Online Experimentation and A/B Testing

Data Science Dojo

datasciencedojo
unleash the data scientist in you

Agenda

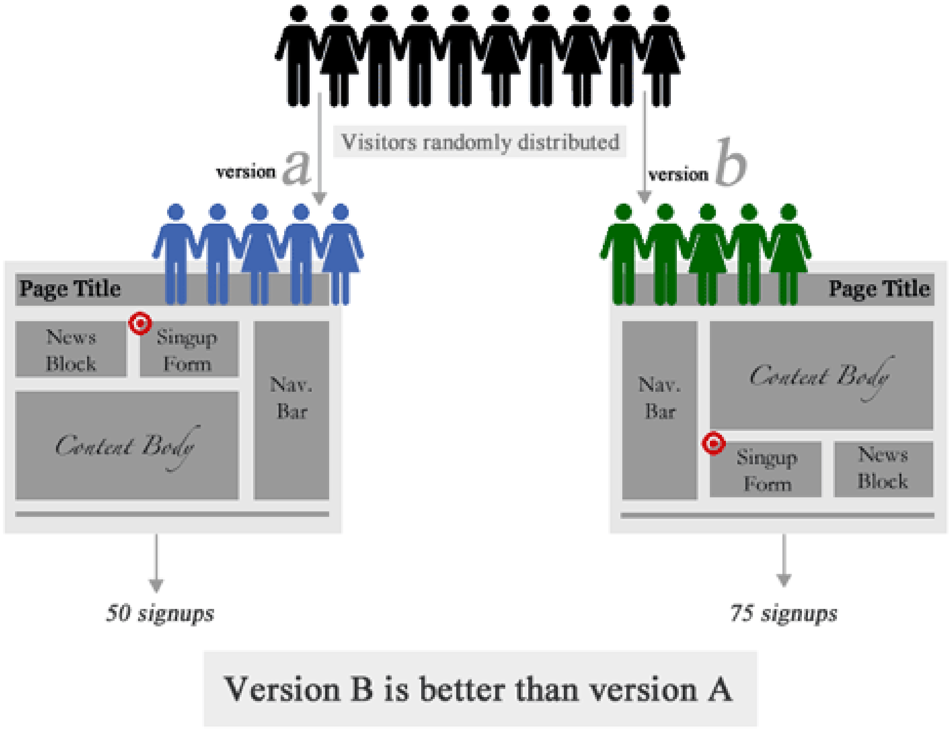
- **Introduction**
  - What is A/B testing?
  - Some interesting A/B tests
- **Fundamentals**
  - Steps in Experimentation
  - Hypothesis testing and related ideas
  - Metrics for A/B testing
  - Focus on intuitive understanding than specific distributions, formulas and tests
- **Common pitfalls**
  - Depth of discussion will depend upon audience engagement and time

Introduction

In God we trust. All others bring data.

W. E. Deming

What is A/B Testing?

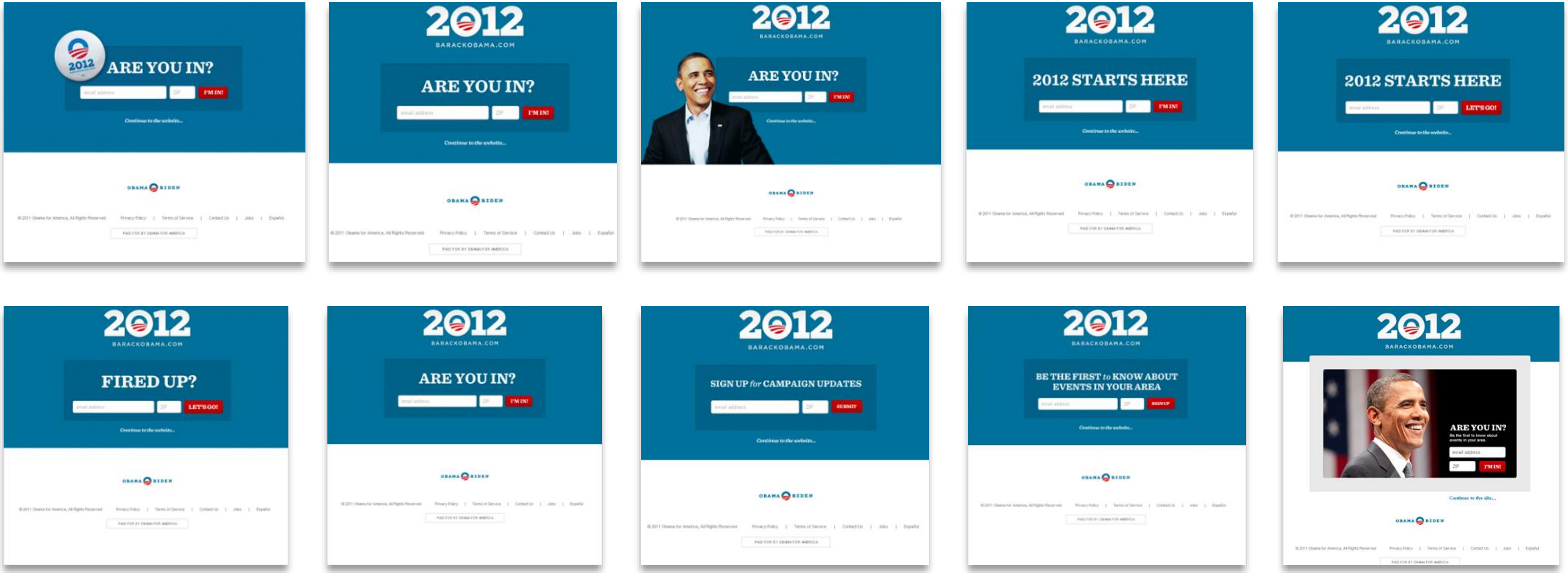


datasciencedojo
unleash the data scientist in you

Obama 2012 Campaign

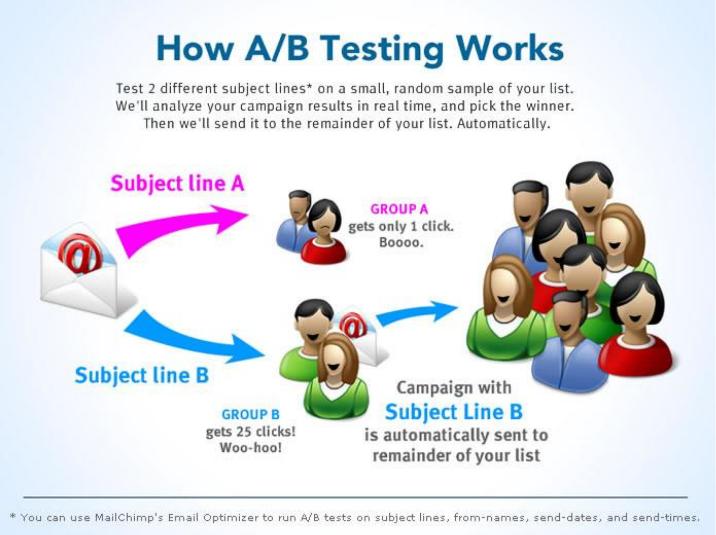
Obama 2012 Campaign

Maximize Sign-Ups And Donations



Source: http://www.nathanielward.net/2011/06/see-ab-testing-in-action-on-barack-obamas-reelection-website/

A/B Testing On Newsletters And Email



How A/B Testing Works

Test 2 different subject lines* on a small, random sample of your list.
We'll analyze your campaign results in real time, and pick the winner.
Then we'll send it to the remainder of your list. Automatically.

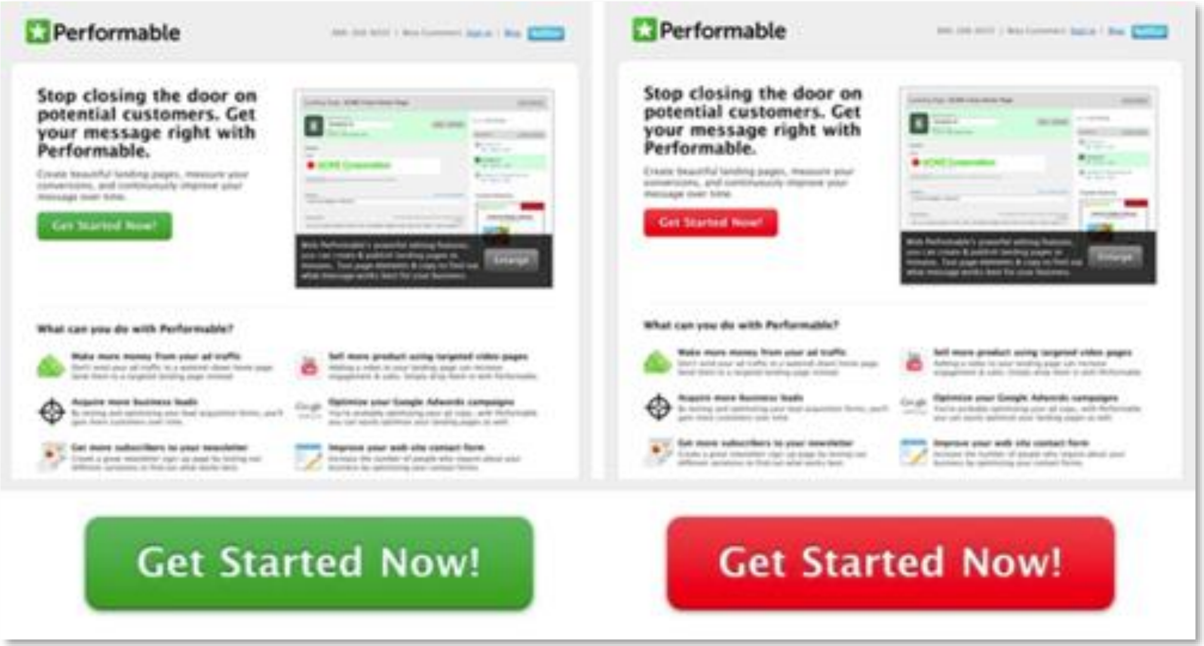
* You can use MailChimp's Email Optimizer to run A/B tests on subject lines, from-names, send-dates, and send-times.

Run tests on many things

- Subject lines
- **From** names
- **Send** dates
- **Send** time

datasciencedojo
unleash the data scientist in you

Testing Call-to-Action Button



Performable

Stop closing the door on potential customers. Get your message right with Performable.

Create beautiful landing pages, measure your conversions, and continuously improve your message over time.

Get Started Now!

What can you do with Performable?

Make more money from your ad traffic

Sell more product using targeted video pages

Acquire more business leads

Optimize your Google Adwords campaigns

Get more subscribers to your newsletter

Improve your web site contact form

Performable

Stop closing the door on potential customers. Get your message right with Performable.

Create beautiful landing pages, measure your conversions, and continuously improve your message over time.

Get Started Now!

What can you do with Performable?

