło nie tak z eksperymentem. Grupa eksperymentalna, czyli ta z nowym układem strony, generowała zbyt duży dochód z reklam. Takie ostrzeżenia o charakterze „to zbyt piękne, by mogło być prawdziwe” już wcześniej okazały się bardzo pożyteczne i z reguły wskazywały poważne defekty, takie jak podwójne rejestrowanie przychodu (podwójne obciążenia) czy wyświetlanie jedynie reklam, podczas gdy reszta strony była uszkodzona.

Jednak w przypadku tego eksperymentu wzrost przychodu nie był błędem. Dochód Binga wzrósł o kosmiczne 12%, co w tamtym okresie odpowiadało ponad 100 milionom dolarów w samych tylko Stanach Zjednoczonych, i to bez szkody dla kluczowych wskaźników z zakresu wrażeń użytkownika. Eksperyment powtarzano wielokrotnie w długich okresach.

Ten przykład odzwierciedla kilka kluczowych prawideł kontrolowanych eksperymentów online:

- Trudno ocenić wartość pomysłu. W tym przypadku prostą zmianę wartą 100 milionów dolarów odkładano miesiącami.
- Małe zmiany mogą mieć olbrzymi wpływ! Zwrot na poziomie 100 milionów dolarów rocznie z inwestycji w postaci kilku dni pracy programisty to jeden z najbardziej wyrazistych przykładów.
- Eksperymenty o dużym wpływie rzadko się zdarzają. Bing przeprowadza ponad 10 tysięcy eksperymentów rocznie, ale proste funkcjonalności stanowiące tak duże usprawnienie zdarzają się zaledwie raz na kilka lat.
- Narzut związany z przeprowadzeniem eksperymentu musi być mały. Inżynierowie Binga mają dostęp do ExP, systemu eksperymentalnego Microsoftu, co sprawia, że naukowa ocena pomysłu jest łatwa.
- Ogólne kryterium ewaluacji (OKE, opisywane szerzej w dalszej części rozdziału) musi być jasne. W tym przypadku dochód był jego głównym składnikiem, ale sam dochód nie jest wystarczający jako OKE. Mogłoby to doprowadzić do oblepiania stron reklamami, co — jak udowodniono

— niekorzystnie wpływa na wrażenia użytkownika. Bing posługuje się kryterium bilansującym dochód i wskaźniki związane z wrażeniami użytkownika, m.in. liczbą sesji na użytkownika (czy użytkownicy porzucają przeglądanie strony, czy też zwiększają swoje zaangażowanie). Zasadnicze znaczenie ma fakt, że wskaźniki UI nie ucierpiały znacząco, podczas gdy dochód poszybował w górę.

Kolejny podrozdział wprowadza pojęcia związane z eksperymentami online.

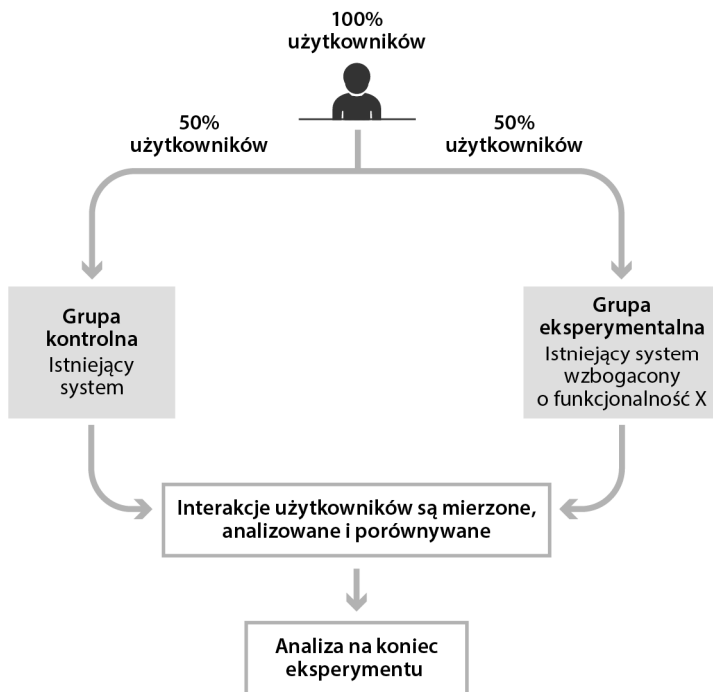
Żargon kontrolowanych eksperymentów online

Kontrolowane eksperymenty online mają długą i fascynującą historię, którą prezentujemy w sieci (Kohavi, Tang i Xu, 2019). Czasami nazywa się je testami *A/B* lub *A/B/n* (podkreślając mnogość wariantów), eksperymentami terenowymi, randomizowanymi eksperymentami kontrolowanymi, testami dzielonymi itd. W tej książce posługujemy się jednak pojęciami *eksperymenty kontrolowane* i *testy A/B* wymiennie, niezależnie od liczby wariantów.

Tego rodzaju eksperymenty są na porządku dziennym w firmach takich jak Airbnb, Amazon, Booking.com, eBay, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yahoo!/Oath czy Yandex (Gupta i inni, 2019). W tych korporacjach przeprowadza się tysiące lub dziesiątki tysięcy eksperymentów rocznie, czasami obejmujących miliony użytkowników i testujących wszystko — od zmian interfejsu użytkownika (UI), przez algorytmy dopasowania (wyszukiwanie, reklamy, personalizację itd.), opóźnienie/ wydajność, systemy zarządzania treścią (CMS), po systemy wsparcia klienta i wiele innych. Eksperymenty przeprowadza się różnymi kanałami: na stronach internetowych, w aplikacjach desktopowych lub mobilnych i poprzez e-mail.

W najbardziej powszechnym schemacie kontrolowanego eksperymentu online użytkownikom losowo, ale trwale przypisuje się któryś z wariantów (użytkownikowi prezentuje się ten sam wariant podczas wielu wizyt). W naszym pierwszym przykładzie z wyszukiwarką Bing grupa kontrolna widziała pierwotny sposób wyświetlania reklam, a grupie eksperymentalnej prezentowano układ z dłuższymi tytułami. Interakcje użytkowników ze stroną Bing podlegały pomiarowi, tzn. były monitorowane i logowane. Z tych zarejestrowanych danych wyliczono wskaźniki, co pozwoliło ocenić różnicę między wariantami w każdym mierzonym zakresie.

W najprostszych eksperymentach kontrolowanych wykorzystuje się dwa warianty: grupę kontrolną (A) i grupę eksperymentalną (B) (zobacz rysunek 1.2).



Rysunek 1.2. Prosty eksperyment kontrolowany: test A/B

Przyjęliśmy terminologię wprowadzoną przez Kohaviego i Longbottoma (2017) oraz Kohaviego, Longbottoma i innych (2009), a kilka pojęć z innych obszarów prezentujemy także poniżej. Wiele innych materiałów na temat eksperymentowania i testów A/B znajdziesz na końcu tego rozdziału, w podrozdziale „Dodatkowe źródła”.

Ogólne kryterium ewaluacji (OKE) — ilościowy wskaźnik celu eksperymentu. Przykładowo Twoim OKE może być liczba dni aktywności na użytkownika, odpowiadająca liczbie dni w trakcie trwania eksperymentu, gdy użytkownik wykazywał aktywność na stronie czy w aplikacji (tzn. odwiedzał ją i wykonywał jakieś działanie). Rosnące OKE oznacza, że użytkownicy odwiedzają Twoją stronę częściej, co jest znakomitym wynikiem. OKE musi być mierzalne w krótkim czasie (okresie trwania eksperymentu), ale jednocześnie powinno pozostawać w relacji przyczynowości w stosunku do długoterminowych celów strategicznych (zobacz podrozdział „Strategia, taktyka i ich związek z eksperymentami” w dalszej części tego rozdziału oraz rozdział 7.). W przypadku silnika wyszukiwarki ogólnym kryterium ewaluacji może być połączenie użycia (np. liczba sesji na użytkownika) i trafności (np. liczba sesji zakończonych powodzeniem lub czas potrzebny na zakończenie sesji sukcesem), a także dochodu z reklam (nie wszystkie silniki wyszukiwań używają wszystkich tych wskaźników czy też wyłącznie ich).

W statystyce nazywa się to często **zmienną zależną** bądź **zmienną odpowiedzi** (Mason, Gunst i Hess, 1989; Box, Hunter i Hunter 2005). Mówi się także czasem o **wyniku**,

ewaluacji czy **funkcji dopasowania** (Quarto-vonTivadar, 2006). Eksperymenty mogą mieć wiele celów, a w analizie można posługiwać się kartą ważonych wyników (Kaplan i Norton, 1996), choć wybranie jednego wskaźnika — choćby ważonej wartości łącznej dla wszystkich celów — jest pożądane i zalecane (Roy, 2001).

