MSc Applied Information and Data Science

HSLU Lucerne University of Applied Sciences and Arts

Data Analytics for Data Scientists

Design of Experiments (DoE)

Lecture 08: A/B Testing

2025

Prof. Dr. Jürg Schwarz

Program: 16:15 until 17:55

16:15	**Start of the lesson**
	Lecture: Jürg Schwarz ◦ Example of A/B testing ◦ What is A/B testing? ◦ Carrying out A/B testing ◦ Error sources and pitfalls ◦ Other aspects ◦ Preview of Lecture 09
	Tutorial: Students / Jürg Schwarz / Assistants ◦ Working on the exercise ◦ Support by Jürg Schwarz / Assistants
17:55	**End of the lesson**

Detailed table of contents → Slide 23

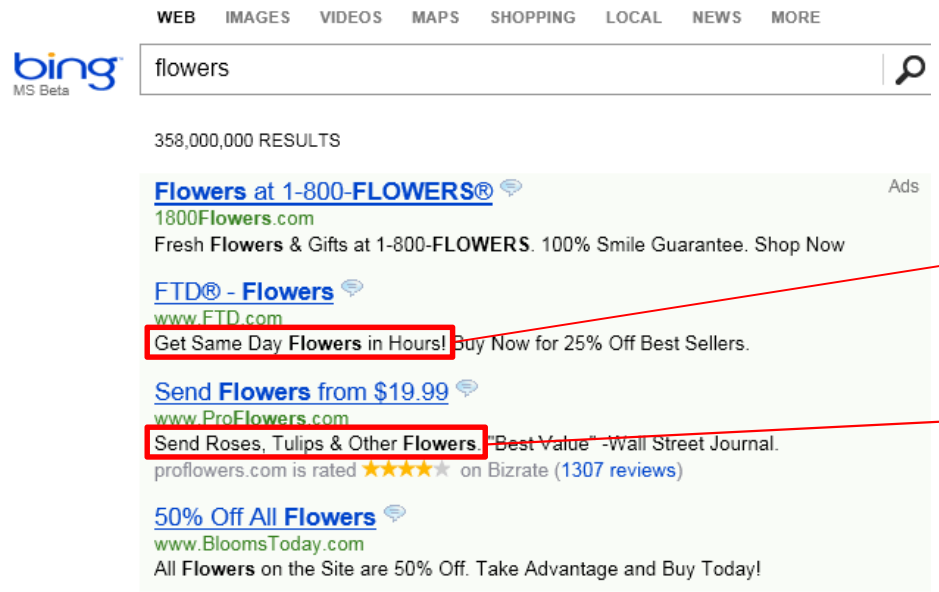
Example of A/B testing – A classic

Bing AdWords

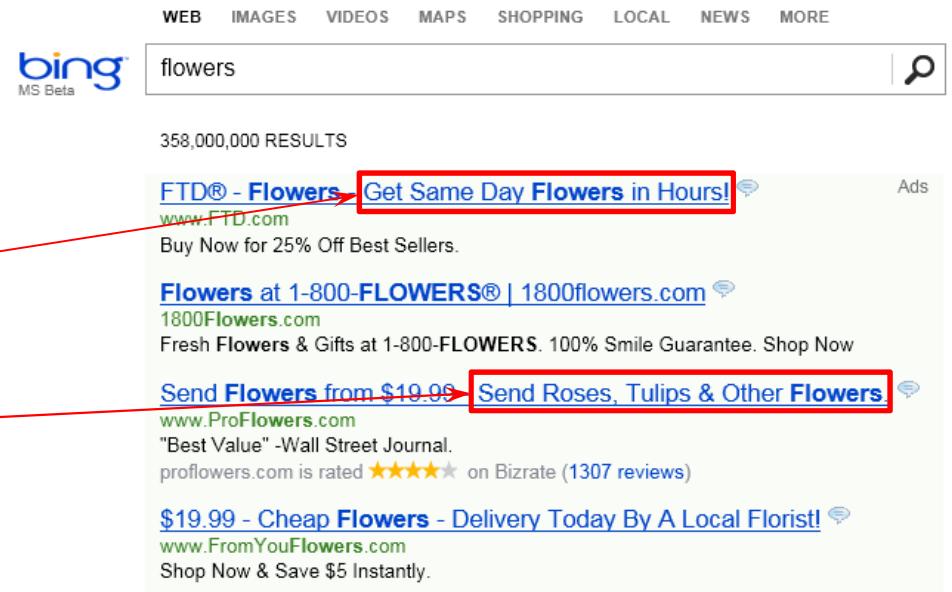
Experiment with advertisements I

Advertisements are moved to the headings

Current view



New view with "Long Ad Titles"



