INFORMACIJE I PRIKAZI

Online eksperimenti – iskustva velikih kompanija[1]

Tvrtko Tadić[2]

Velike kompanije koriste **podatke** kako bi napravile što bolje proizvode za svoje korisnike i ostvarile svoje poslovne ciljeve. No u široj javnosti, osim kada izbije neki skandal, nije jasno kako se to zaista ostvaruje. Dobar dio prakse pokriven je velom poslovne tajne ili često ne izlazi izvan stručnih krugova. Čak i ljudi poput mene, koji rade u tim kompanijama, zbog toga što rade samo na dijelu proizvoda nemaju puni uvid u sve tehnike koje se koriste. **Online eksperiment** jedna je od najvažnijih tehnika koju primjenjuju tehnološki divovi, pa bi bio bi red da (stručna) javnost dozna o čemu je zapravo riječ.

Slika 1. Američko i japansko izdanje knjige

O knjizi

Godine 2020. izašla je knjiga *Thrustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*, o tome kako velike tehnološke kompanije koriste tehniku *online eksperimentiranja*. Na knjigu sam slučajno naišao posjetom web-stranice

[1]Predavanje pod istim naslovom i u nešto širem opsegu održano je u organizaciji Inženjerske sekcije HMD-a 1. 12. 2021. Snimka predavanja dostupna je na stranicama društva.
[2]Tvrtko Tadić, Microsoft Corporation, Redmond, SAD

https://exp-platform.com/ koju održava Microsoftova organizacija za podršku eksperimentiranju – ExP tim. Knjigu su napisali vodeći ljudi Microsofta, Googlea i LinkedIna zaduženi za online eksperimente. Kako je cijena knjige bila 35 dolara, odlučio sam je kupiti i vidjeti što u njoj piše. Knjiga me se iznimno dojmila i proširila mi vidike pa sam o njenu sadržaju izlagao čak četiri puta. Napisana je jednostavnim stilom, većinu poglavlja mogu razumjeti svi, sadrži napredna poglavlja za stručnjake i otvorene probleme na kraju. U knjizi su izloženi vrlo konkretni primjeri iz poslovne prakse. Izdavač knjige je Cambridge University Press, a u međuvremenu je objavljena na kineskom i japanskom jeziku. Knjiga je podijeljena u 5 dijelova:

1. Uvodne teme za svakoga
2. Odabrane teme za svakoga
3. Dopunjujuće i alternativne tehnike kontroliranim eksperimentima
4. Napredne teme za izgradnju platforme za eksperimente
5. Napredne teme za analizu eksperimenata

Cjeline 1 i 2 daju motivacijske primjere i objašnjavaju problematiku online eksperimentiranja. Cjelina 3 govori o izazovima kada nemamo mogućnost izvođenja online eksperimenata te o drugim mogućnostima. Cjelina 4 je tehnička i govori o tome kako se u praksi izvode eksperimenti kada se izvode često. Cjelina 5 govori o matematici iz eksperimenata i raznim statističkim metodama koje se koriste.

O autorima

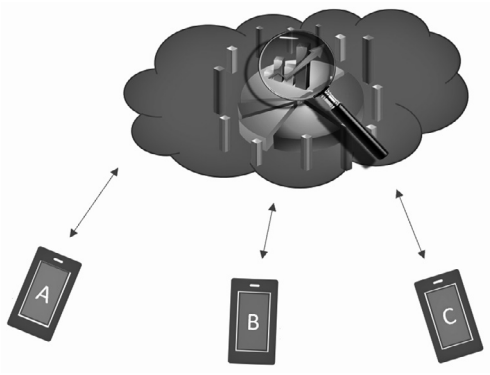
Kao što je već spomenuto, autori su veterani tehnike online eksperimentiranja koji **zajedno imaju više od 30 godina iskustva** u ovom području. Ovo su podatci o autorima u trenutku pisanja knjige:

- **Ron Kohavi** korporativni je potpredsjednik u Microsoftu gdje vodi ExP organizaciju koja se bavi razvojem platforme za provođenje eksperimenata. Prije toga je bio direktor rudarenja podataka i personalizacije u Amazonu. Stekao je doktorat iz računarstva na sveučilištu Stanford.
- **Diane Tang** je *Fellow* u Googleu gdje je radila na razvoju tehnika i infrastrukture za analizu velikih količina podataka. Stekla je doktorat iz računarstva na sveučilištu Stanford.
- **Ya Xu** voditeljica je tima za podatkovne znanosti i eksperimentiranje u LinkedInu. Prije toga je radila u Microsoftu i stekla doktorat iz statistike na sveučilištu Stanford.