Szerzej opisujemy dobór OKE dla eksperymentów w rozdziale 7.

Parametr — kontrolowana zmienna eksperymentalna, o której przypuszcza się, że wpływa na OKE lub inne wskaźniki brane pod uwagę. Parametry czasami nazywa się *czynnikiem* lub *zmiennymi*. Przypisuje się im *wartości*, znane też jako *poziomy*. W prostych testach A/B występuje zazwyczaj jeden parametr o dwóch wartościach. W świecie działań online powszechnie używa się projektów o jednym parametrze, ale wielu wartościach (np. A/B/C/D). Testy z wieloma zmiennymi (*testy wielowymiarowe*, MVT) służą ocenie kilku parametrów (zmiennych) łącznie, np. koloru czcionki i jej wielkości, i pozwalają znaleźć optymalne rozwiązanie w przypadku współoddziaływania parametrów (zobacz rozdział 4.).

Wariant — wrażenia użytkownika podlegające testom, zwykle wynikającym z przypisania wartości parametrom. W prostych testach A/B występują zazwyczaj dwa warianty, nazywane grupą kontrolną i grupą eksperymentalną. W niektórych źródłach pojęcie wariantu odnosi się wyłącznie do grupy eksperymentalnej. My natomiast uznajemy grupę kontrolną za specjalny rodzaj wariantu — bieżącą wersję, wobec której następuje porównanie. Przykładowo jeśli w wyniku eksperymentu wykryje się defekt, eksperyment należy przerwać i upewnić się, że wszyscy użytkownicy znajdują się w grupie kontrolnej.

Randomizacja — proces pseudolosowego podziału (np. przy użyciu haszowania) stosowany do badanych jednostek (np. użytkowników lub stron) w celu przypisania ich do wariantów. Odpowiednia randomizacja jest ważna, aby populacje przypisane do różnych wariantów były statystycznie podobne, umożliwiając wskazanie przyczynowości z dużym prawdopodobieństwem. Randomizacja powinna przebiegać w sposób spójny i niezależny (np. jeśli badaną jednostką jest użytkownik, to powinien on spójnie, cały czas widzieć ten sam wariant, a jego przypisanie nie powinno mówić nic na temat przypisania innego użytkownika do innego wariantu). Posługiwanie się użytkownikiem jako badaną jednostką w przypadku kontrolowanych eksperymentów przeprowadzanych na odbiorcach online jest powszechne i my także zalecamy takie podejście. Niektóre schematy eksperymentowania stosują randomizację na podstawie strony, sesji czy dnia użytkowania (tzn. eksperyment przedstawia użytkownikowi spójną wersję w 24-godzinny oknie czasu określanego przez serwer). Więcej informacji na ten temat znajdziesz w rozdziale 14.

Właściwa randomizacja jest kluczowa! Jeśli projekt eksperymentu przypisuje taki sam procent użytkowników do każdego z wariantów, to każdy użytkownik powinien mieć taką samą szansę bycia przypisanym do każdego wariantu. Nie traktuj tego tematu

z pobłażaniem! Przykłady poniżej pokazują wagę odpowiedniej randomizacji i związane z nią wyzwania:

- RAND Corporation w latach 40. XX w. potrzebowała liczb losowych do metody Monte Carlo. Stworzyła więc tablicę miliona liczb losowych za pomocą maszyny generującej impulsy binarne. Jednak okazało się, że ze względu na błąd sprzętowy pierwotna tablica była mocno tendencyjna i cyfry trzeba było wygenerować ponownie w nowym wydaniu tablicy (RAND, 1955).
- Eksperymentów kontrolowanych używano początkowo w medycynie. Rządowa jednostka Stanów Zjednoczonych — Veterans Administration — przeprowadzała eksperyment (testy leków) dotyczący stosowania streptomycyny w leczeniu gruźlicy. Testy dały jednak negatywny wynik, ponieważ lekarze wpływali na proces doboru pacjentów i wprowadzali w ten sposób przekłamanie wyników (Marks, 1997). Podobne badania wykonano w Wielkiej Brytanii ze ślepą próbą i dały one pozytywny wynik, co stanowiło punkt zwrotny w badaniach kontrolowanych (Doll, 1998).

Nie należy dopuszczać, aby jakikolwiek czynnik miał wpływ na przypisanie wariantów. Użytkownicy (jednostki badane) nie mogą być przypisywani „jak popadnie” (Weiss, 1997). Należy także podkreślić, że losowość nie oznacza sposobu „pospiesznego czy nieplanowanego, lecz świadomy wybór dokonany na podstawie prawdopodobieństwa” (Mosteller, Gilbert i McPeck, 1983). Senn (2012) rozprawia się z innymi mitami na temat randomizacji.

Po co eksperymentować?

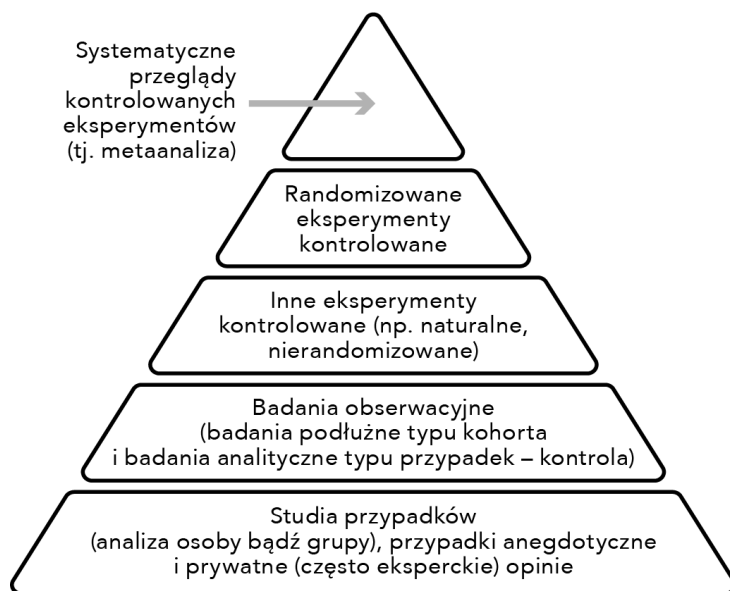
Korelacja, przyczynowość i wiarygodność

Przypuśćmy, że pracujesz w firmie, której model biznesowy opiera się na subskrypcjach — tak jak w Netflixie — gdzie $X\%$ użytkowników rezygnuje każdego miesiąca. Decydujesz się wprowadzić pewną nową funkcjonalność i zauważasz, że wskaźnik rezygnacji dla użytkowników, którzy mają z nią styczność, wynosi $X\%/2$, czyli połowę. Może kusić Cię myśl o wykrytej przyczynowości: funkcjonalność zmniejsza pułap rezygnacji o połowę. Prowadzi to do konkluzji, że jeśli tę funkcjonalność łatwiej będzie znaleźć, a użytkownicy będą z niej częściej korzystać, to liczba subskrypcji wzrośnie. Błąd! Biorąc pod uwagę dane, jakimi dysponujesz, nie możesz wyciągać żadnego wniosku na temat tego, czy dana funkcjonalność zwiększa, czy zmniejsza liczbę rezygnacji — jedno i drugie jest możliwe.

Przykład obrazujący ten błąd pochodzi z Microsoft Office 365, produktu opartego na subskrypcji. Wśród użytkowników Office 365, którzy widzą komunikaty błędów i doświadczają problemów z funkcjonowaniem aplikacji, wskaźnik rezygnacji jest niższy.

Nie oznacza to jednak, że Office 365 powinien wyświetlać więcej błędów, a Microsoft obniżyć jakość kodu, by powodować więcej awarii aplikacji. Okazuje się, że wszystkie trzy zjawiska są spowodowane tym samym czynnikiem: używaniem. Użytkownicy, którzy dużo korzystają z tego oprogramowania, widzą więcej błędów, aplikacja częściej nagle się im wyłącza, a oni rzadziej rezygnują z subskrypcji. Korelacja nie zakłada przyczynowości i przesadne poleganie na tego typu obserwacjach prowadzi do błędnych decyzji.

W 1995 r. Guyatt i inni (1995) wprowadzili hierarchię dowodów jako sposób oceny zaleceń w literaturze medycznej, a Greenhalgh pogłębiła tę propozycję w swoim opisie praktykowania medycyny opartej na dowodach (1997, 2014). Rysunek 1.3 przedstawia prostą hierarchię dowodów przetłumaczoną na nasz żargon w oparciu o pracę Bailara (1983). Randomizowane eksperymenty kontrolowane to złoty standard wykrywania przyczynowości. Systematyczne przeglądy, czyli metaanaliza, kontrolowanych eksperymentów udostępniają więcej dowodów i lepszą możliwość uogólnienia.