Experiment with advertisements II

The idea of "Long Ad Titles" came from a Microsoft employee in 2012.

It was ignored initially.

Later, the idea was implemented with little effort and an experiment was started

→ **A/B testing** (control = current view / treatment = new view)

The new view led to a 12% increase in Bing's sales in one year (USD 120 million)

There were no changes in the key figures that capture users' experience.

Two famous examples from the beginnings – among many ...

Barack Obama's election campaign in 2008 → optimization of online donation forms

3 labels for the donation button (*SIGN UP NOW*, *JOIN US NOW*, etc.)

3 pictures / 3 videos

Amazon's Website

- Boot Camp in January 1997 → statement by Amazon CEO Jeff Bezos
  "At Amazon, we will have a Culture of Metrics"

  A/B testing was / is used for: New home page design, moving features around the page, different algorithms for recommendations, changing search relevance rankings, …

Kohavi & Thomke (2017) / Marcus (2005)

What is A/B testing?

Context I – Design of Experiments / Statistics

A/B testing (also *bucket testing* or *split-run testing*) is an experiment often used in labs.

The research questions are applied to two (A/B) or more randomized groups.

The statistical analysis is done by t-test, ANOVA and more advanced methods.

Context II – From laboratory study to web experiment

The digital transformation from 2000 onwards has shifted the context of communication and behavior towards websites, e-commerce, and (social) communities with their own reality.

Natural human behavior can be investigated directly and experimentally on the web without the subjects having to be informed about the experiment.

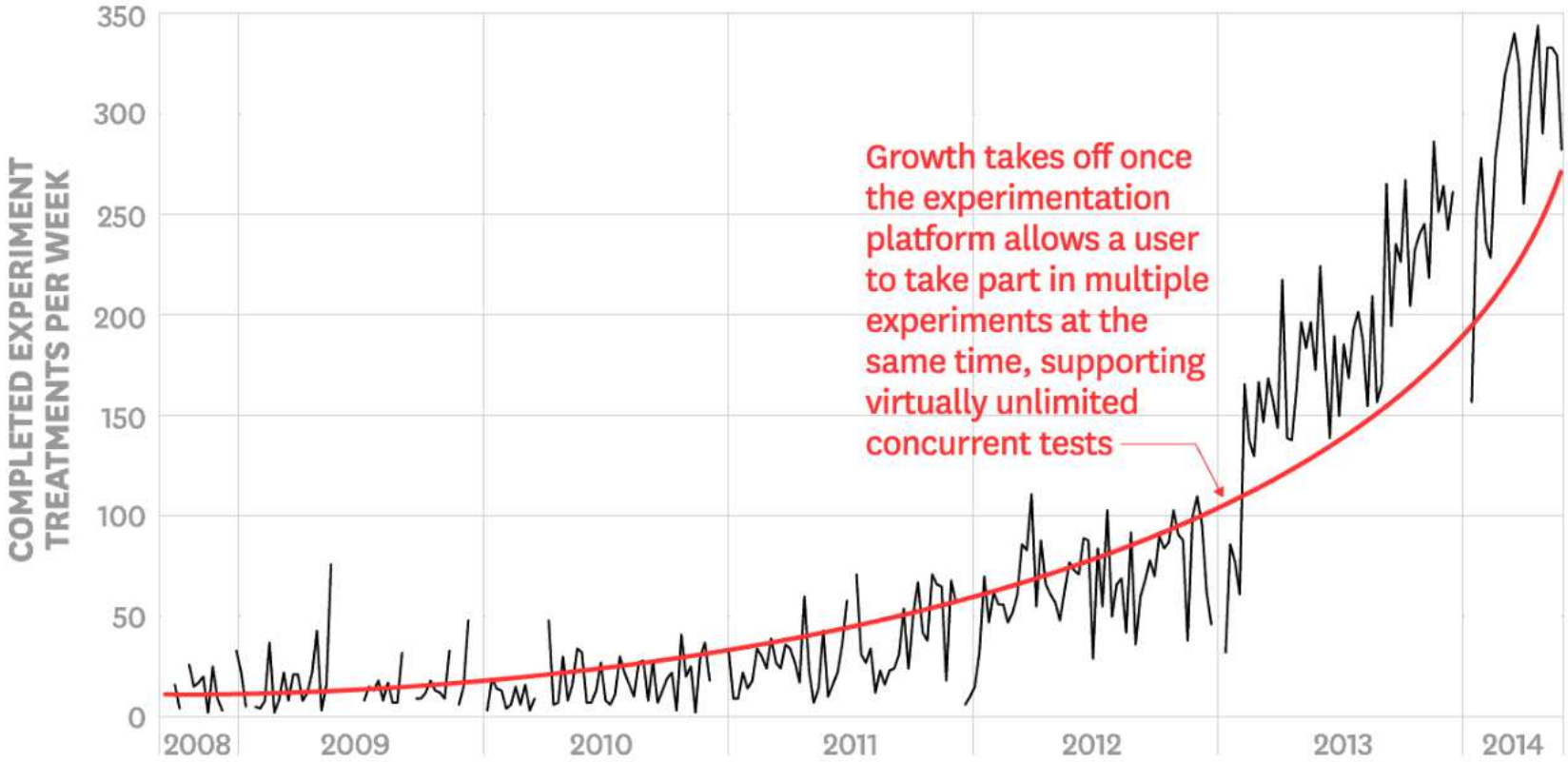
Tools – controlled by scripts and cookies – make it possible to serve different versions of a website to different visitors in parallel, and to measure key indicators such as purchases, registrations, downloads, etc.