Make more money from your ad traffic

Sell more product using targeted video pages

Acquire more business leads

Optimize your Google Adwords campaigns

Get more subscribers to your newsletter

Improve your web site contact form

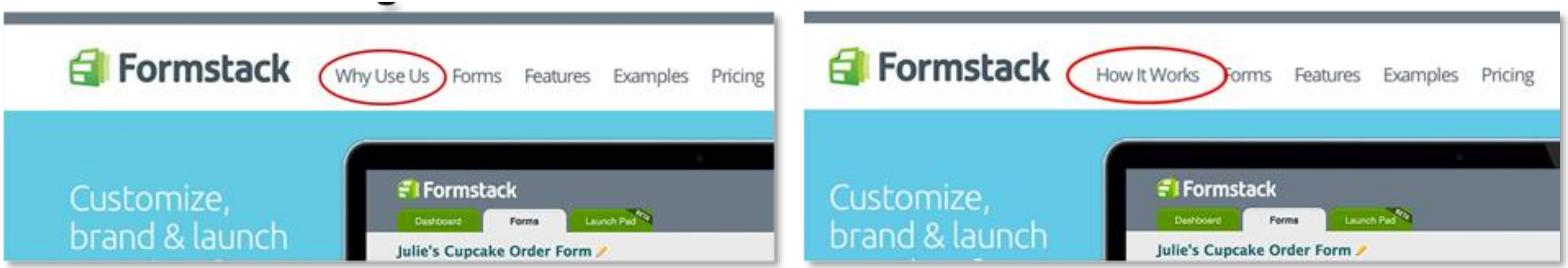
Get Started Now!

Get Started Now!

Red button increased clicks by **21%**

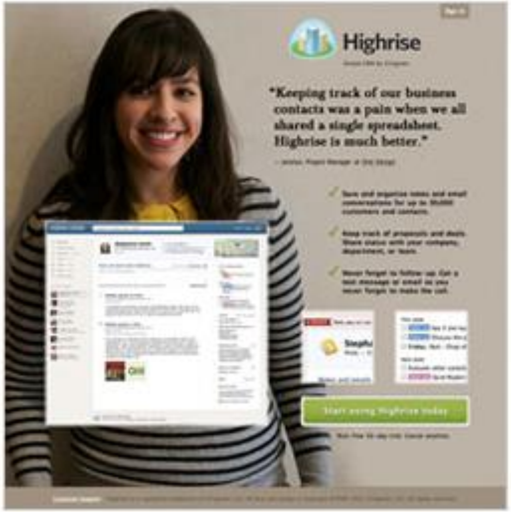
datasciencedojo
unleash the data scientist in you

Testing Navigation Bar



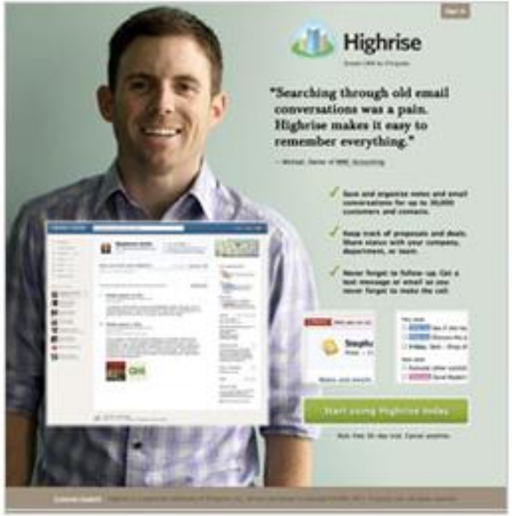
‘How It Works‘ increased clicks by 47.7%

Jocelyn or Michael?



"Keeping track of our business contacts was a pain when we all shared a single spreadsheet. Highrise is much better."

JOCELYN

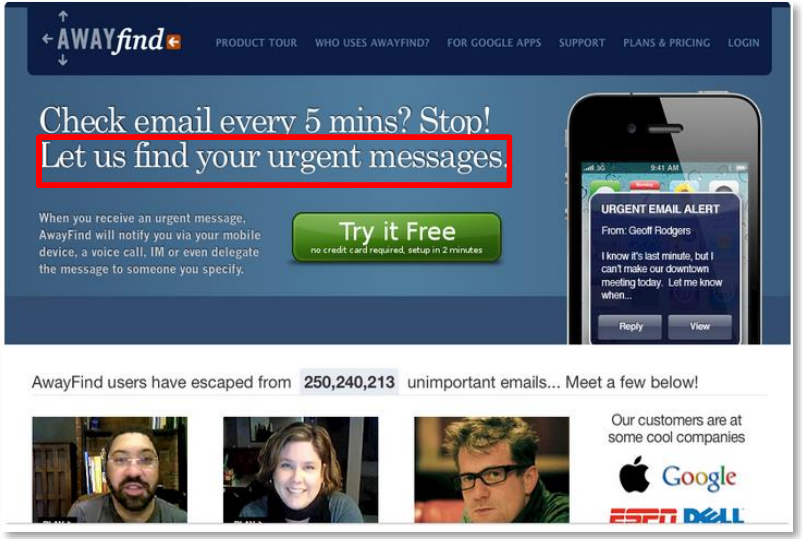


"Searching through old email conversations was a pain. Highrise makes it easy to remember everything."

MICHAEL

Michael increased conversions by 21%

AwayFind - Mobile notifications for priority messages



Check email every 5 mins? Stop!
Let us find your urgent messages.

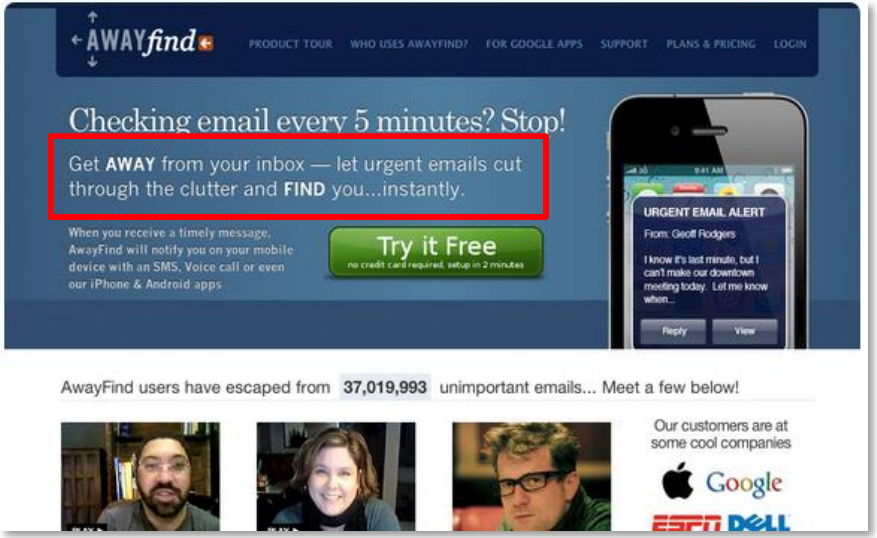
When you receive an urgent message, AwayFind will notify you via your mobile device, a voice call, IM or even delegate the message to someone you specify.

URGENT EMAIL ALERT
From: Geoff Rodgers
I know it's last minute, but I can't make our downtown meeting today. Let me know when...

AwayFind users have escaped from 250,240,213 unimportant emails... Meet a few below!

Our customers are at some cool companies

Version A



Checking email every 5 minutes? Stop!

Get **AWAY** from your inbox — let urgent emails cut through the clutter and **FIND** you...instantly.

When you receive a timely message, AwayFind will notify you on your mobile device with an SMS, Voice call or even our iPhone & Android apps

URGENT EMAIL ALERT
From: Geoff Rodgers
I know it's last minute, but I can't make our downtown meeting today. Let me know when...

AwayFind users have escaped from 37,019,993 unimportant emails... Meet a few below!

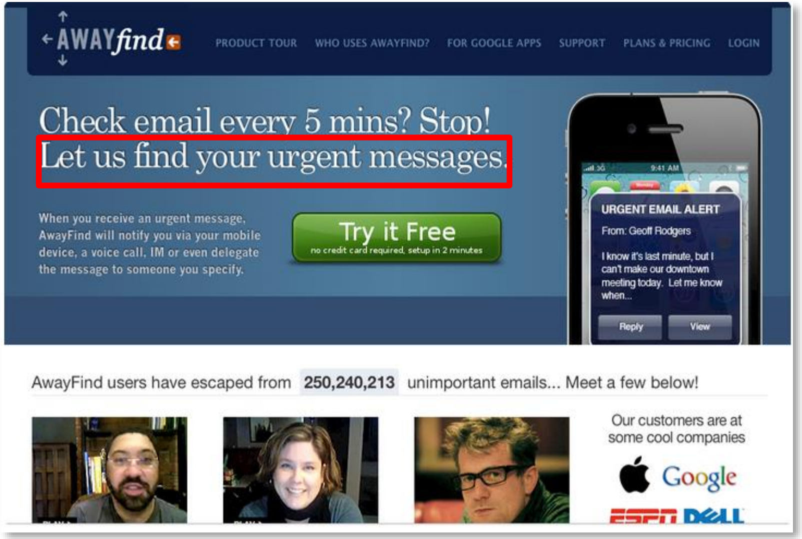
Our customers are at some cool companies

Version B

Which version increased sign-ups by 38%?

datasciencedojo
unleash the data scientist in you

AwayFind - Mobile notifications for priority messages



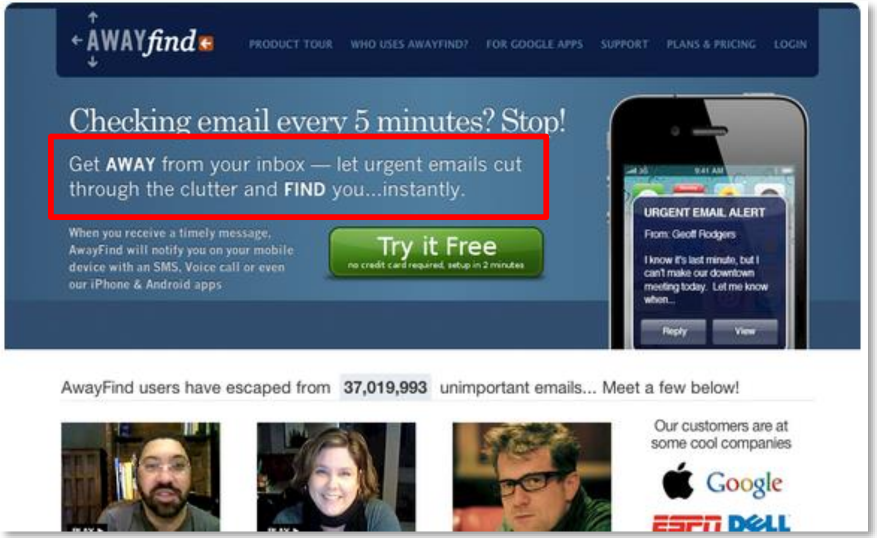
Check email every 5 mins? Stop!
Let us find your urgent messages.