Rysunek 1.3. Prosta hierarchia dowodów służąca ocenie projektu próby (Greenhalgh, 2014)

Funkcjonują również bardziej złożone modele, takie jak *Levels of Evidence* (poziomy dowodów) przedstawiony przez Oxford Centre for Evidence-based Medicine (2009).

Platformy do eksperymentowania w naszych firmach umożliwiają pracownikom Google’a, LinkedIna czy Microsoftu przeprowadzanie dziesiątek tysięcy kontrolowanych eksperymentów online rocznie z wysokim zaufaniem co do ich wyników. Wierzymy, że kontrolowane eksperymenty online:

- Są najlepszą naukową metodą określania przyczynowości z dużym prawdopodobieństwem.
- Wykrywają małe zmiany, które trudno uchwycić w przypadku innych technik, np. zmiany w czasie (czułość).
- Wykrywają nieoczekiwane zmiany. Często się tego nie docenia, ale wiele eksperymentów naświatla zaskakujący wpływ na różne wskaźniki, np. spadek wydajności, wzrost awaryjności czy żerowanie jednej funkcjonalności na innej.

Zasadniczym celem tej książki jest podkreślenie możliwych pułapek w eksperymentach i sugerowanie metod budujących uzasadnione zaufanie do wyników. Kontrolowane eksperymenty online przedstawiają jedyną w swoim rodzaju szansę na elektroniczne zbieranie rzetelnych danych na dużą skalę, właściwe randomizowanie i unikanie czy wykrywanie błędów (zobacz rozdział 11.). Zalecamy używanie innych, mniej wiarygodnych technik, łącznie z badaniami obserwacyjnymi, gdy przeprowadzenie kontrolowanych eksperymentów online jest niemożliwe.

Niezbędne elementy przeprowadzania użytecznych eksperymentów kontrolowanych

Nie każdą decyzję można podjąć z rygiorem naukowym, jaki cechuje eksperymenty kontrolowane. Przykładowo nie możesz przeprowadzić takiego eksperymentu względem procesu fuzji (ang. *merger and acquisition*, M&A), ponieważ nie można jednocześnie pozostawać w stanie fuzji i jego przeciwieństwie. Przyjrzymy się teraz niezbędnym aspektom technicznym dotyczącym użytecznych eksperymentów kontrolowanych (Kohavi, Crook i Longbotham, 2009), a później przejdziemy do zasad organizacyjnych. W rozdziale 4. opisujemy model dojrzałości badawczej.

1. Istnieją badane jednostki (np. użytkownicy), którym można przypisać różne warianty bez (większych) utrudnień, np. grupa eksperymentalna nie ma wpływu na grupę kontrolną (zobacz rozdział 22.).
2. Badanych jednostek (np. użytkowników) jest wystarczająco dużo.
Aby kontrolowane eksperymenty były użyteczne, zalecamy tysiące badanych jednostek: im większa grupa, tym drobniejsze efekty można wykryć. Dobra wiadomość jest taka, że nawet niewielkie startupy związane z oprogramowaniem szybko zdobywają na tyle liczną grupę odbiorców, że mogą zacząć przeprowadzać kontrolowane eksperymenty w poszukiwaniu wpływowych zmian. W miarę jak biznes się rozrasta, coraz ważniejsze staje się rejestrowanie drobnych zmian (np. duże strony internetowe powinny być w stanie wykryć niewielkie wahnięcia wskaźników związanych z wrażeniami użytkownika i zmiany rzędu ułamka procenta w zakresie dochodu). Ta czułość rośnie wraz z poszerzeniem bazy odbiorców.

3. Kluczowe wskaźniki, a najlepiej OKE, są ustalone i można je w praktyce ocenić. Jeśli cele są zbyt trudne do zmierzenia wprost, ważne jest, by zgodzić się co do zamienników (zobacz rozdział 7.). Rzetelne dane można zebrać, najlepiej niewielkim kosztem i w szerokim spektrum. W kontekście oprogramowania zwykle łatwo rejestruje się zdarzenia systemowe i działania użytkownika (zobacz rozdział 13.).
4. Zmian dokonuje się łatwo. Oprogramowanie zwykle prościej zmienić niż sprzęt, ale nawet w oprogramowaniu pewne dziedziny wymagają szczególnej uwagi pod względem zapewnienia jakości. Zmiany w algorytmach rekomendacji łatwo wprowadzić i ocenić. Zmiany w systemach kontroli lotu statków powietrznych wymagają w USA całkowicie innego procesu akceptacji zmiany przez Federal Aviation Administration (FAA). Oprogramowanie po stronie serwera łatwiej zmienić niż to po stronie klienta (rozdział 12.), dlatego wywoływanie usług z oprogramowania klienta staje się coraz bardziej powszechne. Dzięki temu aktualizacje wersji i zmiany w usługach można przeprowadzać szybciej i z wykorzystaniem kontrolowanych eksperymentów online.

Większość nietrywialnych usług online spełnia lub może spełniać warunki niezbędne do przeprowadzania zwinnego procesu wytwarzania opierającego się na kontrolowanych eksperymentach. Wiele implementacji oprogramowania i usług również może sprostać tym wymaganiom względnie łatwo. Thomke pisał, że organizacje rozumieją największy zysk z eksperymentów, gdy używa się ich w połączeniu z „systemem innowacji” (Thomke, 2003). Zwinne wytwarzanie oprogramowania to właśnie taki system.

Gdy nie można przeprowadzić kontrolowanych eksperymentów, należy skorzystać z modelowania, a także posiłkować się innymi technikami eksperymentalnymi (zobacz rozdział 10.). Istotne jest, że gdy można przeprowadzić kontrolowane eksperymenty, to właśnie one są najbardziej rzetelnym i czułym mechanizmem oceny zmiany.

Zasady

Istnieją trzy kluczowe zasady, które obowiązują organizacje chcące przeprowadzać kontrolowane eksperymenty online (Kohavi i inni, 2013):

1. Organizacja chce podejmować decyzje na podstawie danych i sformalizowała OKE.
2. Organizacja jest skłonna zainwestować w infrastrukturę i testy potrzebne do przeprowadzania kontrolowanych eksperymentów, a także upewnić się, że wyniki są wiarygodne.
3. Organizacja zdaje sobie sprawę, że nie radzi sobie z oceną wartości pomysłów.

Zasada 1: organizacja chce podejmować decyzje na podstawie danych i sformalizowała OKE

Rzadko można spotkać przedstawiciela kadry zarządzającej, który oświadczy, że jego organizacja nie chce opierać się na danych. (Pomijając istotny wyjątek, jakim był Apple za czasów Steve'a Jobsa, gdy Ken Segall przyznał: „Nie testowaliśmy żadnej reklamy. Ani w druku, ani w telewizji, ani na billboardach, ani w sieci, ani w sklepach, ani nigdzie” (Segall, 2012)). Niemniej jednak mierzenie rosnących korzyści płynących z nowych funkcjonalności dla użytkownika ma swoją cenę, a obiektywne badania pokazują, że postęp realnie nie wygląda tak różowo, jak można to sobie początkowo wyobrażać. Wiele organizacji nie będzie zużywać zasobów potrzebnych do definiowania i mierzenia postępów. Często łatwiej jest wygenerować plan, postępować zgodnie z nim, odtrąbić sukces, gdy głównym wskaźnikiem jest „procent zrealizowanego planu”, i całkowicie zignorować to, czy funkcjonalność ma pozytywny wpływ na jakikolwiek inny, naprawdę ważny wskaźnik.

Aby organizacja rzeczywiście opierała się na danych, musi ona zdefiniować ogólne kryterium ewaluacji, które można łatwo mierzyć w niewielkich odstępach czasu (np. co tydzień lub dwa). Olbrzymie korporacje mogą posługiwać się kilkoma ogólnymi kryteriami ewaluacji lub szeregiem kluczowych wskaźników, które współdzielili się w różnych obszarach. Wyzwaniem jest znalezienie wskaźnika, który można mierzyć często, ale będzie on na tyle czuły, by pokazać różnice, i to takie, które będą obrazowały wpływ na długoterminowe cele. Przykładowo „zysk” nie jest prawidłowym OKE, ponieważ krótkofalowe zjawiska (np. podniesienie cen) mogą w krótkiej perspektywie zwiększyć zysk, ale zaszkodzić w dłuższym terminie. Wartość cyklu życia klienta jest strategicznie silnym OKE (Kohavi, Longbottom i inni, 2009). Nie da się przecenić wagi dojścia do porozumienia na temat dobrego ogólnego kryterium oceny, które będzie popierane w całej organizacji (zobacz rozdział 6.).