The way that the experiments are designed allows for decisions for or against introducing certain design variants in web systems and emails.

Context III – Development

All the main players (Amazon, Bing, Facebook, Google, LinkedIn, etc.) conduct thousands to tens of thousands of experiments each year to test changes in user interfaces, improvements in algorithms (search, display, personalization, recommendation, etc.), changes of apps, content management system, etc.

Example*: Experiments with advertisements in Bing per week (see → Slide 3ff)



In 2022**, Microsoft conducted more than 20,000 controlled experiments per year on Bing.

*Kohavi et al. (2020) / *Kohavi & Thomke (2017) / **Kohavi et al. (2022)

Carrying out A/B testing

Principles

A/B testing is a method to compare two versions of a website, an app or ...
The aim is to determine which version achieves a better result, in terms of click-through rate etc.

Two versions A and B are tested in parallel in a live environment.
The generated data becomes the basis for decisions. Versions are not changed during the test.

Which elements of websites and apps can be tested?

All of them!

Extensions

It is also possible to test more than two versions (multivariate tests).
There are also dynamic-algorithmic procedures → see from Slide 9

Challenge

Achieve enough traffic / conversion

Costs

The costs are likely to be low compared to lab experiments / surveys.
There is a large market for tools and services.

Selection of user groups

Population / Triggering

Depending on the experiment: All users of a website or only a subset.
Examples: New web design → all / Checkout process → subset with purchase intention

Reduction of nuisance variance through event-triggered filtering
→ Include only those users who were exposed to the treatment for a certain time / intensity.

Sampling / Sample size

Ideal case → Lecture 05: Sampling & Lecture 06: Effect size & Power analysis

Calculation of the sample size with an application

- Choose a suitable metric – e.g. click-through rate (CTR)*
- Make an educated guess for the CTR → existing, successful site → e.g. 3%
- A/B testing calculator → e.g. clevertap.com/ab-testing-calculator

*CTR = Number of clicks on advertising banners or sponsor links in relation to the total impressions / Normal case: 3% to 15%

Advanced methods: Bandit algorithm

Introduction

Besides answering the classic research question in A/B testing
"Do variants A and B differ significantly?"
a test can also be used to dynamically search for optimal solutions.