When you receive an urgent message, AwayFind will notify you via your mobile device, a voice call, IM or even delegate the message to someone you specify.

URGENT EMAIL ALERT
From: Geoff Rodgers
I know it's last minute, but I can't make our downtown meeting today. Let me know when...

AwayFind users have escaped from 250,240,213 unimportant emails... Meet a few below!

Our customers are at some cool companies



Checking email every 5 minutes? Stop!
Get **AWAY** from your inbox — let urgent emails cut through the clutter and **FIND** you...instantly.

When you receive a timely message, AwayFind will notify you on your mobile device with an SMS, Voice call or even our iPhone & Android apps

URGENT EMAIL ALERT
From: Geoff Rodgers
I know it's last minute, but I can't make our downtown meeting today. Let me know when...

AwayFind users have escaped from 37,019,993 unimportant emails... Meet a few below!

Our customers are at some cool companies

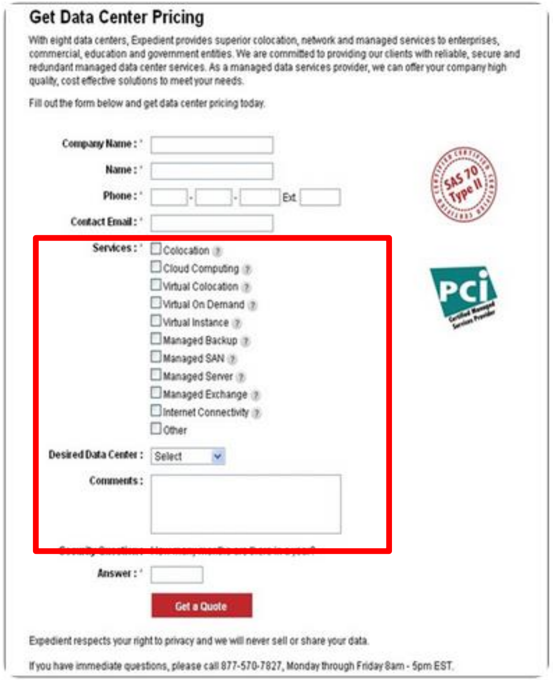
Version A

Version B

Version B!
A longer yet clearer message is more effective.

datasciencedojo
unleash the data scientist in you

Online Form



Get Data Center Pricing

With eight data centers, Expedient provides superior colocation, network and managed services to enterprises, commercial, education and government entities. We are committed to providing our clients with reliable, secure and redundant managed data center services. As a managed data services provider, we can offer your company high quality, cost effective solutions to meet your needs.

Fill out the form below and get data center pricing today.

Company Name : *

Name : *

Phone : * - - Ext

Contact Email : *

Services : *
☐ Colocation ?
☐ Cloud Computing ?
☐ Virtual Colocation ?
☐ Virtual On Demand ?
☐ Virtual Instance ?
☐ Managed Backup ?
☐ Managed SAN ?
☐ Managed Server ?
☐ Managed Exchange ?
☐ Internet Connectivity ?
☐ Other

Desired Data Center : Select

Comments :

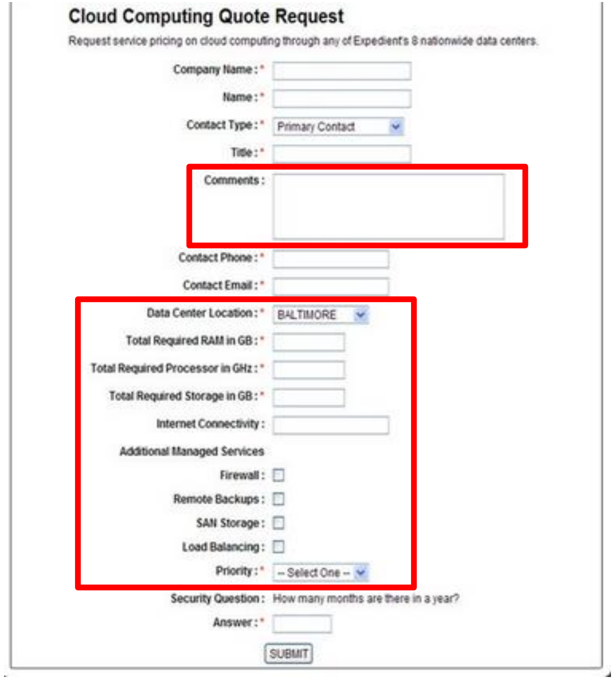
Security Question : How many months are there in a year?

Answer : *

Expedient respects your right to privacy and we will never sell or share your data.

If you have immediate questions, please call 877-570-7827, Monday through Friday 8am - 5pm EST.

Version A



Cloud Computing Quote Request

Request service pricing on cloud computing through any of Expedient's 8 nationwide data centers.

Company Name : *

Name : *

Contact Type : * Primary Contact

Title : *

Comments :

Contact Phone : *

Contact Email : *

Data Center Location : * BALTIMORE

Total Required RAM in GB : *

Total Required Processor in GHz : *

Total Required Storage in GB : *

Internet Connectivity :

Additional Managed Services

Firewall : ☐

Remote Backups : ☐

SAN Storage : ☐

Load Balancing : ☐

Priority : * -- Select One --

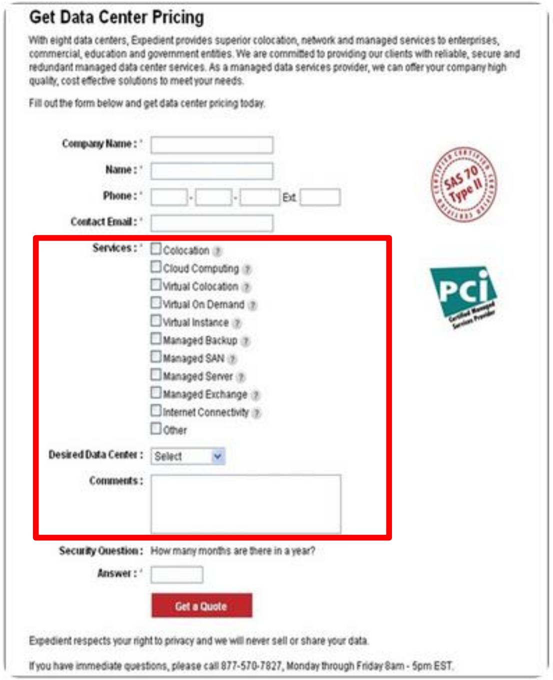
Security Question : How many months are there in a year?

Answer : *

Version B

Which Radically Redesigned Form Increased B2B Leads By 368.5%?

Online Form



Get Data Center Pricing

With eight data centers, Expedient provides superior colocation, network and managed services to enterprises, commercial, education and government entities. We are committed to providing our clients with reliable, secure and redundant managed data center services. As a managed data services provider, we can offer your company high quality, cost effective solutions to meet your needs.

Fill out the form below and get data center pricing today.

Company Name : *

Name : *

Phone : * - - Ext

Contact Email : *

Services : *
□ Colocation ?
□ Cloud Computing ?
□ Virtual Colocation ?
□ Virtual On Demand ?
□ Virtual Instance ?
□ Managed Backup ?
□ Managed SAN ?
□ Managed Server ?
□ Managed Exchange ?
□ Internet Connectivity ?
□ Other

Desired Data Center : Select

Comments :

Security Question : How many months are there in a year?

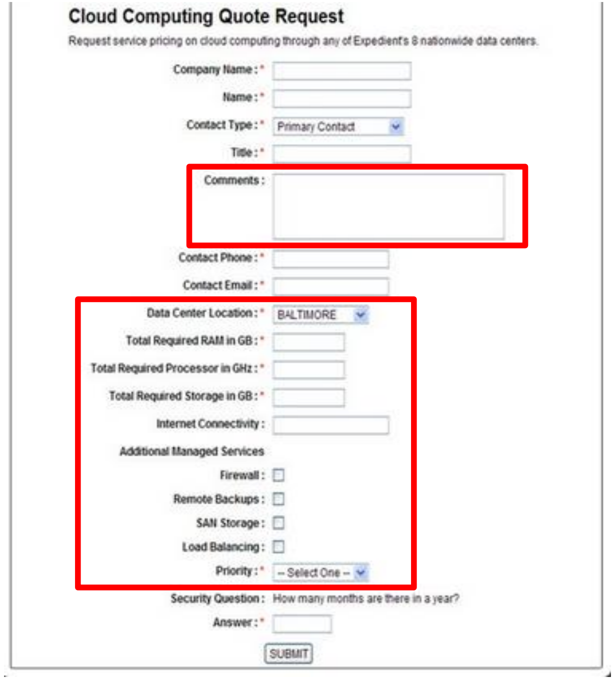
Answer : *

Get a Quote

Expedient respects your right to privacy and we will never sell or share your data.

If you have immediate questions, please call 877-570-7827, Monday through Friday 8am - 5pm EST.

Version A



Cloud Computing Quote Request

Request service pricing on cloud computing through any of Expedient's 8 nationwide data centers.

Company Name : *

Name : *

Contact Type : * Primary Contact

Title : *

Comments :

Contact Phone : *

Contact Email : *

Data Center Location : * BALTIMORE

Total Required RAM in GB : *

Total Required Processor in GHz : *

Total Required Storage in GB : *

Internet Connectivity :

Additional Managed Services

Firewall : □

Remote Backups : □

SAN Storage : □

Load Balancing : □

Priority : * -- Select One --

Security Question : How many months are there in a year?

Answer : *

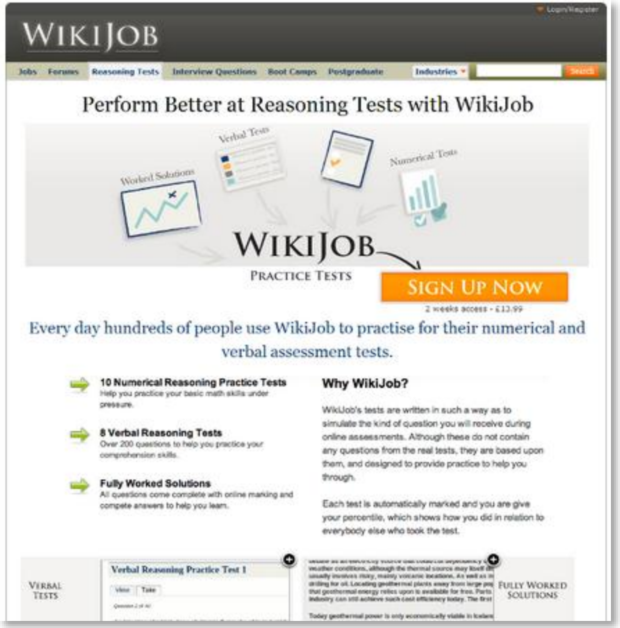
SUBMIT

Version B

Version A!
Better be to the point

datasciencedojo
unleash the data scientist in you

WIKIJOB



Perform Better at Reasoning Tests with WikiJob

2 weeks access - £13.99

Every day hundreds of people use WikiJob to practise for their numerical and verbal assessment tests.