Czasami używa się pojęć takich jak organizacja „zorientowana na dane” czy „świadoma danych”, by uniknąć implikacji stwierdzenia, że cała organizacja opiera się na jednym źródle danych (np. kontrolowanych eksperymentach) (King, Churchill i Tan, 2017; Knapp i in., 2006). My jednak w tej książce stosujemy pojęcia „oparta na danych” i „zorientowana na dane” wymiennie. Ostatecznie decyzje powinno się podejmować na podstawie różnych źródeł danych, m.in. kontrolowanych eksperymentów, ankiet, szacunkowych kosztów utrzymania nowego kodu. Organizacja oparta na danych czy też zorientowana na nie zbiera odpowiednie dane, by kształtować decyzje i uświadamiać „najlepiej opłacaną osobę na sali”, zamiast polegać wyłącznie na intuicji (Kohavi, 2019).

Zasada 2: organizacja jest skłonna zainwestować w infrastrukturę i testy potrzebne do przeprowadzania kontrolowanych eksperymentów, a także upewnić się, że wyniki są wiarygodne

W obszarze oprogramowania online (stron internetowych, technologii mobilnych, aplikacji desktopowych i usług sieciowych) warunki niezbędne do przeprowadzania kontrolowanych eksperymentów można spełnić poprzez pracę inżynierów (zobacz podrozdział „Niezbędne elementy przeprowadzania użytecznych eksperymentów kontrolowanych”). Możliwe jest rzetelne randomizowanie użytkowników, tak samo jak korzystanie z telemetrii. Ponadto wprowadzanie zmian takich jak dodanie nowej funkcjonalności jest względnie łatwe (zobacz rozdział 4.). Nawet nieduże strony internetowe mają wystarczającą liczbę użytkowników, by przeprowadzić potrzebne testy statystyczne (Kohavi, Crook i Longbotham, 2009).

Kontrolowane eksperymenty są szczególnie użyteczne w zwinnych metodykach wytwarzania oprogramowania (Martin, 2008; K.S. Rubin, 2012), procesie Customer Development (Blank, 2005) i budowaniu MVP spopularyzowanym dzięki Ericowi Riesowi w książce *Metoda Lean Startup* (Ries, 2011).

W innych dziedzinach rzetelne przeprowadzanie kontrolowanych eksperymentów może być trudne lub wręcz niemożliwe. Pewne działania, jakie byłyby wymagane w eksperymentach kontrolowanych z zakresu medycyny, mogłyby być nieetyczne lub nielegalne. Sprzęt zwykle wytwarzany jest w długotrwałym i kłopotliwym do zmodyfikowania procesie, więc rzadko wykonuje się kontrolowane eksperymenty z użytkownikami nowych urządzeń (np. nowych telefonów komórkowych). W tych sytuacjach może zaistnieć potrzeba sięgnięcia po inne techniki, w tym opisane w podrozdziale „Techniki uzupełniające” rozdziału 10.

Zakładając, że przeprowadzenie kontrolowanych eksperymentów jest możliwe, na pierwszy plan wysuwa się zapewnienie ich wiarygodności. Jeśli chodzi o eksperymenty online, przytoczenie liczb jest proste. Przytoczenie liczb, którym można zaufać, jest trudne. Wiarygodności wyników poświęcony jest rozdział 3.

Zasada 3: organizacja zdaje sobie sprawę, że nie radzi sobie z oceną wartości pomysłów

Funkcjonalności tworzy się, ponieważ zespoły wierzą, że są one użyteczne, a jednak w wielu obszarach większość pomysłów nie przynosi poprawy kluczowych wskaźników. Zaledwie jedna trzecia pomysłów testowanych w Microsoftzie wpłynęła pozytywnie na wskaźnik, który zgodnie z projektem miała poprawić (Kohavi, Crook i Longbotham, 2009). Sukces jeszcze trudniej osiągnąć w mocno optymalizowanych dziedzinach, takich jak funkcjonowanie Binga czy Google’a, gdzie poziom powodzenia w przypadku niektórych wskaźników nie przekracza 10–20% (Manzi, 2012).

Fareed Mosavat, dyrektor produktu i cyklu życia w Slacku, napisał w mediach społecznościowych, że przy całym doświadczeniu Slacka i tak zaledwie 30% eksperymentów monetizacyjnych przynosi pozytywne rezultaty: „Jeśli pracujesz w zespole zorientowanym na eksperymenty, przyzwyczaisz się do tego, że w najlepszym przypadku 70% twojej pracy trafi do kosza. Twórz więc swoje procesy zgodnie z tą zasadą” (Mosavat, 2019).

Avinash Kaushik napisał w swoim elementarzu eksperymentowania i testów (Kaushik, 2006), że „w 80% przypadków mylimy się co do tego, czego chce klient”. Mike Moran (Moran, 2007, 2014) napisał, że Netflix uznaje około 90% swoich testów za nieudane. Regis Hadian z Quicken Loans przyznał: „Przez pięć lat przeprowadzania testów moja skuteczność w zgadywaniu wyników jest na poziomie skuteczności uderzenia piłki u zawodowego baseballisty. Dokładnie tak! Zajmuję się tym od pięciu lat i potrafię »zgadnąć« wynik testu w zaledwie 33% przypadków” (Moran, 2008). Dan McKinley z Etsy (McKinley, 2013) napisał, że „niemal wszystko zawodzi”, a jeśli chodzi o funkcjonalności: „Lekcją pokory jest obserwowanie, jak rzadko któraś odnosi sukces za pierwszym razem. Mam mocne przeczucie, że to doświadczenie powszechne, ale nikt nie chce tego przyznać”. Wreszcie Colin McFarland napisał w swojej książce *Experiment!* (McFarland, 2012): „Niezależnie od tego, jak często myślisz, że masz pewnik, ile badań zrobiłeś i którzy konkurenci już czegoś używają, czasem — częściej, niż myślisz — eksperymentalne pomysły po prostu zawodzą”.

Nie w każdej dziedzinie statystyki są tak słabe, ale większość osób, które przeprowadzały kontrolowane eksperymenty na aplikacjach i stronach internetowych kierowanych do szerokiej grupy odbiorców, doświadczyła tej lekcji pokory: *nie radzimy sobie z oceną wartości pomysłów*.

Poprawa w czasie

W praktyce do poprawy najważniejszych wskaźników dochodzi przez wiele małych zmian — wielkości 0,1 do 2%. Wiele eksperymentów ma wpływ jedynie na pewien segment użytkowników, więc musisz przeliczyć wzrost 5% w grupie stanowiącej 10%, co oczywiście daje znacznie mniejszy wpływ zmiany (np. 0,5%, jeśli *reagująca* populacja jest podobna do reszty użytkowników; zobacz rozdział 3.). Jak Al Pacino mówił w filmie *Męska gra*, „wygrywa się cal po calu”.

Przykład reklam Google’a

W 2011 r. Google wydał ulepszony mechanizm oceny reklam. Stało się to po przeszło roku pracy i przeprowadzanych przyrostowo eksperymentów (Google, 2011). Inżynierowie stworzyli i testowali nowy, lepszy model mierzenia poziomu jakości reklam w ramach bieżącego mechanizmu oceny rankingowej reklam, a także zmiany całego rankingu reklam. Przeprowadzili setki kontrolowanych eksperymentów w wielu iteracjach, niektóre

na wszystkich rynkach, inne długoterminowo na wybranych rynkach, by dogłębnie zrozumieć wpływ zmian na reklamodawców. Ta olbrzymia zmiana serwerowa — oraz przeprowadzone eksperymenty — zweryfikowała, jak planowanie różnych zmian i wprowadzanie ich warstwowo poprawi wrażenia użytkownika dzięki prezentowaniu reklam o wysokiej jakości, a także wrażenia reklamodawcy poprzez obniżenie średniej ceny za reklamy lepszej jakości.

Przykład trafności Binga

Zespół ds. trafności Binga zatrudnia kilkaset osób, których zadaniem jest poprawianie jednego wskaźnika OKE o 2% rocznie. Te 2% to suma efektu eksperymentalnego (tzn. delty w OKE) we wszystkich kontrolowanych eksperymentach udostępnionych użytkownikom w danym roku, przy założeniu, że można je dodać. Ponieważ zespół przeprowadza tysiące eksperymentów i niektóre pozytywne wyniki mogą być dziełem przypadku (Lee i Shen, 2018), wkład w docelowe 2% jest przyznawany dopiero po eksperymentcie replikacji: po udanej implementacji, a nawet po szeregu iteracji i poprawek, przeprowadza się eksperyment *certyfikujący* z jedną tylko grupą eksperymentalną. Efekt eksperymentalny w tym teście określa, jaki wkład dana zmiana ma do docelowych 2%. Ostatnie prace sugerują zmniejszenie efektu eksperymentalnego w celu większej precyzji (Coey i Cunningham, 2019).