Bandit algorithms make it possible to test several treatments at once and dynamically drawing faster conclusions than conventional study designs.

A treatment is considered as the "arm" of a slot machine ("one-armed bandit").
Several treatments correspond to several arms with various payment probabilities.

Description

Several variants (treatments A, B, C, ...) are offered at the same time.

The variant with the highest "success" (metrics ...) gets the most data traffic.
Other variants are refined and tested with less traffic.

Adjustments are made based on the actual success.
As the test progresses, more and more information about the success becomes known and thus the most successful variants can be determined dynamically.

A key advantage of the algorithm is that it achieves a higher overall profit while allowing the algorithm to collect data on other variants.

Examples: Comparison of Classic A/B test vs. Multi-armed bandit algorithm

Classic A/B test

Control group (**A**) of a current website has a 4% conversion rate.
A new version of the website (**B**) is tested and probably achieves a 5% conversion rate.

A test should be performed that can detect a significant difference.
According to a power analysis, 22,330 observations are required → 11,165 per variant.

If 100 visits per day can be tested, the test will take 223 days.

Multi-armed bandit algorithm

Day 1

- 100 visits are distributed proportionally: ~ 50 (e.g. ~ 50%) visits per version (A & B)
- Success probabilities for both versions are calculated

Day 2

- Traffic is distributed proportional to the probabilities (e.g. 70% to version B if B is better)
- Total results from day 1 & 2 are cumulated

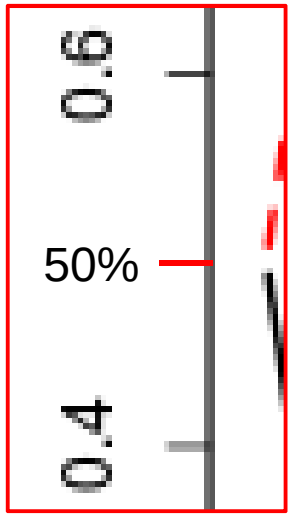
Following days

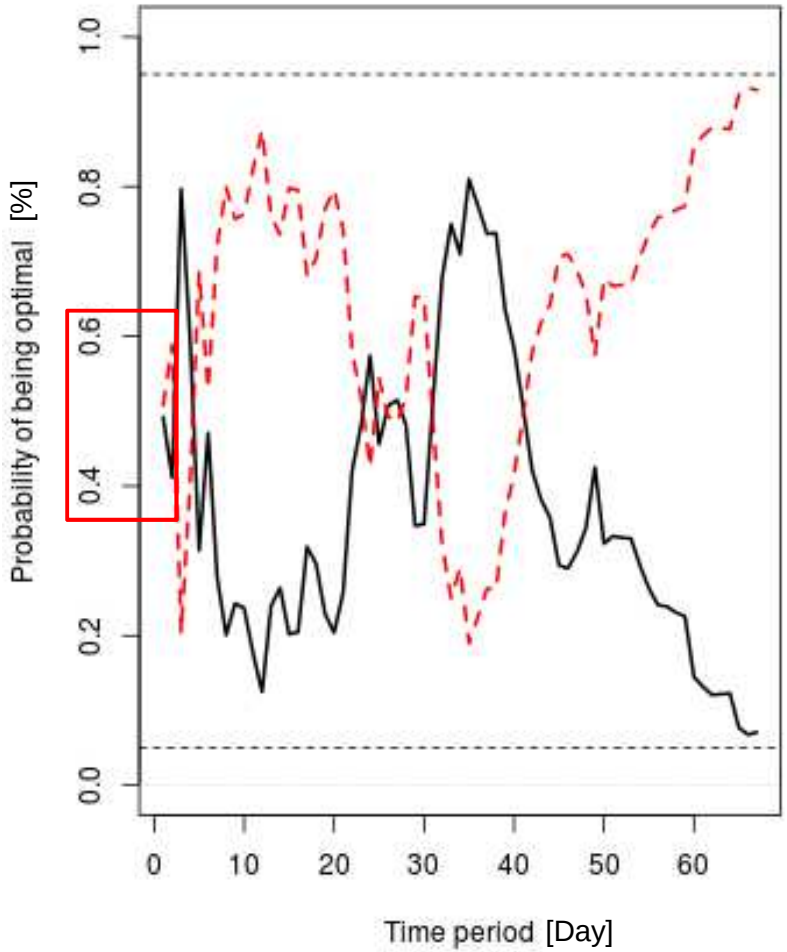
- Traffic allocation is continuously updated based on new results
- Success probabilities are recalculated daily