Autori se poznaju s brojnih konferencija iz područja online eksperimentiranja na kojima izlažu stručnjaci iz brojnih kompanija koje koriste ovu tehniku za razvoj proizvoda.

Online eksperiment

U današnjem svijetu većina je softwarea spojena na Internet. Dio programa izvršava se na samom računalu (mobitelu, tabletu ili pravom računalu), dok se dio radi u *oblaku*. Ovakav pristup omogućava **osvježivanje i izmjenu sadržaja** u programu prema potrebi, kao i prikupljanje podataka o načinu na koji se korisnici služe programom.



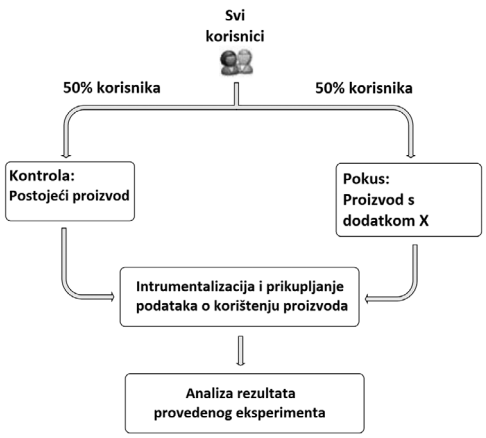
Slika 2. Ista aplikacija s tri različite verzije A, B i C

Ovo je velika razlika u odnosu na prvobitan razvoj softwarea do ranih 2000-ih. Zbog slabe ili nikakve internetske veze bilo je nemoguće napraviti veće izmjene programa i prikupiti potrebne podatke. Ovo je onemogućavalo bitna unapređenja softwarea i usklađivanje s novim spoznajama i znanstvenim otkrićima.

Ono što novi pristup omogućava je da možemo raznim korisnicima pokazati različite verzije programa. Slika 2. ilustrira jedan takav primjer gdje tri mobitela imaju tri različite verzije iste aplikacije – A, B i C. Na ovaj način moguće je provesti **eksperiment** na većem broju korisnika i utvrditi koja verzija bolje odgovara potrebama korisnika i poslovnim ciljevima izdavača programa.

Ovakva istraživanja obično se svode na A/B testove koji se često koriste u biomedicinskim i raznim drugim istraživanjima. Uobičajeno se na postojeći proizvod dodaje novi dodatak (funkcionalnost) X. Na skupu korisnika onda se provede uobičajeni A/B test:

- **50 %** korisnika u **kontrolnoj grupi** bit će izloženo postojećem proizvodu.
- **50 %** korisnika u **pokusnoj grupi** bit će izloženo proizvodu s dodatkom X.
- Prikupit će se razna mjerenja i na kraju analizirati i usporediti rezultati ovih dviju grupa korisnika.



Slika 3. Shema A/B testa

Neka od pitanja na koja se želi odgovoriti su:

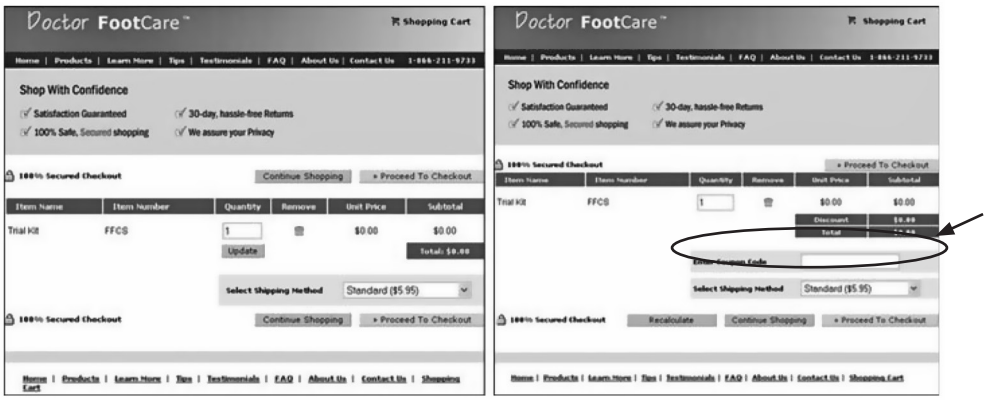
- Koristi li se dodatak X uspješno?
- Je li došlo do poremećaja u korištenju proizvoda nakon dodavanja dodatka X? Koristi li se više ili manje proizvod sa ili bez dodatka X?

Dva primjera

Dano je više primjera iz prakse korištenja online eksperimenata. Ovdje ćemo navesti dva.