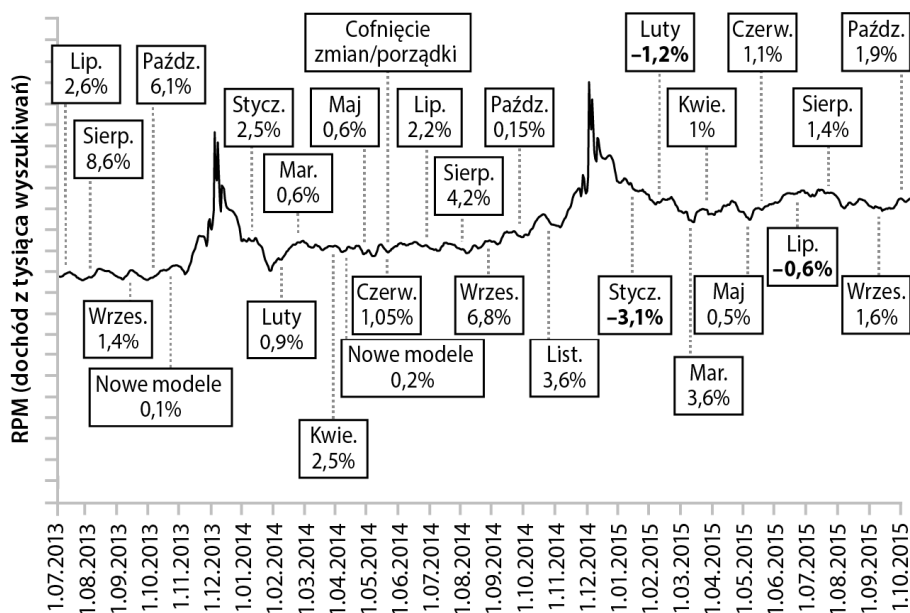
Przykład reklam Binga

Zespół ds. reklam Binga stale generował dochód rosnący o 15–25% rocznie (eMarketer, 2016), ale większość zmian zachodziła w drobnych krokach. Każdego miesiąca wydawano „pakiet” — efekt wielu eksperymentów (zobacz rysunek 1.4). Większość poprawek była niewielka, niektóre pakiety miały wręcz negatywny wpływ ze względu na ograniczenia miejsca czy zobowiązania prawne.

Cenną obserwacją jest odnotowanie sezonowych wahań w grudniu, gdy radykalnie rośnie zapotrzebowanie zakupowe użytkowników, więc miejsce na reklamy jest zwiększane, a dochód z tysiąca wyszukiwań rośnie.

Przykłady ciekawych kontrolowanych eksperymentów online

Ciekawe eksperymenty to te, w których różnica między wynikiem oczekiwanym a rzeczywistym jest duża. Jeśli spodziewałeś się, że coś się stanie, i rzeczywiście się wydarzyło, to nie nauczyłeś się wiele. Jeśli jednak spodziewałeś się, że coś się stanie, a to się nie wydarzyło — była to ważna lekcja dla Ciebie. Z kolei jeśli spodziewałeś się, że wydarzy się coś niedużego, a rzeczywisty wynik jest potężną niespodzianką prowadzącą do przełomu — dowiedziałeś się czegoś bardzo wartościowego.



(*) Liczby zmieniono z oczywistych względów

Rysunek 1.4. Dochód zespołu ds. reklam Binga w czasie (os Y przedstawia wzrost o ok. 20% rok do roku). Poszczególne liczby nie mają znaczenia

Przykład Binga z początku tego rozdziału oraz kolejne w tej sekcji to nieczęste przypadki sukcesów z zaskakującymi, bardzo pozytywnymi wynikami. Próba integracji Binga z Facebookiem i Twitterem to z kolei przykład, w którym oczekiwano lawinowego efektu i się go nie doczekano — po dwóch latach eksperymentów niewykazujących wartości pomysłu inicjatywę porzucono.

Choć utrzymywany postęp jest kwestią ciągłego eksperymentowania i wielu małych zmian, jak wskazaliśmy w punkcie „Przykład reklam Binga”, oto kilka przykładów podkreślających duże i zaskakujące efekty, które tylko potwierdzają, jak kiepsko oceniamy wartość pomysłów.

Przykład interfejsu użytkownika: 41 odcieni niebieskiego

Małe zmiany w projekcie mogą mieć spore znaczenie, co ustawicznie widać w Google'u i Microsoftzie. Google testował 41 odcieni niebieskiego na stronie wyników wyszukiwania Google'a (Holson, 2009), doprowadzając tym do frustracji kierownika zespołu projektantów. Niemniej jednak usprawnienia w palecie kolorów miały ostatecznie znaczący, pozytywny wpływ na zaangażowanie użytkowników (zwróć uwagę, że Google nie przedstawia danych na temat rezultatów pojedynczych zmian) i poskutkowały dalszą ścisłą współpracą projektantów i specjalistów od eksperymentów. Podobne modyfikacje

kolorów w Bing-u pokazały, że więcej użytkowników kończyło swoje wyszukiwania sukcesem, poprawiał się czas zakończenia działania powodzeniem, a monetyzacja rozkwitła o 10 milionów dolarów rocznie w Stanach Zjednoczonych (Kohavi i inni, 2014; Kohavi i Thomke, 2017).

Choć są to świetne przykłady drobnych zmian o gigantycznym wpływie, biorąc pod uwagę szeroką paletę testowanych kolorów, mało prawdopodobne jest, by dalsze manipulacje z kolorami w eksperymentach przyniosły kolejne znaczne poprawki.

Propozycja we właściwym momencie

W 2004 r. Amazon umieścił ofertę swojej karty kredytowej na głównej stronie sklepu. Była ona bardzo zyskowna, ale współczynnik użytkowników, którzy przeszli przez cały proces (ang. *click-through rate*, CTR), był bardzo niski. Zespół przeprowadził eksperyment, w którym przeniósł ofertę do widoku koszyka, który jest prezentowany po tym, jak użytkownik doda tam produkt. Prezentowane było także proste równanie pokazujące kwotę, jaką użytkownik mógłby zaoszczędzić, gdyby skorzystał z oferty — widać to na rysunku 1.5 (Kohavi i inni, 2014).

You could save \$30 today with the Amazon Visa® Card:



Your current subtotal: \$32.20

Amazon Visa discount: - \$30.00

Your new subtotal: \$2.20

[Find out how](#)

Save \$30 off your first purchase, earn **3% rewards**, get a **0% APR***, and pay **no annual fee**.

Rysunek 1.5. Oferta karty kredytowej Amazon przedstawiająca oszczędności w widoku podsumowania koszyka

Ponieważ użytkownicy, którzy dodali produkt do koszyka, mają zamiar dokonać zakupów, sugestia Amazona pojawia się we właściwym momencie. Kontrolowany eksperyment udowodnił, że ta prosta zmiana wzbogaci budżet firmy o dziesiątki milionów dolarów rocznie.

Spersonalizowane rekomendacje

Greg Linden stworzył w Amazonie prototyp służący do pokazywania użytkownikowi spersonalizowanych rekomendacji w oparciu o produkty znajdujące się w koszyku użytkownika (Linden, 2006; Kohavi, Longbottom i inni, 2009). Kiedy dodajesz produkt, pojawiają się rekomendacje. Dodaj coś innego — zobaczysz nowe sugestie. Linden zauważył, że choć prototyp wyglądał obiecująco, „wiceprezes ds. marketingu był stanowczo przeciwny temu pomysłowi”, twierdząc, że odwróci to uwagę użytkowników od dokończenia procesu zakupowego. Gregowi „zabroniono dalszej pracy nad

tym projektem”. Mimo to przeprowadził kontrolowany eksperyment, „a [jego] funkcjonalność zostawiła grupę kontrolną tak daleko w tyle, że okazało się, iż bez niej Amazon traci sporą kwotę pieniędzy. Z nowym priorytetem uruchomiono więc rekomendacje w koszyku zakupowym”. Obecnie z podobnego rozwiązania korzysta wiele stron internetowych.