End of experiment

- A stop rule (e.g. ≥ 95% probability that one version is best) is applied to declare a winner

By default, a multi-armed bandit experiment runs for at least two weeks.
For example, a stop rule can be defined by taking into account the probability that a variant will generate the most traffic in the future. If this probability is ≥ 95%, there is a "winner".

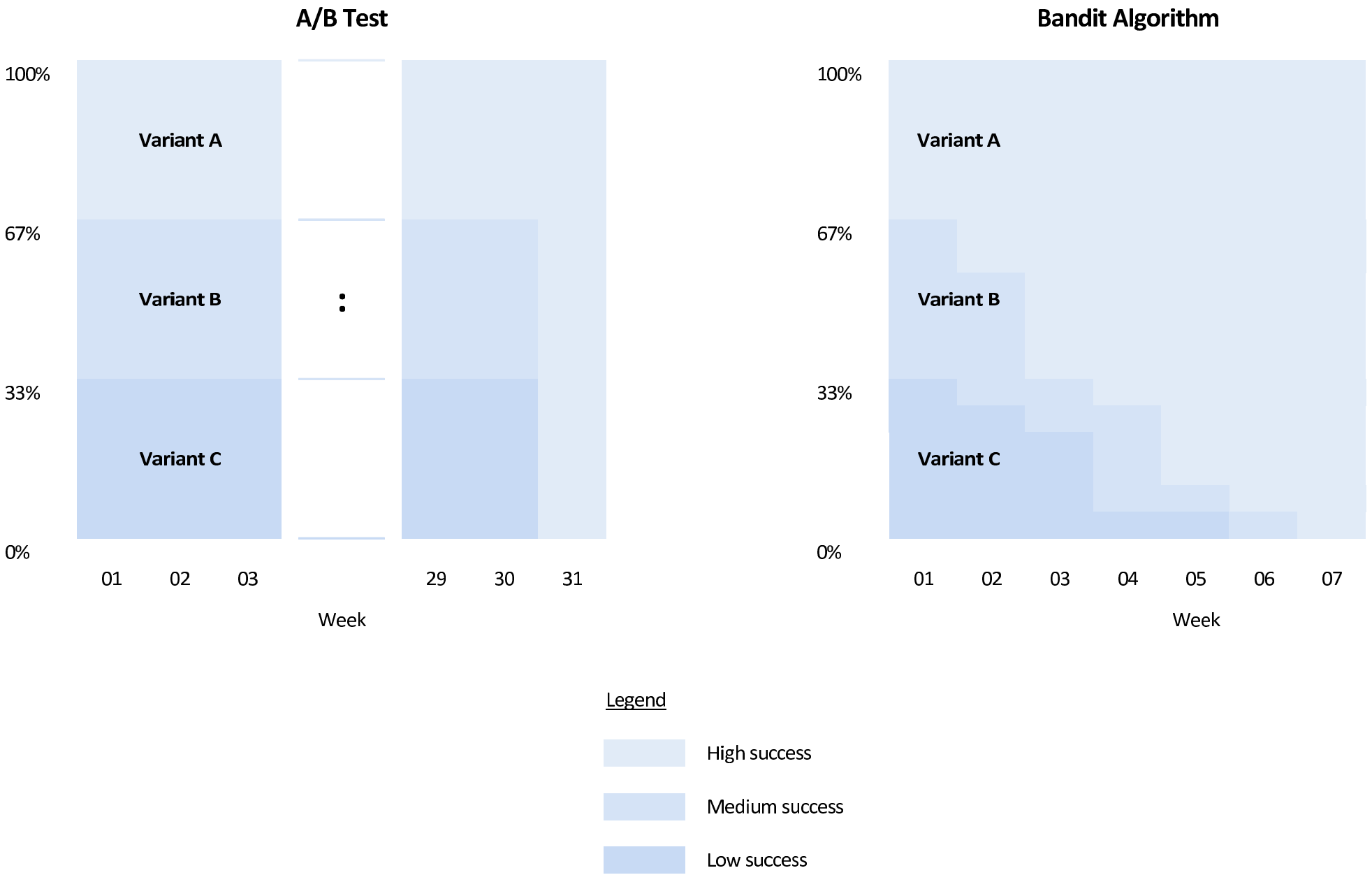




Graph: Simulation of a simple multi-armed bandit experiment with two variants.
The "Probability of being optimal" reflects how the traffic is allocated to the two variants.

analytics.googleblog.com/2013/01/multi-armed-bandit-experiments.html

Comparison


A/B Test
100%
67%
33%
0%
Variant A
Variant B
Variant C
01
02
03
29
30
31
Week
Bandit Algorithm
100%
67%
33%
0%
Variant A
Variant B
Variant C
01
02
03
04
05
06
07
Week
Legend
High success
Medium success
Low success

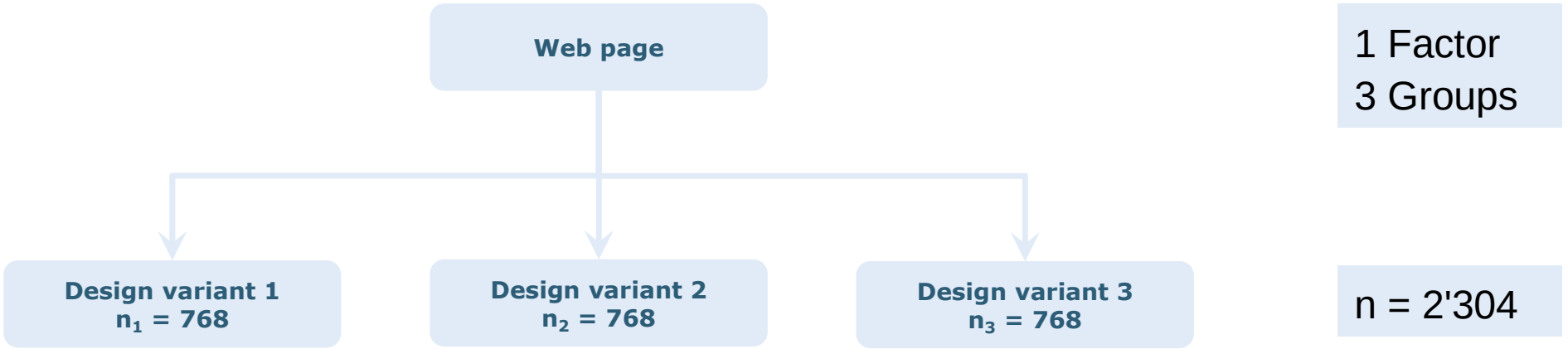
An example of A/B testing

Average time spent on a web page (*dwell time*)

How does the factor IV1 affect the dwell time DV, measured in seconds?

DV = dwell time [s] → output with metric scaling

IV1 = design variants [1,2,3] → factor with 3 levels (↔ 3 groups)



$\overline{A}_i$ = mean of the dwell time in group i

	IV1		
	1	2	3
	$\overline{A}_1$	$\overline{A}_2$	$\overline{A}_3$

	IV1		
	1	2	3
	36.2	38.8	42.1

ANOVA with data set *dwelltime* and R file *average_time*

```
library(readxl)
dwelltime <- read_excel("dwelltime.xlsx")

fit  <-  aov(DV  ~  factor(IV1),  data  =  dwelltime)
summary(fit)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
factor(IV1)    2  13256    6628   77.93 <2e-16 ***
Residuals   2301 195708      85
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is a main effect of IV1 (levels 1, 2, 3) on DV, $F(2, 2301) = 77.93$, $p = .000$.

Design variants (1, 2, 3) have a significant effect on dwell time.

The mean of the dwell time in design variant 1 is 36.2 seconds,
38.8 seconds in variant 2, and 42.1 seconds in variant 3.

All mean values differ in pairs (Post-hoc Bonferroni).

What is the effect size? → run ANOVA with "partial eta squared"