10 Numerical Reasoning Practice Tests
Help you practice your basic math skills under pressure.

8 Verbal Reasoning Tests
Over 200 questions to help you practice your comprehension skills.

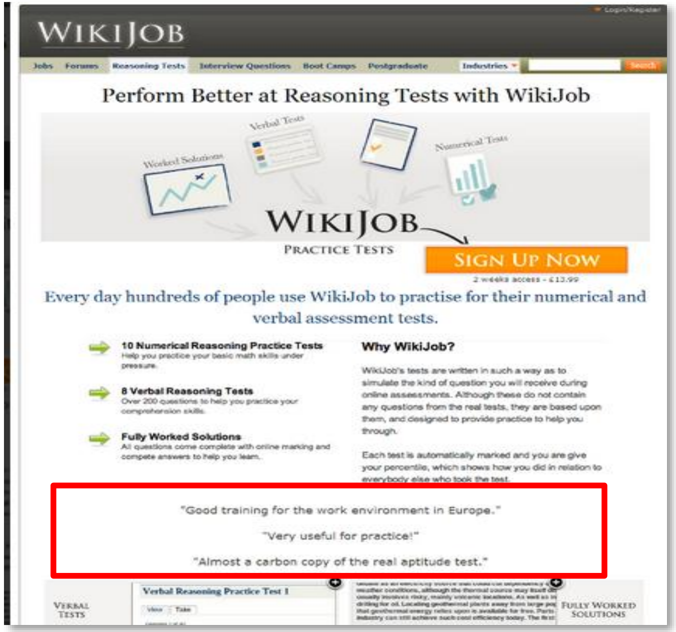
Fully Worked Solutions
All questions come complete with online marking and compete answers to help you learn.

Why WikiJob?

WikiJob's tests are written in such a way as to simulate the kind of question you will receive during online assessments. Although these do not contain any questions from the real tests, they are based upon them, and designed to provide practice to help you through.

Each test is automatically marked and you are give your percentile, which shows how you did in relation to everybody else who took the test.

Version A



Perform Better at Reasoning Tests with WikiJob

2 weeks access - £13.99

Every day hundreds of people use WikiJob to practise for their numerical and verbal assessment tests.

10 Numerical Reasoning Practice Tests
Help you practice your basic math skills under pressure.

8 Verbal Reasoning Tests
Over 200 questions to help you practice your comprehension skills.

Fully Worked Solutions
All questions come complete with online marking and compete answers to help you learn.

Why WikiJob?

WikiJob's tests are written in such a way as to simulate the kind of question you will receive during online assessments. Although these do not contain any questions from the real tests, they are based upon them, and designed to provide practice to help you through.

Each test is automatically marked and you are give your percentile, which shows how you did in relation to everybody else who took the test.

"Good training for the work environment in Europe."

"Very useful for practice!"

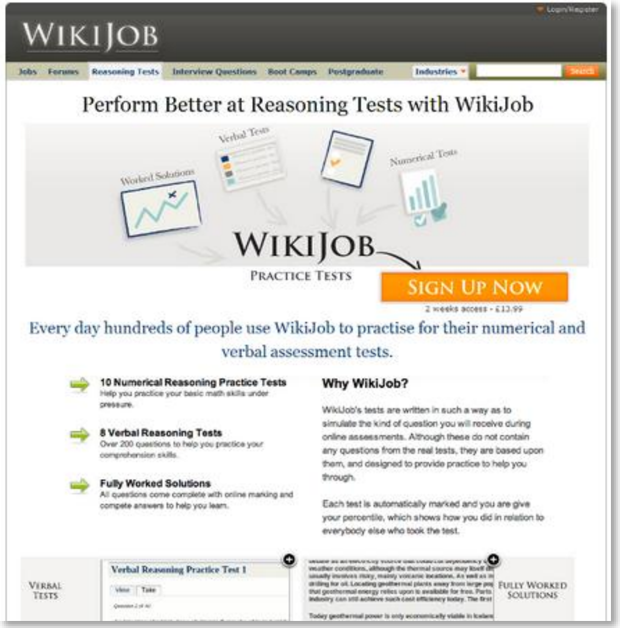
"Almost a carbon copy of the real aptitude test."

Testimonials

Version B

Version B has **testimonials**, does it work?

WIKIJOB



Perform Better at Reasoning Tests with WikiJob

2 weeks access - £13.99

Every day hundreds of people use WikiJob to practise for their numerical and verbal assessment tests.

10 Numerical Reasoning Practice Tests
Help you practice your basic math skills under pressure.

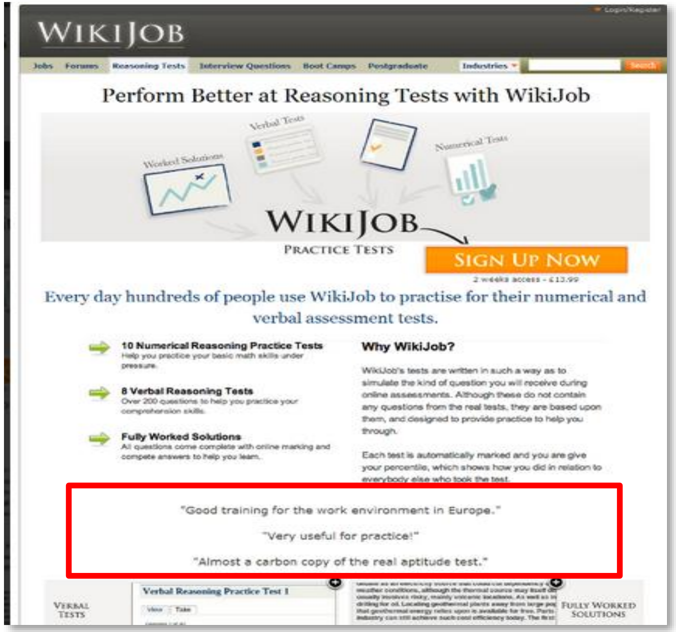
8 Verbal Reasoning Tests
Over 200 questions to help you practice your comprehension skills.

Fully Worked Solutions
All questions come complete with online marking and compete answers to help you learn.

Why WikiJob?

WikiJob's tests are written in such a way as to simulate the kind of question you will receive during online assessments. Although these do not contain any questions from the real tests, they are based upon them, and designed to provide practice to help you through.

Each test is automatically marked and you are give your percentile, which shows how you did in relation to everybody else who took the test.



Perform Better at Reasoning Tests with WikiJob

2 weeks access - £13.99

Every day hundreds of people use WikiJob to practise for their numerical and verbal assessment tests.

10 Numerical Reasoning Practice Tests
Help you practice your basic math skills under pressure.

8 Verbal Reasoning Tests
Over 200 questions to help you practice your comprehension skills.

Fully Worked Solutions
All questions come complete with online marking and compete answers to help you learn.

Why WikiJob?

WikiJob's tests are written in such a way as to simulate the kind of question you will receive during online assessments. Although these do not contain any questions from the real tests, they are based upon them, and designed to provide practice to help you through.

Each test is automatically marked and you are give your percentile, which shows how you did in relation to everybody else who took the test.

"Good training for the work environment in Europe."

"Very useful for practice!"

"Almost a carbon copy of the real aptitude test."

Testimonials

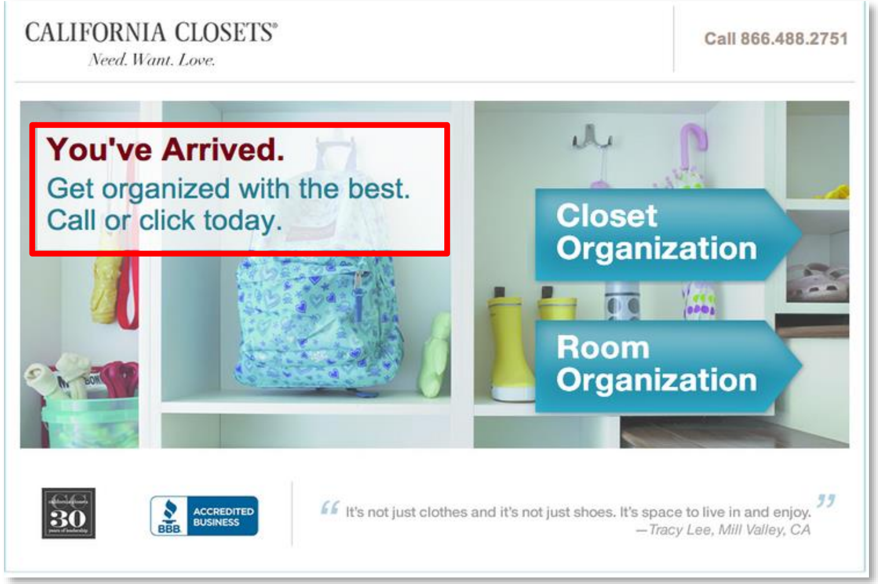
Version A

Version B

Yes, testimonials increased sales by 34%

datasciencedojo
unleash the data scientist in you

CALIFORNIA CLOSET



CALIFORNIA CLOSETS®
Need. Want. Love.

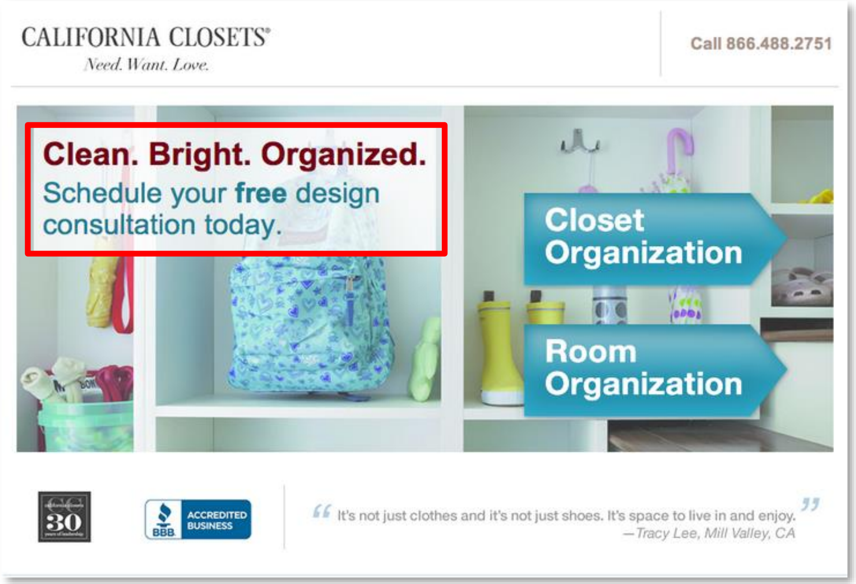
Call 866.488.2751

You've Arrived.
Get organized with the best.
Call or click today.