Szybkość ma duże znaczenie

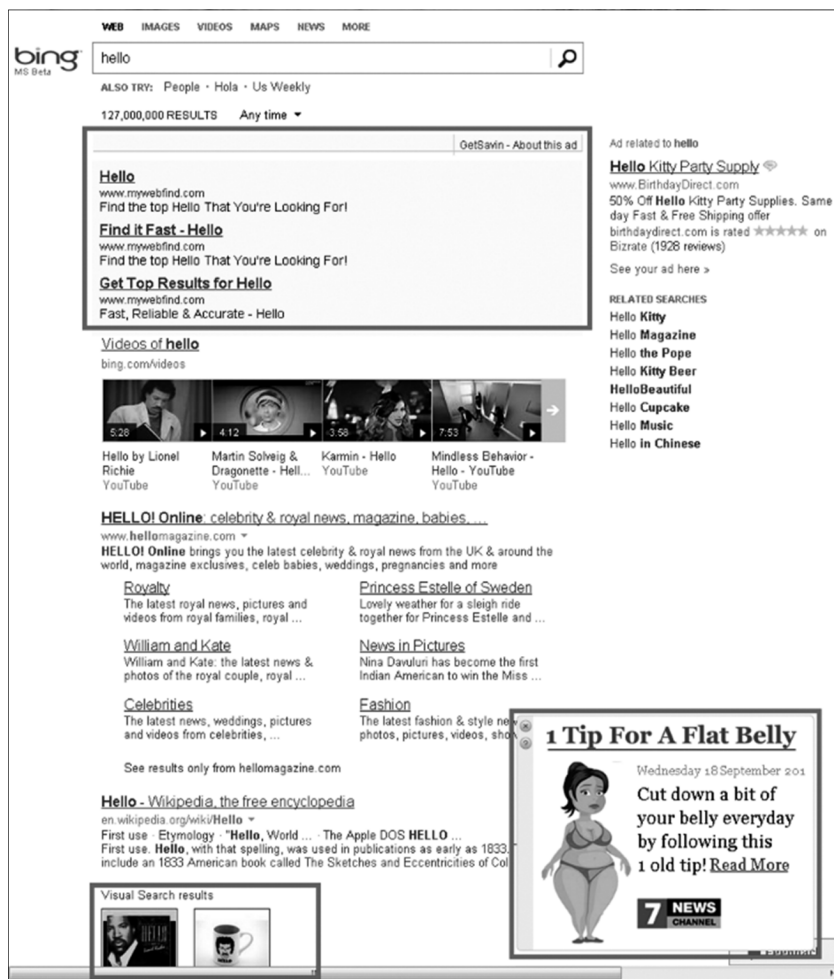
W 2012 r. pewien programista Binga dokonał zmiany sposobu generowania kodu w JavaScriptcie, co znacznie skróciło kod w HTML-u wysyłany do klienta, a dzięki temu bardzo poprawiło wydajność. Dotyczący tego kontrolowany eksperyment wykazał zaskakującą poprawę wielu wskaźników. Zespół przeprowadził kolejny eksperyment służący oszacowaniu wpływu na wydajność serwera. Wyniki pokazały, że wzrost wydajności znacznie poprawia także kluczowe wskaźniki dotyczące użytkowników, takie jak procent interakcji zakończonych powodzeniem i czas zakończenia działania z powodzeniem, a każde 10 milisekund zaoszczędzone w wydajności (czyli $\frac{1}{30}$ czasu mruknięcia okiem) pokrywa pełny koszt rocznego zatrudnienia inżyniera (Kohavi i inni, 2013).

Do 2015 r. wydajność Binga rosła, co zrodziło pytania o celowość dalszej optymalizacji, gdy serwer zwraca wyniki w czasie poniżej sekundy dla 95. centyla (tzn. dla 95% zapytań). Zespół Binga przeprowadził dalsze badania i kluczowe wskaźniki dotyczące użytkowników nadal znacznie się poprawiały. Choć względny wpływ na dochód nieco zmalał, to dochód Binga wzrósł w tym czasie tak bardzo, że każda milisekunda lepszej wydajności była warta więcej niż w przeszłości — już tylko 4 milisekundy poprawy sponsorowały roczny kontrakt programisty! W rozdziale 5. znajdziesz pełny opis tego eksperymentu i refleksję nad wagą wydajności.

Eksperymenty dotyczące wydajności przeprowadzano w różnych firmach, a ich wyniki wskazywały, jak ważny jest to aspekt. W Amazonie eksperyment zakładający 100 milisekund opóźnienia zmniejszył sprzedaż o 1% (Linden, 2006). Wspólne wystąpienie prelegentów z Binga i Google’a (Schurman i Brutlag, 2009) przedstawiało znaczący wpływ wydajności na kluczowe wskaźniki, takie jak liczba zapytań, dochód, liczba kliknięć, satysfakcja czy czas kliknięcia (ang. *time-to-click*).

Ograniczenie złośliwego oprogramowania

Reklamy to lukratywny biznes i duża część „darmowego oprogramowania”, które instalują użytkownicy, zawiera wirusy rozsiewające reklamy na stronach internetowych. Rysunek 1.6 pokazuje, jak wygląda strona z Binga dla użytkownika z takim złośliwym oprogramowaniem. Zwróć uwagę, że na stronie pojawiło się wiele reklam (oznaczonych ramką) (Kohavi i inni, 2014).



Rysunek 1.6. Gdy użytkownik padł ofiarą złośliwego oprogramowania, strona Binga wyświetla wiele reklam

Nie tylko usunięto reklamy Binga, pozbawiając w ten sposób firmę dochodu, ale też wyświetlono reklamy o niskiej jakości, często nietrafione, generujące złe wrażenia dla użytkownika, który może być nieświadomy powodu, dla którego widzi tak wiele reklam.

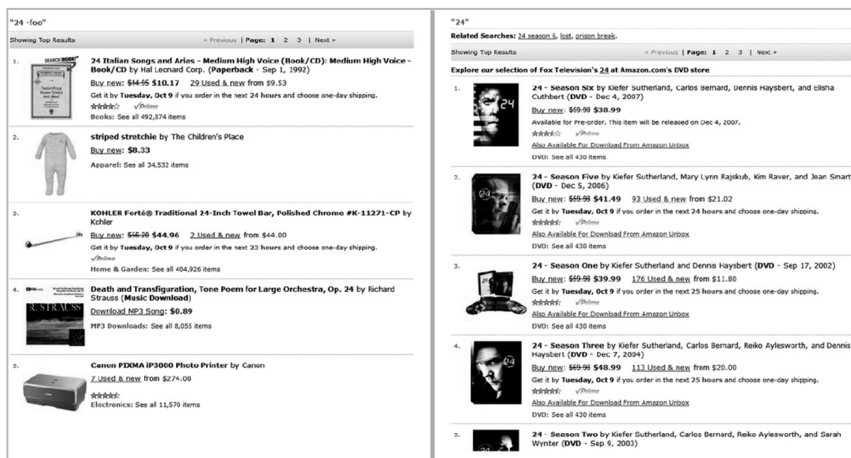
Microsoft przeprowadził kontrolowany eksperyment obejmujący potencjalnie 3,8 miliona użytkowników, podczas którego podstawowe procesy, które modyfikują obiektowy model strony (model DOM), zostały nadpisane tak, by zezwalać jedynie na pewne zmiany z zaufanych źródeł (Kohavi i inni, 2014). Wyniki wykazały poprawę wszystkich kluczowych wskaźników Binga, łącznie z liczbą sesji na użytkownika, sugerując, że użytkownicy chętniej odwiedzali strony lub rzadziej porzucali przeglądanie. Ponadto użytkownicy dokonywali wyszukiwań z większym powodzeniem, szybciej klikali użyteczne odnośniki, a roczny dochód wzrósł o kilka milionów dolarów. Ponadto czas

ładowania strony, podstawowy wskaźnik wydajności, o której była wcześniej mowa, poprawił się o setki milisekund w przypadku stron objętych eksperymentem.

Zmiany serwerowe

Serwerowe zmiany w algorytmach to obszar, który często pomija się w kontrolowanych eksperymentach (Kohavi, Longbottom i inni, 2009), a może on wiązać się ze znaczącymi wynikami. Widać to w sposobie pracy zespołów Google’a, LinkedIna i Microsoftu nad wieloma małymi, przyrostowymi zmianami opisanymi wyżej oraz w przykładzie zaczerpniętym z Amazona.

Już w 2004 r. istniał dobry algorytm rekomendacji oparty na dwóch zbiorach. Znakiem rozpoznawczym rekomendacji w Amazonie było „osoby, które kupiły produkt X, kupiły także produkt Y”, ale uogólniono to do twierdzeń „osoby, które *oglądały* produkt X, kupiły produkt Y” oraz „osoby, które *oglądały* produkt X, oglądały także produkt Y”. Zaproponowano więc, by użyć tego samego algorytmu dla przypadku „osoby, które *szukały* produktu X, kupiły produkt Y”. Zwolennicy algorytmu przytaczali przykłady niedookreślonych wyszukiwań, takich jak „24”, które większości Amerykanów kojarzy się z programem telewizyjnym z Kieferem Sutherlandem. Wyszukiwarka Amazona zwracała nieadekwatne wyniki (lewa strona rysunku 1.7), takie jak płyty CD z 24 włoskimi piosenkami, odzież dla 24-miesięcznych dzieci, 24-calowy wieszak na ręczniki itd. Nowy algorytm zwracał bezbłędne dopasowania (prawa strona rysunku 1.7), prezentując płyty DVD z serialem i powiązane książki, w oparciu o to, co ludzie, którzy wyszukiwali frazę „24”, rzeczywiście kupili. Jedną ze słabości algorytmu był jednak fakt, że część sugerowanych produktów nie miała w tytule wyszukiwanego zwrotu. Niemniej jednak Amazon przeprowadził kontrolowany eksperyment i mimo tej ułomności zmiana zwiększyła ogólny dochód firmy o 3% — setki milionów dolarów.