```
library(effectsize)
eta_squared(fit)$Eta2
```

```
...
[1] 0.06343711
```

Partial eta squared η_p^2 relates the variance *explained by one factor* to the variance *not explained by other factors* in the model.

Effect size *f* for the *one-way analysis of variance* according to Cohen (1992), calculated from η_p^2

$$f = \sqrt{\frac{\eta_p^2}{1-\eta_p^2}} = \sqrt{\frac{0.063}{1-0.063}} = \sqrt{\frac{0.063}{0.937}} = 0.26$$

	Small	Medium	Large
Effect size *f*	0.10	0.25	0.40

Design variants (1, 2, 3) have a significant effect on dwell time.
Effect size is *f* = 0.26 (Cohen 1992)

The effect size is medium.

Table from the article *A Power Primer* by Cohen (1992)

Test	ES index	Effect size		
		Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. $P = .5$ and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

In our case the equation is different because R does not provide σ_m and σ

Measurement / Metrics / KPI (Key performance indicator)

Measurement	Metrics	KPI (Key performance indicator)
Method for obtaining one or more measured values that can be assigned to a quantity.	Calculation from measured values	Quantifiable metric that shows how effectively the most important company goals are achieved.
Number of clicks on a web site	Click through rate	Conversion rate

Description / Examples

Click-through rate (CTR) is the ratio of users who click on a specific link to the number of total users who view a page, email, or advertisement. It is commonly used to measure the success of an online advertising campaign for a particular website.

Conversion rate is defined as the percentage of visitors who complete a goal, as set by the site owner. It is calculated as the total number of conversions, divided by the total number people who visited your website:

$$\text{Conversion rate} = \frac{\text{Number of visitors who achieved the goal}}{\text{All visitors}}$$

A website receives 100 visitors in a day and 15 visitors sign up for the email newsletter (the chosen conversion to measure). The conversion rate would be 15% for that day.

Error sources and pitfalls

Research design / Suitability of A/B testing

A/B testing cannot be used in all research questions → Example: For a complete website redesign, further elements have to be included (qualitative aspects, survey, UX testing ...)

Sampling → Lecture 06: Sampling

Population is unknown / sampling procedure is not suitable / sampling bias
Target audience is unsuitable → Example: Conducting an experiment on *page layout* that includes only heavy users of the page.

Sample size / Hypothesis testing → Lecture 07: Effect size & Power analysis

Large samples → Hypothesis test is significant, but information on effect sizes is missing.
Unsuitable statistical tests → Example: Skewed distributions are not taken into account.

Metrics

The measurement uses an unsuitable metric → Topic / examples on exp-platform.com*

Other

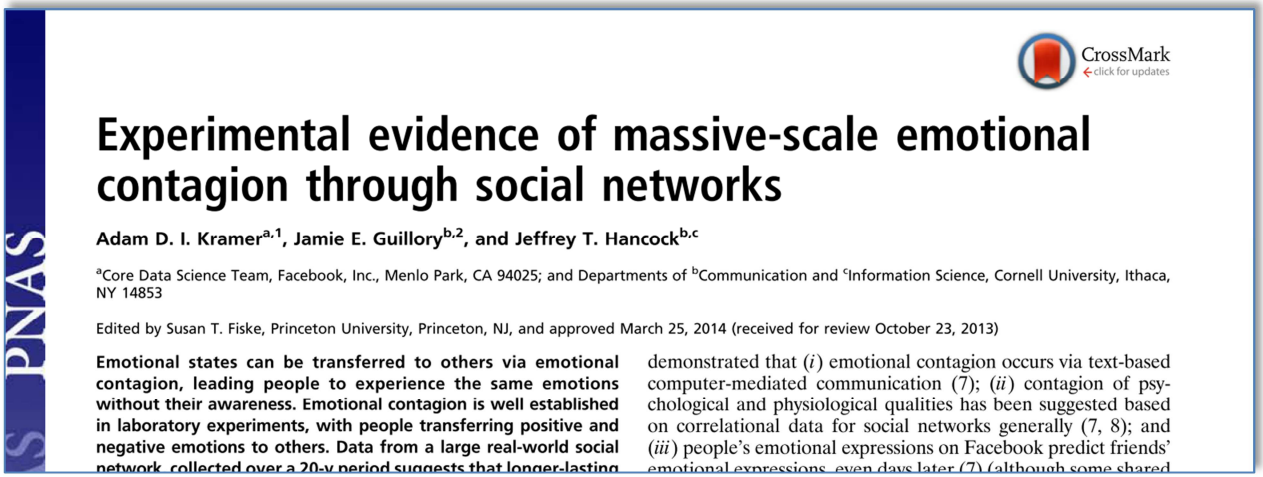
Multiple testing → Topic / examples exp-platform.com*

…

*For example: Kohavi et al. (2022) A/B Testing Intuition Busters
*For example: Larsen et al. (2023) Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology

Other aspects

Manipulation / Dark Pattern



PNAS

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared

Example of manipulation: Facebook – Article by Kramer et al. (2014)