ACCREDITED BUSINESS BBB

"It's not just clothes and it's not just shoes. It's space to live in and enjoy."
—Tracy Lee, Mill Valley, CA

Version A



CALIFORNIA CLOSETS®
Need. Want. Love.

Call 866.488.2751

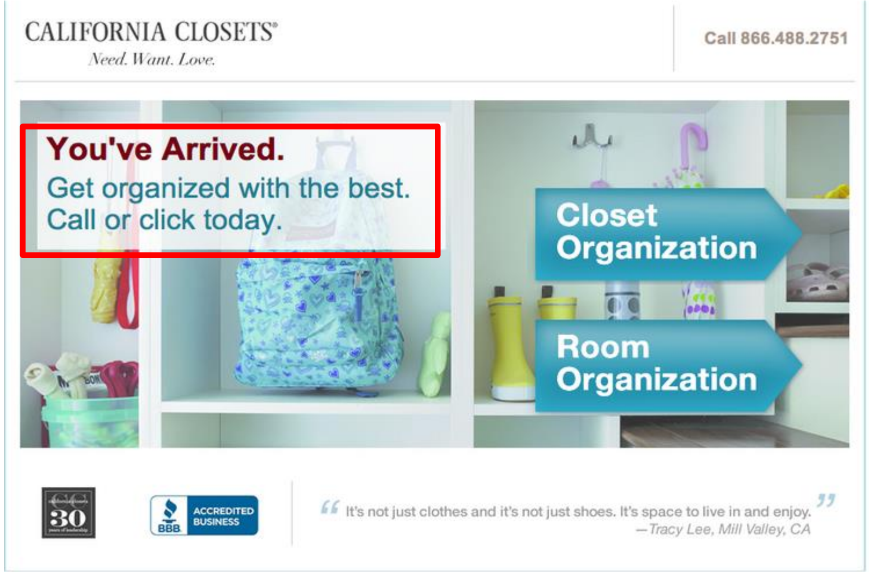
Clean. Bright. Organized.
Schedule your **free** design consultation today.

ACCREDITED BUSINESS BBB

"It's not just clothes and it's not just shoes. It's space to live in and enjoy."
—Tracy Lee, Mill Valley, CA

Version B

CALIFORNIA CLOSET



CALIFORNIA CLOSETS®
Need. Want. Love.

Call 866.488.2751

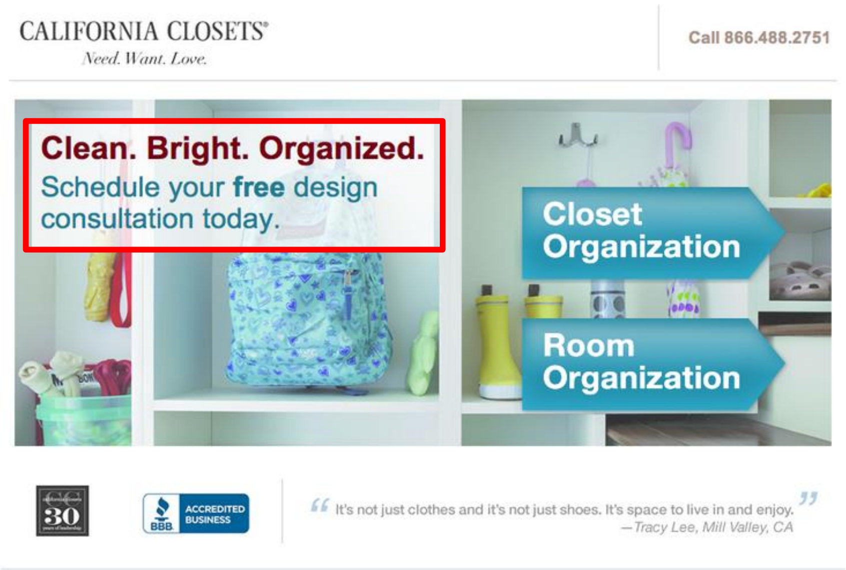
You've Arrived.
Get organized with the best.
Call or click today.

Closet Organization

Room Organization

"It's not just clothes and it's not just shoes. It's space to live in and enjoy."
—*Tracy Lee, Mill Valley, CA*

Version A



CALIFORNIA CLOSETS®
Need. Want. Love.

Call 866.488.2751

Clean. Bright. Organized.
Schedule your **free** design consultation today.

Closet Organization

Room Organization

"It's not just clothes and it's not just shoes. It's space to live in and enjoy."
—*Tracy Lee, Mill Valley, CA*

Version B

Version A increased leads by 115%.
This is why you should test…!

datasciencedojo
unleash the data scientist in you

Fundamentals

Why We Use A/B Testing

Problem

- Users are complex and our intuition is often wrong

- Rolling out a feature to all the users at the same time is risky

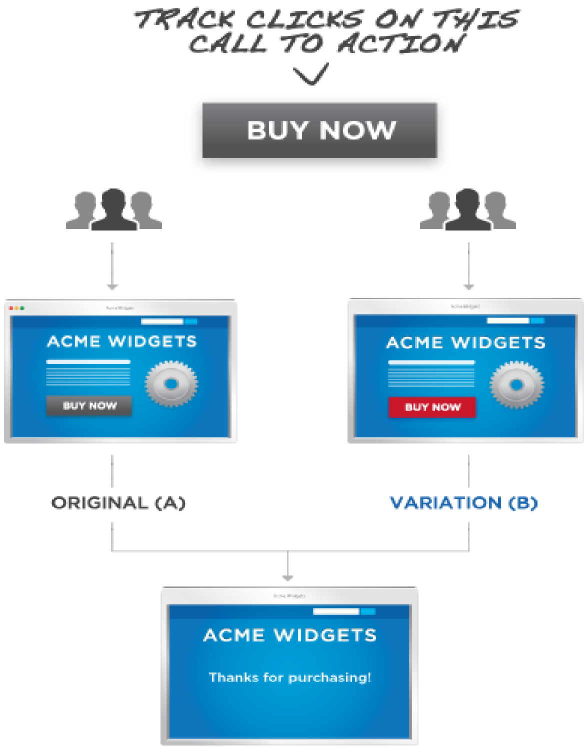
A/B testing purpose

- Know what the users want subconsciously or otherwise.

- Helps to fail fast and move on

Impact is always expected to be positive, but outcome is often humbling

A/B Testing vs. Multivariate Testing



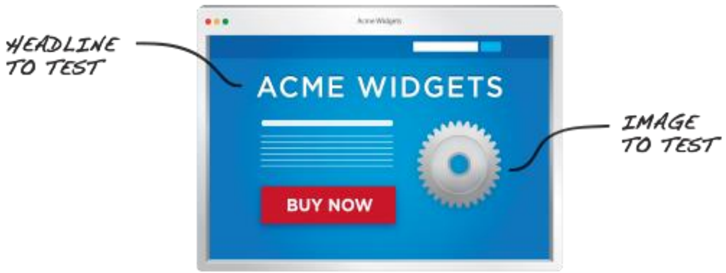


	Image		Headline
VERSION 1	⚙	+	"ACME WIDGETS"
VERSION 2	💡	+	"ACME WIDGETS"
VERSION 3	⚙	+	"THE ONE AND ONLY ACME WIDGETS"
VERSION 4	💡	+	"THE ONE AND ONLY ACME WIDGETS"

A/B Testing vs Multivariate Testing

	A/B Testing	Multivariate Testing
Common use	Compare two very different designs with each other	Several minor variations are up for debate: ➢ Two colors of button with three different headlines ⓘ Also called full factorial testing
Advantages	➢ Simple in design ➢ Small sample size may be ok	A lot of different combinations tried at once.
Limitations	Trying only one alternative	➢ Bigger sample size ➢ Complex ➢ Need better understanding of interactions

Terminology

Control and Treatment

Control

Default experience, the way things are now.

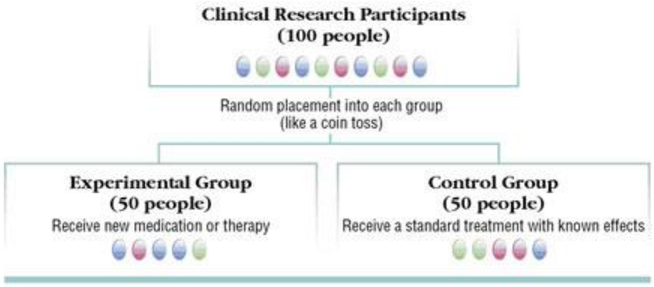
Example: Current look and feel of your 'Buy Now' button

Treatment

The change we want to make.

Example: Change the button from green to blue

Illustration



Clinical Research Participants
(100 people)

Random placement into each group
(like a coin toss)

Experimental Group
(50 people)
Receive new medication or therapy

Control Group
(50 people)
Receive a standard treatment with known effects

Factor and Level

Factor

- The item we want change

Level

- The variations of factor

ADD TO WISHLIST
Buy It Now
Add to Cart
proceed to che   ut
Add to
PAY SECURELY

datasciencedojo
unleash the data scientist in you

Metrics Used For A/B Testing

- **Search engines**

  Queries/UU, Session length, Sessions/UU, Page views, Bounce rate

- **Online Retailers**

  Conversion rate, revenue/UU, Avg Cart Value and so on

- **Other websites**

  CTR, signup for newsletter

Each business is different

Brainstorming

OEC: Overall evaluation Criteria

- Summarizes the primary indicator of success
- May be one of the metrics or a combination of metrics

Null vs Alternate Hypothesis

- Null Hypothesis (H_o)
  - Control and treatment are similar (in terms of the parameter we are estimating)
- Alternate Hypothesis (H_a)
  - Treatment is different from control

Null vs Alternate Hypothesis

Control

Treatment

- **Null Hypothesis (H_o)**
  - Green and blue buttons have the same CTR
- **Alternate Hypothesis (H_a)**
  - Each button has a different CTR

Type I and Type II Error

Type I Error

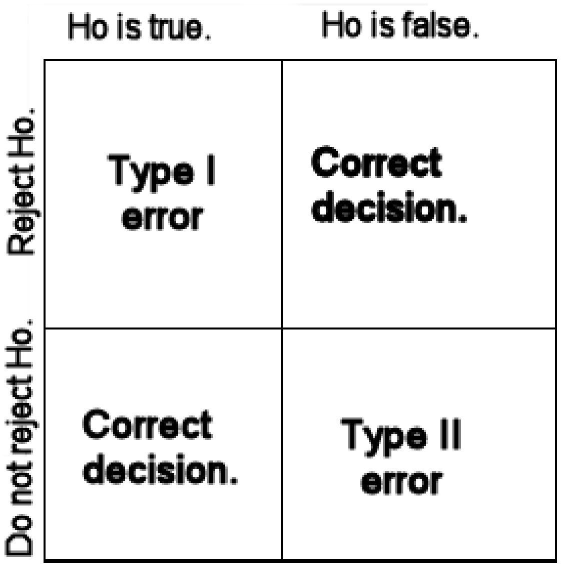
The probability of **falsely rejecting** null hypothesis