Kuponi za popust u web-trgovini

Prvi nije iz knjige nego iz članka dostupnog na prije spomenutoj web-stranici. Slični primjeri dostupni su u knjizi.



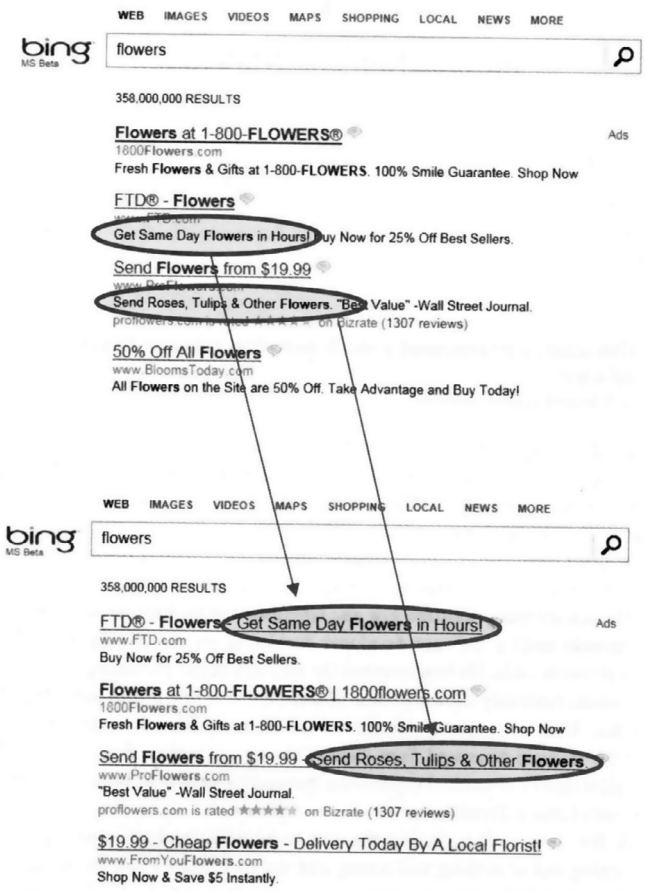
Slika 4. Primjer eksperimenta u kojem se testiralo dodavanje mogućnosti unosa kupona za popuste

Web-trgovina razmatrala je uvođenje kupona za popust. Poznato je da to može imati negativne utjecaje. Kupac će početi razmišljati je li cijena koju plaća prevelika jer nema kupon, i može dovesti do toga da određen broj kupaca odustane od namjere kupnje. Kako bi se *procijenili* efekti dodavanja kućice u koju se može unijeti kôd kupona kod plaćanja, obično se ona uvede i prije nego kuponi zaista postoje. Na slici 4. prikazan je eksperiment u kojem je na lijevoj strani kontrolna grupa s postojećim rješenjem, a na desnoj pokusna grupa s dodanim rješenjem za unos kupona.

Ovaj eksperiment proveden je na manjem dijelu korisnika. U pokusnoj skupini zabilježen je **promet manji za 90 %**. Logično, odustalo se od ovoga rješenja.

Online oglasi

Idući je primjer iz povijesti razvoja Microsoftove tražilice Bing. Oglasi na Bingu pokušavali su razviti efikasan način informiranja korisnika o svom sadržaju. Postojao je prijedlog da se uz naslov oglasa doda i prva rečenica iz oglasa. Predložena promjena prikazana je na slici 5.



Slika 5. Eksperiment s Bing oglasima – u poksunoj grupi naslov sadrži prvu rečenicu iz oglasa

Ovu ideju nitko nije smatrao ozbiljnom i nije ju se namjeravalo provesti u praksi. Tijekom *hackathon tjedna*[3] jedan je zaposlenik odlučio eksperimentirati s ovim rješenjem. Rezultati eksperimenta bili su nevjerojatni. Korisnici su počeli mnogo češće *klikati* na oglase, tako da je promet od prodaje oglasa **povećan za čak 10 %**. Godine 2012., kad je eksperiment napravljen, to je povećanje iznosilo **100 milijuna dolara**.

Niska uspješnost eksperimenata

Često se stječe dojam da velike kompanije nižu uspjehe. Međutim, riječ je o tome da mali broj eksperimenata bude uspješan, dok većina eksperimenata, unatoč velikoj vjeri u njih, ne pokaže nikakvu vrijednost. Primjer razine uspješnosti s Bing oglasima dogodi se jednom u nekoliko godina, dok većina uspješnih eksperimenata ima mnogo skromnije rezultate. Knjiga donosi više citata iz industrije s raznih konferencija, a ovdje ćemo izdvojiti nekoliko:

[3]Hackathon tjedan u IT kompanijama vrijeme je kada zaposlenici mogu slobodno raditi što žele. Mnogi iskoriste mogućnost da isprobaju nešto novo.