Experimental evidence of massive-scale emotional contagion through social networks*

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness.

The experiment was conducted **without participants knowing** that they were part of it.

Later, an "Editorial Expression of Concern" was published (red colored by the MSc author)

Obtaining informed consent and allowing participants to opt out are best practices in most instances under the US Department of Health and Human Services Policy for the Protection of Human Research Subjects (the "Common Rule").

It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.

*emotional contagion → Spreading of emotions (DE: Gefühlsansteckung)

Dark Pattern / Deceptive Pattern

Dark / deceptive patterns are patterns of persuasion and influence.

In general, they may prove promising for the provider, but they can have a negative effect in terms of image, user experience, customer satisfaction, etc.

Example of hotel bookings

- Experiments with hotel bookings show that posting information about a room that one customer is interested in and then mentioning that other customers are currently also viewing it will increase the booking rate.

→ The resulting pressure has a negative effect on the user experience.

Example of software

- Microsoft wants Windows 11 users to go with Edge as their default browser.

  C'mon Microsoft. I thought your Evil Empire days were over...
  Mary Jo Foley* on www.zdnet.com in November 2021

Web page www.deceptive.design (former darkpatterns.org)

- Types of dark pattern – an example

  Privacy Zuckering → You are tricked into publicly sharing more information about yourself than you really intended to. Named after Facebook CEO Mark Zuckerberg.

*Mary Jo Foley calls herself a "Microsoft watcher"

Ethical aspects

First overview of guidelines and voluntary commitment

Europe: General Data Protection Regulation (Datenschutz-Grundverordnung DSGVO)

www.datenschutz-grundverordnung.eu (Version in English: gdpr-info.eu)

Chapter 1 General provisions

Article 1 – Subject-matter and objectives

:

Switzerland: Federal Act on Data Protection (Bundesgesetz über den Datenschutz)

www.admin.ch/opc/en/classified-compilation/19920153/index.html

Art. 1 Aim: This Act aims to protect the privacy and the fundamental rights of persons when their data is processed.

Voluntary commitment of the Verband Schweizer Markt- und Sozialforschung (vsms)

swiss-insights.ch/label-market-social-research/datenschutz

Art. 4 Principles

1 Personal data may only be processed lawfully.

2 Processing must be carried out in good faith and must be proportionate.

3 Personal data may only be processed for the purpose stated when it was obtained, …

:

Preview of Lecture 09

What has happened so far

A/B testing and more

It is worth testing versions of a web site or app to improve the result.

Methodical knowledge is the basis and further knowledge of metrics, procedures, etc. is needed.

As always in the methodological environment, sources of error and pitfalls are not far away!

Last but not least, it is useful to know the dark sides of tests and simulations.

What follows in Lecture 09

A/B testing was just the beginning!

In lecture "Factorial Experimental Design" the procedures shown in lecture "A/B Testing" are systematized and extended.

The methods are introduced into the context of experimental design.

Table of contents