Type II Error

The probability of **falsely accepting** null hypothesis

Ground Truth

Experiment Outcome



	Ho is true.	Ho is false.
Reject Ho.	Type I error	Correct decision.
Do not reject Ho.	Correct decision.	Type II error

Power

- Power of an online experiment is the probability of **not** rejecting the null hypothesis false
- Which is really 1 – Probability (Type II Error)

Can you tell me in simple words…

datasciencedojo
unleash the data scientist in you

The Cook and Smoke Detector

- Null Hypothesis (Ho): There is no fire
- Alternate Hypothesis (Ha): There is fire

The Cook and Smoke Detector

- **Type I Error:** There is no fire but smoke detector goes off.
- The cook removes the alarm to prevent type I error.
- This increases the chance of Type II Error i.e. a fire without an alarm

datasciencedojo
unleash the data scientist in you

The Boy Who Cried Wolf

- Null Hypothesis (Ho): There is no wolf
- Alternate Hypothesis (Ha): There is a wolf

The Boy Who Cried Wolf

- **Type I Error:** Villagers believe the boy when there is no wolf
- **Type II Error:** Villagers do not believe the boy when the wolf is really there

Confidence Intervals

Problem: On a 5-point scale, a product has an average review of 4.32 and a standard deviation of 0.845 based on 62 participants in the study. What is the 95% confidence interval?

$$\bar{x} \pm 1.96\,\sigma/\sqrt{n}$$

Confidence Intervals

Mean $\bar{X} = 4.32$

Standard deviation $\sigma = 0.845$

Standard error SE = $\frac{0.845}{\sqrt{n}} = \frac{0.845}{\sqrt{62}} = 0.11$

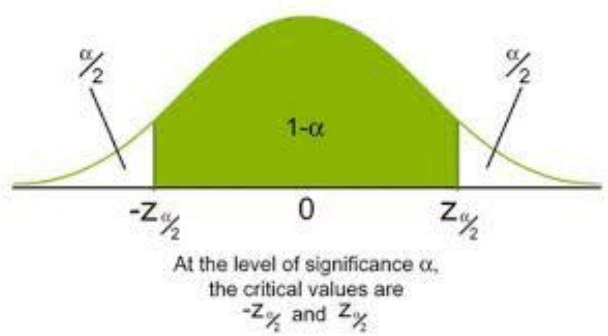
Margin or error is 2 x 0.11 = 0.22

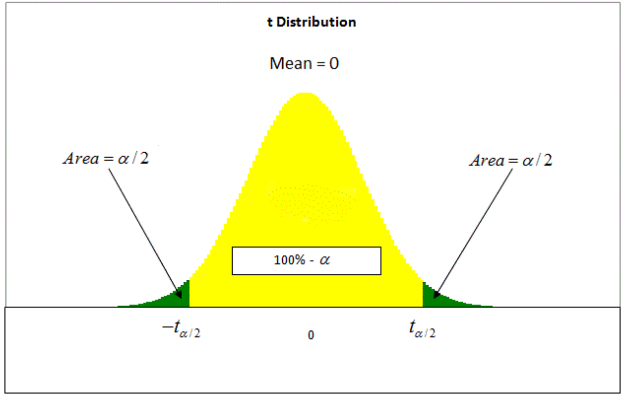
The confidence interval is

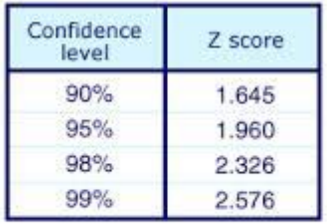
4.32+0.22 = 4.54

4.32 – 0.22 = 4.10

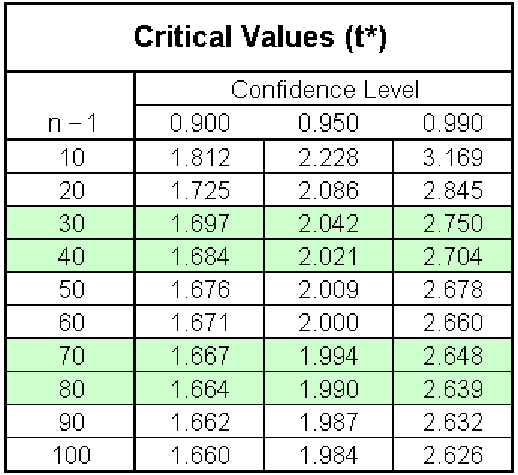
Calculating Confidence Interval


$\alpha/2$
$\alpha/2$
1-α
$-Z_{\alpha/2}$
0
$Z_{\alpha/2}$
At the level of significance α, the critical values are $-Z_{\alpha/2}$ and $Z_{\alpha/2}$





Confidence level	Z score
90%	1.645
95%	1.960
98%	2.326
99%	2.576



Critical Values (t*)

	Confidence Level		
n − 1	0.900	0.950	0.990
10	1.812	2.228	3.169
20	1.725	2.086	2.845
30	1.697	2.042	2.750
40	1.684	2.021	2.704
50	1.676	2.009	2.678
60	1.671	2.000	2.660
70	1.667	1.994	2.648
80	1.664	1.990	2.639
90	1.662	1.987	2.632
100	1.660	1.984	2.626

Type I and Type II Error

Type I Error

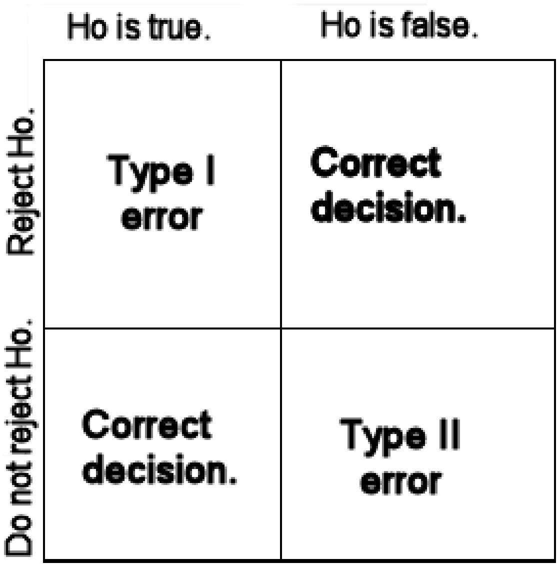
The probability of **falsely rejecting** null hypothesis

Type II Error

The probability of **falsely accepting** null hypothesis

Ground Truth

Experiment Outcome



	Ho is true.	Ho is false.
Reject Ho.	**Type I error**	**Correct decision.**
Do not reject Ho.	**Correct decision.**	**Type II error**

Type I and Type II Error

Type I Error

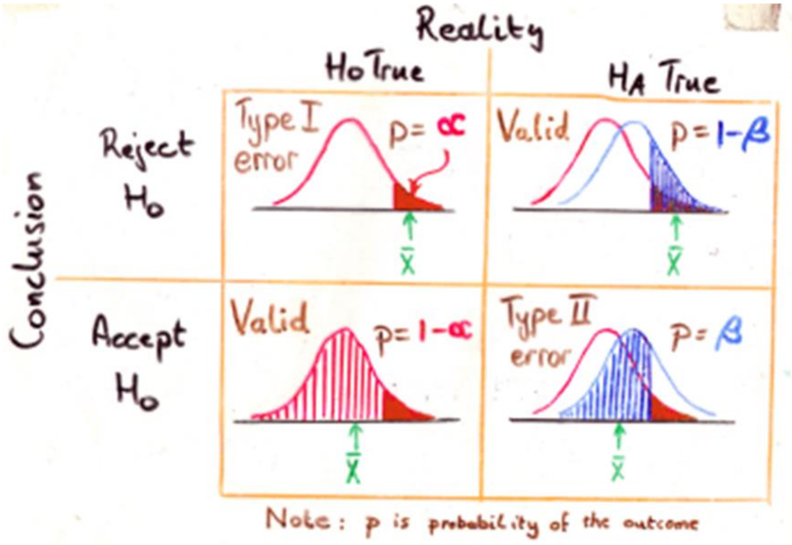
The probability of **falsely accepting** null hypothesis

Type II Error

The probability of **falsely rejecting** null hypothesis

Ground Truth

Experiment Outcome



Note: p is probability of the outcome

Confidence Interval

- Range of plausible values of parameter being estimated given the sample data

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1989 PAWS, INC. All Rights Reserved.
JIM DAVIS 3-12

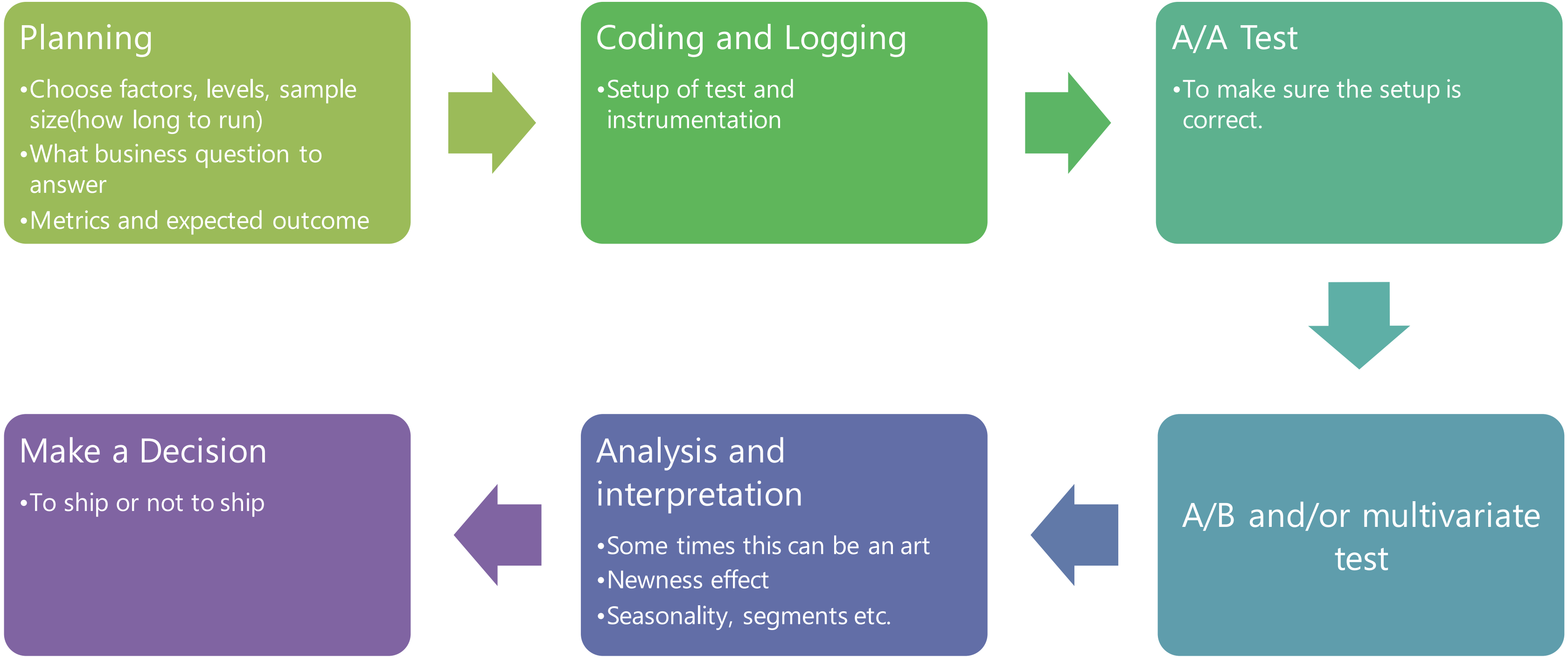
A/A Test

- Comparing the identical experience on different random sets of users
- Used for validation of setup

Control

Treatment

Steps in Experimentation



Planning

- Choose factors, levels, sample size(how long to run)
- What business question to answer
- Metrics and expected outcome

Coding and Logging

- Setup of test and instrumentation

A/A Test

- To make sure the setup is correct.