Rysunek 1.7. Wyszukiwanie hasła „24” na portalu Amazon z funkcjonalnością BBS i bez niej

Strategia, taktyka i ich związek z eksperymentami

Uważamy, że gdy wymagania dotyczące przeprowadzania kontrolowanych eksperymentów online zostaną spełnione, należy realizować te działania, by dostarczyć informacji na wszystkich poziomach decyzji organizacyjnych od strategii po taktykę.

Pomiędzy strategią (Porter, 1996, 1998) a kontrolowanymi eksperymentami zachodzi synergia. David Collins z Lean Strategy napisał, że „zamiast tłumić działania przedsiębiorcze, efektywna strategia zachęca do nich — poprzez wskazywanie ram, w których powinny zachodzić eksperymentowanie i innowacja” (Collins, 2016). Ten sam autor definiuje *proces strategii lean*, który chroni przed skrajnościami — ścisłym planowaniem z jednej strony i nieograniczonym eksperymentowaniem z drugiej.

Dobrze prowadzone eksperymenty z odpowiednimi wskaźnikami uzupełniają strategię biznesową i projektowanie produktu oraz poprawiają efektywność organizacyjną przez zorientowanie organizacji na dane. Przez ujęcie strategii w OKE kontrolowane eksperymenty mogą stanowić świetną informację zwrotną na jej temat. Czy pomysły są oceniane za pomocą eksperymentów właściwie podwyższających OKE? Ponadto zaskakujące wyniki eksperymentów mogą rzucić światło na alternatywne szanse strategiczne, prowadząc do zwrotu orientacji w tej kwestii (Ries, 2011). Decyzje z zakresu projektowania produktu są ważne ze względu na spójność, a wypróbowanie różnych wariantów projektu daje cenną informację zwrotną projektantom. Wreszcie wiele taktycznych zmian jest w stanie usprawnić efektywność operacyjną, którą Porter definiuje jako „wykonywanie podobnych działań lepiej niż konkurenci” (Porter, 1996).

Przyjrzyjmy się teraz dwóm zasadniczym scenariuszom.

Scenariusz 1: masz strategię biznesową i masz produkt z wystarczającą liczbą użytkowników

W tym scenariuszu eksperymenty mogą pomóc w osiągnięciu lokalnego optimum w oparciu o Twój bieżący produkt i strategię:

- Eksperymenty mogą wskazać obszary wysokiego zwrotu z inwestycji (ROI) — te, które najbardziej poprawiają OKE w stosunku do wymaganego nakładu. Wypróbowywanie różnych dziedzin z MVP pomaga zbadać większy zakres rynku szybciej, przed podjęciem znaczących zobowiązań.
- Eksperymentowanie może również pomóc z optymalizacjami, które mogą nie być oczywiste dla projektantów, ale mają olbrzymie znaczenie (np. kolory, interlinia, wydajność).
- Eksperymenty pomagają w ciągłych iteracjach prowadzących do skuteczniejszego przeprojektowywania strony zamiast skazywać zespoły na pracę nad jedną, całkowitą reorganizacją, która często wystawia

użytkowników na efekt pierwszeństwa (użytkownicy są *przygotowani* na poprzednią funkcjonalność, czyli to, jak strona działała kiedyś) i przez to uniemożliwia nie tylko osiągnięcie założonych celów, ale nawet dorównanie poprzedniej wersji pod względem kluczowych wskaźników (Goward, 2015; Rawat, 2018; Wolf, 2018; Laja, 2019).

- Eksperymenty mogą mieć kluczowe znaczenie w optymalizacji algorytmów serwerowych i infrastruktury, np. algorytmów rekomendacji i rankingu.

Posiadanie strategii ma zasadnicze znaczenie dla przeprowadzania eksperymentów — to ona kieruje wyborem OKE. Po zdefiniowaniu tego wskaźnika kontrolowane eksperymenty pomagają przyspieszyć innowacje przez danie zespołom mocy do optymalizacji i poprawy OKE. Tam, gdzie obserwowaliśmy nieprawidłowości w posługiwaniu się eksperymentami, wynikało to z wadliwie wybranego wskaźnika OKE. Jego składowe powinny odzwierciedlać kluczowe charakterystyki i nie podlegać manipulacji (zobacz rozdział 7.).

W naszych firmach mamy nie tylko zespoły zajmujące się prawidłowym przeprowadzaniem eksperymentów, ale też zespoły odpowiedzialne za wskaźniki: wybieranie ich, walidowanie i zmienianie w czasie. Ewolucja wskaźników będzie zachodziła zarówno dlatego, że z czasem będzie zmieniać się Twoja strategia, jak i dlatego, że poznasz ograniczenia swoich obecnych metryk, jak choćby ich podatność na manipulację. Zespoły zajmujące się wskaźnikami pracują też nad tym, które wskaźniki mierzalne w krótkim czasie wpływają na długofalowe cele, bo zwykle przecież eksperymenty realizuje się w węższych ramach czasowych. Hausner i Katz (1998) pisali, że „firma musi zidentyfikować wskaźniki, na które zespół może wpłynąć dziś, ale które ostatecznie wpłyną na długofalowe cele firmy” (zobacz rozdział 7.).

Łączenie strategii i OKE tworzy też **strategiczną integralność** (Sinofsky i Iansiti, 2009). Autorzy wskazują, że „w integralności strategicznej nie chodzi o stworzenie błyskotliwej strategii czy posiadanie perfekcyjnej organizacji. Chodzi o wprowadzenie właściwych strategii w organizacji, która je popiera i wie, jak je wprowadzić. Chodzi o dopasowanie odgórnjej perspektywy z oddolnymi zadaniami”. Wskaźnik OKE to znakomity mechanizm ujawnienia strategii i uzgodnienia, które funkcjonalności się z nią łączą.

W ostatecznym rachunku bez dobrego OKE marnujesz zasoby — to tak, jakbyś eksperymentował z poprawą wyżywienia czy oświetlenia na tonącym statku. Waga bezpieczeństwa pasażerów we wskaźniku OKE powinna być bardzo wysoka — tak wysoka, że nie powinniśmy być skłonni do obniżania bezpieczeństwa pasażerów. Można to uchwycić poprzez nadanie wysokiej wagi temu wskaźnikowi w OKE lub używanie bezpieczeństwa pasażerów jako wskaźnika ochronnego (zobacz rozdział 21.). W dziedzinie oprogramowania analogią do bezpieczeństwa pasażerów statku są awarie oprogramowania. Jeśli funkcjonalność zwiększa awaryjność produktu, to wrażenia uważane są za tak złe, że ewentualne zalety i tak w tym kontekście błędne.

Zdefiniowanie wskaźnika ochronnego w eksperymencie jest ważne ze względu na konieczność jasnego stwierdzenia, czego organizacja *nie jest* skłonna zmienić, ponieważ strategia „zmusza Cię także do dokonywania kompromisów — decydowania, czego nie robić” (Porter, 1996). Feralny lot 401 Eastern Air Lines skończył się katastrofą, ponieważ załoga skupiła się na przepalanej kontrolce blokady przednich kół i nie zauważyła, że mechanizm autopilota był przez przypadek wyłączony. Wysokość — kluczowy wskaźnik ochronny — spadała stopniowo i samolot rozbił się w bagnach Everglades na Florydzie w 1972 r., wskutek czego śmierć poniosło 101 pasażerów (Wikipedia, *Katastrofa lotu Eastern Air Lines 401*, 2019).

Poprawa wydolności organizacyjnej może przynieść długoterminową wyróżniającą się przewagę nad konkurencją, jak zauważył Porter w sekcji zatytułowanej „Japanese Companies Rarely have Strategies” (1996) oraz Varian w swoim artykule na temat metody kaizen (2007).

Scenariusz 2: masz produkt i strategię, ale wyniki wskazują, że należy rozważyć zmianę kierunku

W scenariuszu 1 kontrolowane eksperymenty są świetną pomocą w zdobywaniu nowych szczytów. Jeśli wyobrazisz sobie wielowymiarową przestrzeń pomysłów, z OKE jako „wysokością”, która podlega optymalizacji, to prawdopodobnie stawiasz kroki ku szczytowi. Jednak czasem — czy to w oparciu o wewnętrzne dane na temat tempa zmian, czy zewnętrzne dane na temat rozwoju, czy inny punkt odniesienia — musisz zastanowić się nad porzuceniem dotychczasowej drogi i zmianą kierunku. By posłużyć się poprzednią analogią: przeskoczeniem w inny obszar tej przestrzeni, który może znajdować się za większą górą — zmianą strategii i OKE (a tym samym zmianą ukształtowania terenu).