Zapanjujuća činjenica o intuiciji je da se u velikom broju slučajeva pokaže potpuno pogrešnom.

John Quarto-von Tivadar, Future Now

Ako ste organizacija koja koristi eksperimente, pripremite se da, u najboljem slučaju, 70 % onog što radite ne uspije.

Fareed Mosavat, Slack

Netflix smatra da 90 % onoga na čemu radimo - neće uspjeti.

Mike Moran, Netflix

U knjizi se govori o važnosti **kulture eksperimentiranja** kao ključne za razumijevanje potreba korisnika. Trebamo biti svjesni da **put do konačnog uspjeha često vodi preko niza neuspjeha**.

Metrike i matematika iza eksperimentiranja

Kako bismo mjerili uspješnost eksperimenta i bili sigurni da on nije poremetio korištenje proizvoda, definiraju se statistike koje to mjere, a nazivaju se **metrike**. Metrika može biti i nekoliko stotina, a trebaju osigurati da promjena ostvari cilj. U tablici je dana česta industrijska klasifikacija metrika:

Naziv	Opis	Primjeri
Metrike cilja	Manji broj metrika koji najbolje opisuje poslovne ciljeve.	Prihod/profit, udio tržišta
Pokretačke metrike	Metrike koje pokazuju kretanje u pravom smjeru, a koje se neće nužno odmah odraziti na metrike cilja.	Mjere poboljšanja kvalitete, broj korisnika
Zaštitne metrike	Metrike kojima vrijednosti moraju zadovoljavati određene uvjete jer, u suprotnom, prijeti gubitak povjerenja u proizvod.	Dostupnost usluge, brzina dostave usluge

Testovi i tablica promjena

U online eksperimentu izračunaju se sve metrike za kontrolnu i pokusnu grupu. Zatim se izvrši usporedba za svaku pojedinačnu metriku i gleda se je li došlo do **značajne promjene**. Kako bismo znali što je *značajno*, trebamo pomoć statističkih testova. U zadnjem dijelu knjige obrađuje se više slučajeva jer metrike mogu biti kompleksne, a ovdje ćemo prikazati jednostavan slučaj t-testa s kojim su se mnogi mogli susresti u uvodnim kolegijima iz statistike.

Izmjerit ćemo metriku u kontrolnoj grupi

$$Y_1^A, Y_2^A, \ldots, Y_{n_A}^A,$$

i pokusnoj grupi

$$Y_1^B, Y_2^B, \ldots, Y_{n_B}^B.$$

U brojnim slučajevima želimo vidjeti:

- je li došlo do **značajne** promjene u *prosječnoj* vrijednosti metrike,
- ako je došlo do značajne promjene, je li ta promjena *poželjna*, tj. u skladu s našim ciljevima.

Kako bismo to izmjerili, računamo aritmetičke sredine obaju skupova podatka

$$\overline{Y^A} = \frac{Y_1^A + \ldots + Y_{n_A}^A}{n_A} \text{ i } \overline{Y^B} = \frac{Y_1^B + \ldots + Y_{n_B}^B}{n_B}$$

te gledamo razliku

$$\Delta = \overline{Y^B} - \overline{Y^A}.$$

Pokazuje li vrijednost Δ da je došlo do značajne promjene? Za to moramo još izračunati vrijednost varijance

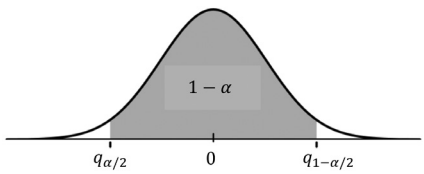
$$\widehat{Var}(\Delta) = \frac{1}{n_A - 1}\sum_{i=1}^{n_A}\left(Y_i^A - \overline{Y^A}\right)^2 + \frac{1}{n_B - 1}\sum_{i=1}^{n_B}\left(Y_i^B - \overline{Y^B}\right)^2.$$

Ako su vrijednosti izmjerenih metrika nezavisne, nije došlo ni do kakve promjene. Za veliki broj mjerenja imat ćemo

$$T = \frac{\Delta}{\sqrt{\widehat{Var}(\Delta)}} \sim N(0,1).$$

Ovo tumačimo na sljedeći način: ako su kontrolna i pokusna grupa izložene istom proizvodu, kada ponovimo pokus n puta, vrijednosti T_1, T_2..., T_n *će* (približno) biti realizacija normalne slučajne varijable s varijancom 1 i očekivanjem 0.