A/B and/or multivariate test

Analysis and interpretation

- Some times this can be an art
- Newness effect
- Seasonality, segments etc.

Make a Decision

- To ship or not to ship

Categories of Metrics

	Short-term	Medium-term	Long-term
Examples	➢ CTR ➢ PVs ➢ Bounce Rate	➢ PVs/user/day ➢ CTR/user /day ➢ Avg session length	**Days with at least one visit:** ➢ Total time on site ➢ Repeat visits/user
What is measured?	Immediate or almost immediate impact	Engagement over hours up to a day	Loyalty

Common Pitfalls

Pitfalls in Online Experimentation

1. Picking an OEC for which it is easy to beat the control
2. Incorrectly computing the confidence intervals
3. Using standard statistical formulas for computation of variance and power
4. Combining metrics over periods where proportions assigned to Control and Treatment vary or over subpopulations sampled at different rates
5. Neglecting to filter bots
6. Failing to validate each step of the analysis pipeline and the OEC components
7. Forgetting to control for all differences, and assuming that humans can keep the variants in sync

Pitfall 1: Picking an Easy-to-Beat Overall Evaluation Criteria (OEC)

- Before running an experiment an OEC is selected
- OEC should be tied to a long term goals as opposed to short term goals. Click-through Rate (CTR) vs. long term revenue
- Loyal/repeat users get more weight?
- Sometimes getting the true metric is hard. High CTR does not necessarily mean high conversion rate

Pitfall 1: Picking an Easy-to-Beat Overall Evaluation Criteria (OEC)

- Measuring click through on a small area of the page, ignoring the impact on other areas
  - What if the small area on the page was bold/flashing/high contrast?
  - What happens to the whole page CTR?
- Is ‘time on site’ a good OEC?
  - What if the treatment has a reduced user’s effectiveness?

Pitfall 2: Incorrect Computation of Confidence Intervals

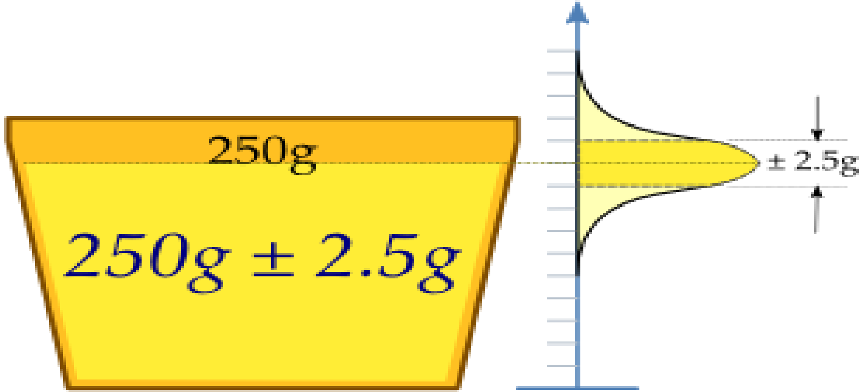
- Hypothesis Test: determines whether there is a statistically significant difference in the means of the control and the treatment
- Confidence Interval: provides a plausible range of the size of the effect (difference in C and T means)

Pitfall 2: Incorrect Computation of Confidence Intervals

$$0.95 = 1 - \alpha = P(-z \leq Z \leq z) = P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right)$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X} - 1.96 \times 0.5 \leq \mu \leq \bar{X} + 1.96 \times 0.5\right)$$



$$(\bar{x} - 0.98; \bar{x} + 0.98) = (250.2 - 0.98; 250.2 + 0.98) = (249.22; 251.18).$$

Confidence interval implies: If we randomly fill a cup from this vending machine, there is a 95% chance that our cup will have this much coffee

Pitfall 2: Incorrect Computation of Confidence Intervals

- Confidence interval should be formed out of absolute difference
- Do not form a confidence interval around percent change. Percentage change involves dividing by a random variable.
- Some techniques to compute CI are mentioned when the OEC is a linear/non-linear combination of metrics that have the same/different basis/experimental unit.

Pitfall 3: Standard Statistical Formulas for Computation of Variance and Power

- Variance of the metric is needed to compute the statistical significance
- Variance estimates using standard statistical formula for some families of metrics are inaccurate
- This happens when the experimental unit used in random assignment is different from the experiment unit used in the calculation of the metric.

Pitfall 3: Standard Statistical Formulas for Computation of Variance and Power

- Variance, Power and Sample size estimates may be wrong if care is not taken
- How to correct this?
  - Bootstrap method: Estimate variance using bootstrap samples and compare with the variance from standard formula
- This should be done for all metrics and especially for the one with different experiment and randomization units

Pitfall 4: Simpson’s Paradox

- Unintuitive but not uncommon
- Simpson’s paradox: ‘A correlation or trend present in different groups is reversed when the groups are combined’.

	Treatment A	Treatment B
Small Stones	Group 1	Group 2
	93% (81/87)	87% (234/270)
Large Stones	Group 3	Group 4
	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

Pitfall 4: Simpson’s Paradox

- 1 million visitors/day
- On Friday the treatment ran with 1% traffic
- On Saturday, the allocation was raised to 50%.
- If we consider Friday and Saturday separately T has a better CTR
- T’s CTR is worse when aggregated over days

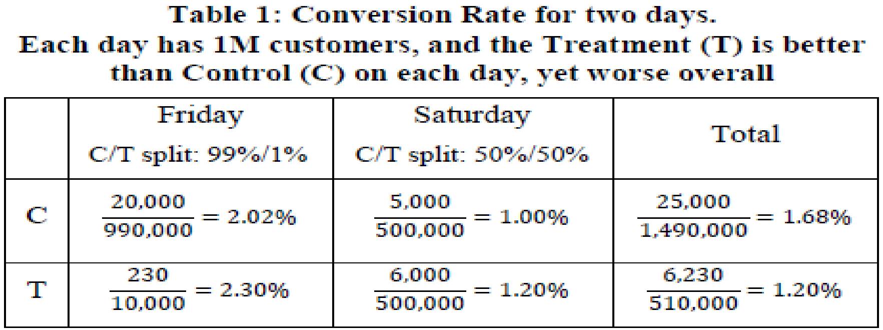


Table 1: Conversion Rate for two days. Each day has 1M customers, and the Treatment (T) is better than Control (C) on each day, yet worse overall

	Friday C/T split: 99%/1%	Saturday C/T split: 50%/50%	Total
C	$\frac{20{,}000}{990{,}000} = 2.02\%$	$\frac{5{,}000}{500{,}000} = 1.00\%$	$\frac{25{,}000}{1{,}490{,}000} = 1.68\%$
T	$\frac{230}{10{,}000} = 2.30\%$	$\frac{6{,}000}{500{,}000} = 1.20\%$	$\frac{6{,}230}{510{,}000} = 1.20\%$

It is possible to have $\frac{a}{b} < \frac{A}{B}$ and $\frac{c}{d} < \frac{C}{D}$ while $\frac{a+c}{b+d} > \frac{A+C}{B+D}$

Pitfall 4: Simpson’s Paradox – A Scenario in Controlled Experiments

- Sampling of users with non uniform sampling to make sure all browsers have a representative sample

- Overall results show treatment is better than control but when segmented by browser, control looks better than treatment for each browser

Pitfall 5: Ignoring Bot Traffic

- For experimentation, we are interested in removing bots/fraud clicks that are not uniformly distributed across the control and treatment

- Uniformly distributed bots will only reduce the power of the experiment

Pitfall 5: Ignoring Bot Traffic

Failing to exclude bot traffic and fraud clicks may invalidate the results of an experiment

Pitfall 6: Failing to Validate Each Step of Analysis

It is important to keep a check on the health of the pipeline

- Assignment of users to experiment variants
- Calculation of metrics
- Any abnormal shift in metrics
- Movement of metrics that are not expected to move
- Broken instrumentation

Pitfall 6: Failing to Validate Each Step of Analysis

Logging Tests:

- Compare with real historical data
- Compare with generated data
- Look for **unexpected patterns**
  - Volume of data over time
  - New and repeat users over time
  - Abnormal shift in any of the metrics
- A/A Tests
- Rich Instrumentation

Pitfall 7: Failing to ‘Control’ the Control

- **Don’t allow any difference** between the Control and the Treatment besides what is actually being tested