Zasadniczo zawsze zalecamy posiadanie całego portfolio pomysłów. Większość z nich powinna dotyczyć inwestycji w próbę optymalizacji „w pobliżu” obecnej lokalizacji, ale warto wypróbować kilka radykalnych pomysłów, by sprawdzić, czy prowadzą do lepszych widoków. Z naszego doświadczenia wynika, że większość wielkich, skokowych zmian się nie sprawdza (np. duże zmiany projektu strony), niemniej jednak bilans między ryzykiem a szansą pozostaje do przemyślenia. Rzadki sukces może prowadzić do dużych zysków, które zrekompensują nawet liczne porażki.

Przy testowaniu radykalnych pomysłów sposób przeprowadzania i oceny eksperymentów nieco się zmienia. Przede wszystkim musisz zastanowić się nad następującymi kwestiami:

- Czas trwania eksperymentu. Przykładowo, podczas testowania dużych zmian interfejsu użytkownika na odbiór zmiany eksperymentalnej może mieć wpływ efekt pierwszeństwa czy niechęć do zmiany. Bezpośrednie porównanie

grupy kontrolnej i grupy eksperymentalnej może nie zmierzyć prawdziwego, długofalowego efektu. W dwustronnej platformie typu marketplace testowanie zmiany, o ile nie będzie wystarczająco duża, może nie wywołać efektu. Dobrą analogią będzie kostka lodu w bardzo zimnym pomieszczeniu: małe zmiany temperatury mogą ująć uwagę, ale po przekroczeniu temperatury topnienia kostka lodu się rozpułynie. W takich okolicznościach mogą być potrzebne dłuższe i większe eksperymenty czy alternatywne projekty, takie jak eksperymenty krajowe w powyższym przykładzie jakości reklam Google'a (zobacz rozdział 23.).

- Liczba testowanych pomysłów. Możesz potrzebować wielu różnych eksperymentów, bo każdy z nich testuje tylko konkretną taktykę, będącą komponentem całej strategii. Niepowodzenie jednego eksperymentu w poprawie OKE może wynikać z tego, że dana taktyka jest nieskuteczna, i nie świadczyć o wadliwości całej strategii. Zgodnie z zasadami sztuki eksperyment testuje jedną hipotezę, a strategia to szerszy byt. Niemniej jednak kontrolowane eksperymenty pomagają doprecyzować strategię lub wykazać jej nieskuteczność i zachęcić do zmiany (Ries, 2011). Jeśli wiele taktyk ocenianych w eksperymentach zawodzi, być może nadszedł czas, by pomyśleć nad powiedzeniem Winstona Churchilla: „Niezależnie od tego, jak piękną ma się strategię, trzeba od czasu do czasu przyjrzeć się jej efektom”. Przez mniej więcej dwa lata Bing utrzymywał strategię integracji z mediami społecznościowymi, zwłaszcza z Facebookiem i Twitterem, otwierając trzeci panel z wynikami wyszukiwania w mediach społecznościowych. Po wydaniu 25 milionów dolarów na strategię, która nie przyniosła żadnych istotnych zmian w kluczowych wskaźnikach, pomysł porzucono (Kohavi i Thomke, 2017). Może być trudno rozstać się z pomysłem, na który bardzo liczono, ale teoria ekonomii mówi, że takie chybione nadzieje to stracone pieniądze. Powinniśmy podejmować decyzje ze spojrzeniem w przód i w oparciu o dostępne dane, które uzyskujemy z przeprowadzania kolejnych eksperymentów.

Eric Ries posługuje się wyrażeniem „osiągnięcie porażki” w stosunku do firm, które z powodzeniem, wiernie i ściśle wykonały plan, który okazał się całkowicie chybiony (Ries, 2011). Zamiast tego zaleca:

Metoda lean startup postrzega wysiłki startupu jako eksperymenty, które testują jego strategię, by przekonać się, które jej części są olśniewające, a które po prostu szalone. Prawdziwy eksperyment posługuje się metodą naukową. Rozpoczyna się od jasnej hipotezy, w której przewiduje się, co się wydarzy. Następnie empirycznie testuje te przewidywania.

Ze względu na czas i wyzwania związane z przeprowadzeniem eksperymentów w celu ewaluacji strategii niektórzy, jak Sinofsky i Iansiti (2009), piszą:

(...) proces wytwarzania produktu [jest] naszpikowany ryzykiem i niepewnością. To dwa bardzo różne pojęcia (...) niepewności nie możemy ograniczyć — nie wiemy, czego nie wiemy.

My się jednak z tą myślą nie zgadzamy. Możliwość przeprowadzania kontrolowanych eksperymentów pozwala znacząco zmniejszyć niepewność przez wypróbowanie wersji MVP produktu (Ries, 2011), zebranie danych i dalszą pracę w iteracjach. Jednak nie każdy może mieć kilka lat, by zainwestować w testowanie strategii, a wtedy rzeczywiście trzeba podejmować decyzje z obciążeniem, jakim jest niepewność.

Pojęciem, które warto mieć w pamięci, jest EVI (ang. *Expected Value of Information* — spodziewana wartość informacji). To pojęcie stworzone przez Douglasa Hubbarda (2011) odnosi się do dodatkowych informacji, które mogą pomóc Ci podjąć decyzję. Możliwość przeprowadzania kontrolowanych eksperymentów pozwala znacząco zmniejszyć niepewność przez wypróbowanie wersji MVP produktu (Ries, 2011), zebranie danych i dalszą pracę w iteracjach.

Dodatkowe źródła

Istnieje wiele książek bezpośrednio poświęconych eksperymentom online i testom A/B (Siroker i Koomen, 2013; Goward, 2012; Schrage, 2014; McFarland, 2012; King i inni, 2017). Większość z nich zawiera bardzo motywujące historie, ale też niedokładne statystyki. Ostatnia książka Georgiego Georgieva zawiera pełne wyjaśnienia statystyczne (Georgiev, 2019).

Literatura na temat eksperymentów kontrolowanych jest olbrzymia (Mason i inni, 1989; Box i inni, 2005; Keppel, Saufley i Tokunaga, 1992; Rossi, Lipsey i Freeman, 2004; Imbens i Rubin, 2015; Pearl, 2009; Angrist i Pischke, 2014; Gerber i Green, 2012).

Pojawiło się także kilka przewodników po przeprowadzaniu kontrolowanych eksperymentów w sieci (Peterson 2004, s. 76–78; Eisenberg, 2005, s. 283–286; Chatham, Temkin i Amato, 2004; Eisenberg, 2005; Eisenberg, 2004; Peterson, 2005, s. 248–253; Tyler i Ledford, 2006, s. 213–219; Sterne, 2002, s. 116–119; Kaushik, 2006).

Wieloręki bandyta to typ eksperymentu, w którym ruch eksperymentalny można dynamicznie aktualizować w miarę postępu testów (Li i inni, 2010; Scott, 2010). Przykładowo można co godzinę odświeżać eksperyment, by przekonać się, jak radzą sobie warianty, i dopasować procent ruchu, który kierowany jest do każdego z wariantów. Ten, który wydaje się radzić sobie dobrze, uzyskuje większe obciążenie, a słabszy — mniejsze.

Eksperymenty tego rodzaju są zwykle bardziej wydajne niż „klasyczne” testy A/B, ponieważ stopniowo kierują ruchu ku wygrywającemu wariantowi zamiast czekać na zakończenie eksperymentu. Choć zakres problemów, które można w ten sposób rozwiązać, jest szeroki (Bakshy, Balandal i Kashin, 2019), do poważnych ograniczeń trzeba zaliczyć konieczność sformułowania jednego OKE jako celu ewaluacji (np. kompromisu pomiędzy różnymi wskaźnikami nie da się łatwo wypracować) oraz fakt, że OKE nie da się zmierzyć względnie dobrze pomiędzy realokacjami (np. współczynnik klikalności CTR a liczba sesji). Istnieje także ryzyko tendencyjności w przypadku przesunięcia użytkowników doświadczających gorszych wariantów i nierównego przypisania ich do wygrywających opcji.

W grudniu 2018 r. troje autorów tej książki zorganizowało wydarzenie pod nazwą First Practical Online Controlled Experiments Summit (dosł. pierwszy praktyczny szczyt ds. kontrolowanych eksperymentów online). Trzyście organizacji, a wśród nich Airbnb, Amazon, Booking.com, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yandex, i Uniwersytet Stanforda, oddelegowało łącznie 34 ekspertów, którzy zaprezentowali wnioski i wyzwania płynące z sesji (Gupta i inni, 2019). Czytelnicy zainteresowani wyzwaniami z pewnością skorzystają na zapoznaniu się z tą pracą.