Sada primjenjujemo standardni način statističkog testiranja. Ako je $T \in [q_{\alpha/2}, q_{1-\alpha/2}]$, gdje je q_x x-ti kvantil normalne razdiobe, zaključujemo da nije došlo do značajne promjene na razini značajnosti α. U suprotnom imamo značajnu promjenu.



Slika 6. Normalna razdioba s testnim intervalom $[q_{\alpha/2}, q_{1-\alpha/2}]$.

Vertical Feature Metrics	Treatment	Control	Delta [%]	Pval
	0.9992	0.9992	–	0.8670
	0.0154	0.0117	0.0036 [30.69%]	< 0.001
	0.9334	0.9344	–	0.1222
	0.0265	0.0300	-0.0035 [-11.60%]	< 0.001
	0.2589	0.2597	–	0.5045
	0.0500	0.0619	-0.0119 [-19.29%]	< 0.001
	1.000	1.000	–	0.1573
	0.6635	0.6674	-0.0040 [-0.59%]	0.0481
	0.7680	0.7688	–	0.2522
	0.0073	0.0074	–	0.2575
	0.5230	0.5281	-0.0051 [-0.96%]	< 0.001

Slika 7. Izvještajna tablica jednog eksperimenta u Microsoftu. Imena metrika su prekrivena.

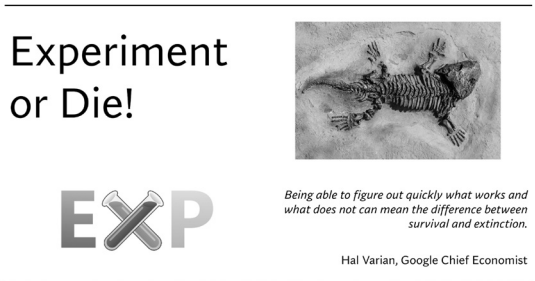
Ovo nam omogućuje da *automatski* pripremimo **izvještaj za sve metrike** koji će nam reći imamo li značajnu promjenu ili ne. Vidi sliku 7. Vrijednost Δ se ne prikazuje ako promjena nije bila značajna, a boja vrijednosti pokazuje smjer promjene. Na temelju ovakvih tablica donosi se odluka treba li novu verziju proizvoda napraviti dostupnom svim korisnicima ili od nje odustati. Za vrijednost α obično se uzima 5 %.

Završni komentari

U ovom pregledu vidjeli smo neke primjere eksperimenata i jedan jednostavni način kako se vrednuje je li došlo do značajne promjene ili ne. Kako je knjiga nastala kroz praksu, u njoj se razmatraju bitno kompliciraniji slučajevi. Spomenimo neke teme i izazove koji se obrađuju:

- Izmjerene vrijednosti metrika u kontrolnoj i pokusnoj grupi ne moraju biti neovisne. Primjerice, u društvenim mrežama, zbog brojnih veza, može doći do efekta prelijevanja.
- Velike kompanije na godišnjoj razini izvode desetke tisuća eksperimenata na istom skupu korisnika. Platforma za eksperimentiranje mora biti izgrađena kako bi podržala istovremeno eksperimentiranje i uzimanje uzorka da bi rezultati eksperimenata bili pouzdani.
- Eksperimenti mogu štetiti korisniku. Ovo otvara brojna etička pitanja.
- Nekad nismo u mogućnosti izvesti online eksperiment. Tada ovisimo o opa*žen*im podatcima koji mogu biti bitno manje pouzdani.
- Brojne metrike podložne su izigravanju. Sustav metrika mora biti takav da onemogući takvo ponašanje.
- Svaki test na razini značajnosti α može krivo pokazati značajnost promjene metrike s vjerojatnošću α. Ako imamo n neovisnih metrika i obje skupine izložene su istom proizvodu, vjerojatnost da bar jedna od njih lažno pokaže značajnu promjenu je $1 - (1 - \alpha)^n$. Za veće n to može dovesti do toga da praktično nijedan eksperiment neće biti moguće provesti, zato metrike u eksperimentu treba grupirati i zadati različite značajnosti.

Online eksperimenti predstavljaju zlatni standard za donošenje odluka i ključni su dio proizvodnog procesa u IT industriji. Ova knjiga zbog ogromnog iskustva koje autori imaju predstavlja neizostavan praktični udžbenik iz ovog područja.



Slika 8. Plakat o važnosti eskperimenata koji je u Microsoftovim zgradama postavila organizacija za podršku eksperimentiranju