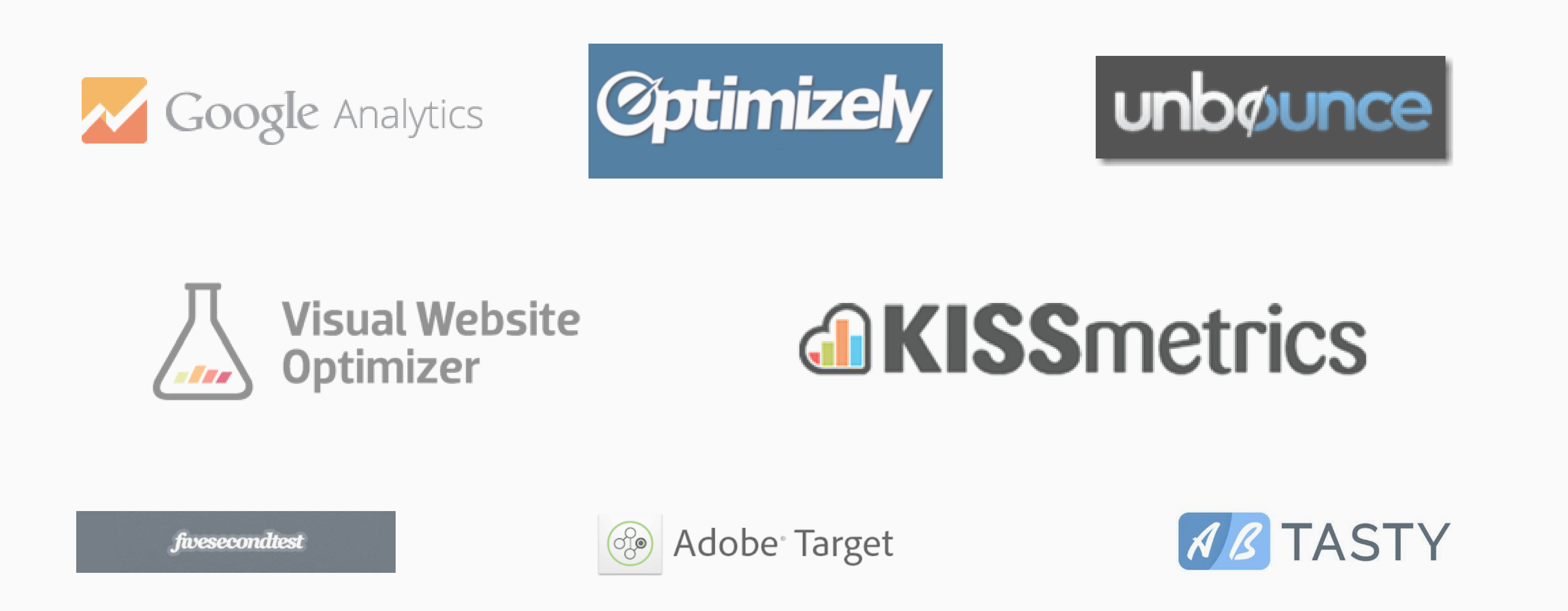
- If the **Treatment** has some **updates**, **Control** should have them too and vice versa

Pitfall 7: Failing to ‘Control’ the Control

- If the site is receiving **frequent updates**, these updates should be **applied equally** to the control and the treatment

- Forgetting to **control for all differences**, and assuming that humans can keep the variants in sync.

A/B Testing Tools



Humor

Have you heard the latest statistics joke?

Probably....

Did you hear about the statistician who was thrown in jail?

He now has zero degrees of freedom.

A statistician's wife has twins. He was delighted, and he called to tell his minister the good news.

"Excellent!", said the minister. "Bring them to church on Sunday and we'll baptize them.“

"No," replied the statistician. "Let's just baptize one. We'll keep the other as control."

Three statisticians go out hunting together. After a while they spot a solitary rabbit.

The first statistician takes aim and overshoots. The second aims and undershoots.

The third shouts out "We got him!"

How many statisticians does it take to change a light bulb?

1 – 3. α=0.05 (.95 confidence)

Questions?

datasciencedojo
unleash the data scientist in you

Enjoying the bootcamp?

We'd love it if you could write a short review of Data Science Dojo!

Switch Up (https://www.switchup.org/bootcamps/data-science-dojo)
Course Report (https://www.coursereport.com/schools/data-science-dojo)

Your reviews help other people find and attend our bootcamp.

datasciencedojo
unleash the data scientist in you

Appendix

Is a drug efficient?

A/B testing is often applied to test the efficiency of a drug, against a placebo, in order to control for the placebo effect in the drug.

CONTROL

DRUG

Examples:

- Betablocker
- Diastolic Blood Pressure (DBP)
- Polypses: we will study this case more specifically

The Polyps dataset (1/3)

- Data from a placebo-controlled trial of a non-steroidal anti-inflammatory drug in the treatment of familial andenomatous polyposis (FAP).

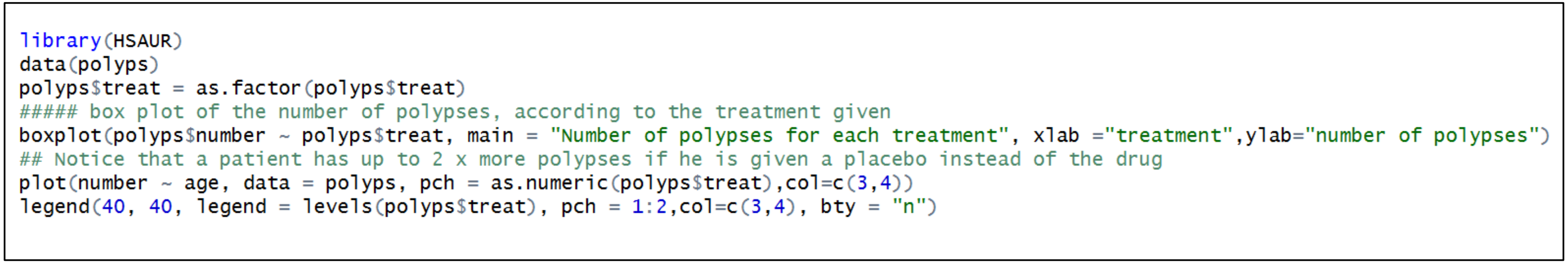
- The trial was halted after a planned interim analysis had suggested compelling evidence in favour of the treatment.

- Here we are interested in assessing whether the number of colonic polyps at 12 months is related to treatment and age of the patient.

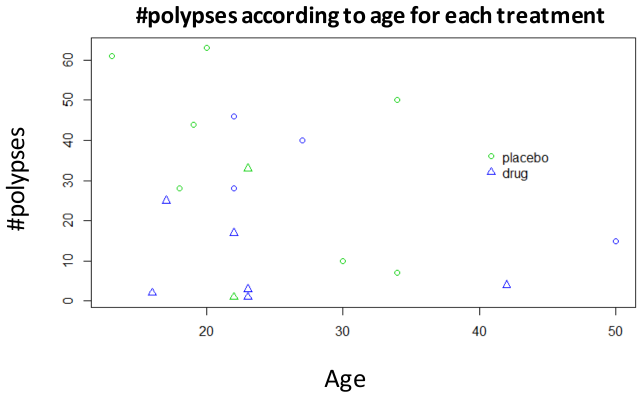
number	treat	age
63	placebo	20
2	drug	16
28	placebo	18
17	drug	22
61	placebo	13
1	drug	23
7	placebo	34
15	placebo	50
44	placebo	19
25	drug	17
3	drug	23
28	placebo	22
10	placebo	30
40	placebo	27
33	drug	23
46	placebo	22
50	placebo	34
3	drug	23

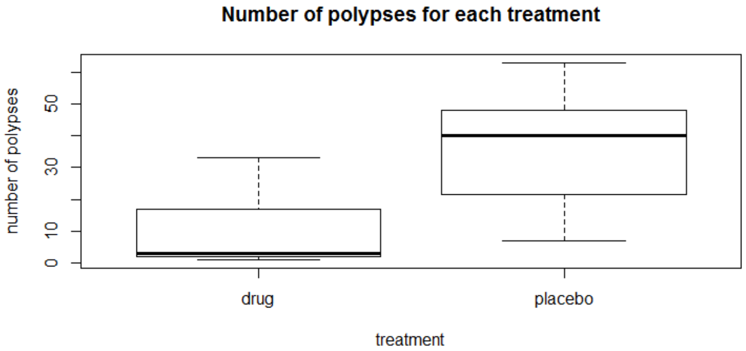
Extract from the dataset

The Polyps dataset (2/3)



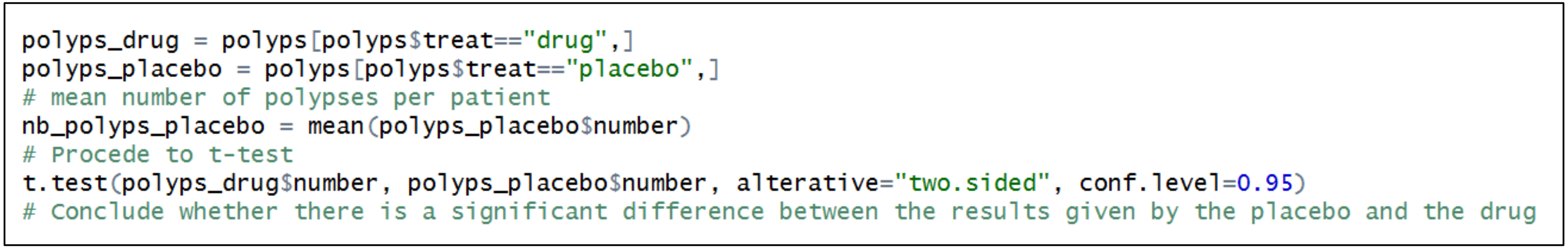
```
library(HSAUR)
data(polyps)
polyps$treat = as.factor(polyps$treat)
##### box plot of the number of polypses, according to the treatment given
boxplot(polyps$number ~ polyps$treat, main = "Number of polypses for each treatment", xlab ="treatment",ylab="number of polypses")
## Notice that a patient has up to 2 x more polypses if he is given a placebo instead of the drug
plot(number ~ age, data = polyps, pch = as.numeric(polyps$treat),col=c(3,4))
legend(40, 40, legend = levels(polyps$treat), pch = 1:2,col=c(3,4), bty = "n")
```





datasciencedojo
unleash the data scientist in you

The Polyps dataset (3/3)



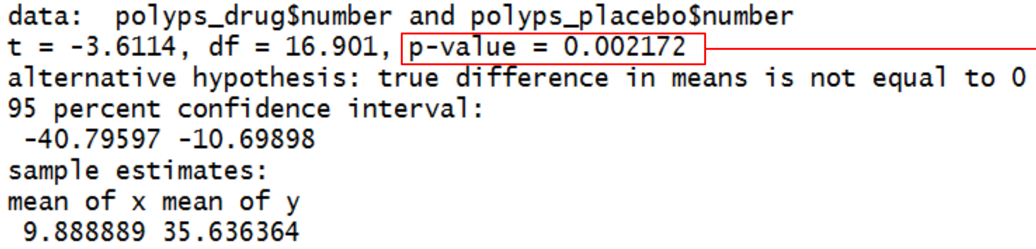
```
polyps_drug = polyps[polyps$treat=="drug",]
polyps_placebo = polyps[polyps$treat=="placebo",]
# mean number of polypses per patient
nb_polyps_placebo = mean(polyps_placebo$number)
# Procede to t-test
t.test(polyps_drug$number, polyps_placebo$number, alterative="two.sided", conf.level=0.95)
# Conclude whether there is a significant difference between the results given by the placebo and the drug
```

Mean number of polyps per individual

```
nb_polyps_placebo [1] 35.63636
```

Welch Two Sample t-test *

*used to test the hypothesis that two populations have equal means



```
data:  polyps_drug$number and polyps_placebo$number
t = -3.6114, df = 16.901, p-value = 0.002172
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -40.79597 -10.69898
sample estimates:
mean of x mean of y
 9.888889 35.636364
```

Which treatment is more efficient?