Web Experiments

Neil Sengupta and Calvin Xavier Gomez
Learning Goals

- A/B Testing
- Ramping Up and Auto abort
- The Facebook Experiment
- Debate on ethics
Controlled experiments on the web

Kohavi, Longbotham, Sommerfield, Henne
Key Ideas to take away from this paper

- A/B tests, Multivariate Tests, Control/Treatments tests, parallel flights.
- Practical Guide to conducting online experiments.
- Significant learning and increase in ROI when development teams listen to customers.
It’s All A/Bout Testing

1. Understand Your Data
2. Test an Alternative
3. Implement the Winner

Repeat
Thanks Obama!
<table>
<thead>
<tr>
<th>Variation</th>
<th>Est. conv. rate</th>
<th>Chance to Beat Orig.</th>
<th>Observed Improvement</th>
<th>Conv./Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Button</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>7.51% ± 0.2%</td>
<td>—</td>
<td>—</td>
<td>5851 / 77858</td>
</tr>
<tr>
<td>Learn More</td>
<td>8.91% ± 0.2%</td>
<td>100%</td>
<td>18.6%</td>
<td>6927 / 77729</td>
</tr>
<tr>
<td>Join Us Now</td>
<td>7.62% ± 0.2%</td>
<td>73.5%</td>
<td>1.37%</td>
<td>5915 / 77644</td>
</tr>
<tr>
<td>Sign Up Now</td>
<td>7.34% ± 0.2%</td>
<td>13.7%</td>
<td>-2.38%</td>
<td>5660 / 77151</td>
</tr>
<tr>
<td><strong>Media</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>8.54% ± 0.2%</td>
<td>—</td>
<td>—</td>
<td>4425 / 51794</td>
</tr>
<tr>
<td>Family Image</td>
<td>9.66% ± 0.2%</td>
<td>100%</td>
<td>13.1%</td>
<td>4996 / 51696</td>
</tr>
<tr>
<td>Change Image</td>
<td>8.87% ± 0.2%</td>
<td>92.2%</td>
<td>3.85%</td>
<td>4595 / 51790</td>
</tr>
<tr>
<td>Barack's Video</td>
<td>7.76% ± 0.2%</td>
<td>0.04%</td>
<td>-9.14%</td>
<td>3992 / 51427</td>
</tr>
<tr>
<td>Sam's Video</td>
<td>6.29% ± 0.2%</td>
<td>0.00%</td>
<td>-26.4%</td>
<td>3261 / 51864</td>
</tr>
<tr>
<td>Springfield Video</td>
<td>5.95% ± 0.2%</td>
<td>0.00%</td>
<td>-30.3%</td>
<td>3084 / 51811</td>
</tr>
<tr>
<td>Combination</td>
<td>Status</td>
<td>Est. conv. rate</td>
<td>Chance to Beat Orig.</td>
<td>Observed Improvement</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>----------------</td>
<td>----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Original</td>
<td>Enabled</td>
<td>8.26% ± 0.5%</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Combination 11</td>
<td>Enabled</td>
<td>11.6% ± 0.6%</td>
<td>100%</td>
<td>40.6%</td>
</tr>
<tr>
<td>Combination 7</td>
<td>Enabled</td>
<td>10.3% ± 0.6%</td>
<td>100%</td>
<td>24.0%</td>
</tr>
<tr>
<td>Combination 3</td>
<td>Enabled</td>
<td>9.80% ± 0.6%</td>
<td>99.7%</td>
<td>18.7%</td>
</tr>
<tr>
<td>Combination 10</td>
<td>Enabled</td>
<td>9.23% ± 0.6%</td>
<td>95.9%</td>
<td>11.7%</td>
</tr>
<tr>
<td>Combination 8</td>
<td>Enabled</td>
<td>9.03% ± 0.6%</td>
<td>91.6%</td>
<td>9.28%</td>
</tr>
<tr>
<td>Combination 9</td>
<td>Enabled</td>
<td>8.77% ± 0.6%</td>
<td>81.8%</td>
<td>6.10%</td>
</tr>
<tr>
<td>Combination 6</td>
<td>Enabled</td>
<td>8.64% ± 0.5%</td>
<td>75.3%</td>
<td>4.58%</td>
</tr>
</tbody>
</table>
Mailchimp’s A/B Test

What would you like to test?
Select the variable you want to test. We’ll generate a campaign for each combination of the variable, up to 3 combinations.

How would you like to test the combinations?
- Distribute the combinations across all recipients
- Test on a percentage of recipients to send a winner

The remaining 60% of your recipients will be sent the winning combination.

What metric should determine the winner?
- By Click Rate after 4 hours

We recommend waiting at least 4 hours to gather all the results.
Mailchimp’s A/B Test

Fall Shirts Are Here

’Tis the season at J.R. Florence for fitted long sleeve button downs! Our full Autumn Series of artisanal, handmade poly cotton blends is debuting in stores this weekend, but mailing list members get early online access.
Mailchimp’s A/B Test

Content setup
Test different messages or templates within your campaign.

1 Column - Banded Template
Edited on Aug 07, 2015 09:48 am
"Testing the order in which the content appears: main alpha image (red), headline,"

1 Column - Banded Template
Edited on Aug 07, 2015 09:37 am
"Testing the order in which the content appears: main alpha image (blue), headline,"

1 Column - Banded Template
Edited on Aug 07, 2015 09:53 am
"Testing the order in which the content appears: main alpha image (yellow), headline,"
Mailchimp’s A/B Test

Early Online Access List

**15,000 Recipients**

- **List:** Early Online Access
- **Tested on:** 40% of 37,500 subscribers
- **Varies tested:** Content
- **Subject:** Fall Shirts Are Here

**Winning combination**

- **|LNAME|, Fall Shirts Are Here**
  - **Sent:** Aug 06, 2015, 3:00 pm from Shelby (fredie@freddiesjokes.com)
  - “Testing the order in which the content appears: main alpha image (blue)”

<table>
<thead>
<tr>
<th>Overall open rate</th>
<th>42.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>List average</td>
<td>45.3%</td>
</tr>
<tr>
<td>Industry average</td>
<td>27.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overall click rate</th>
<th>21.3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>List average</td>
<td>20.4%</td>
</tr>
<tr>
<td>Industry average</td>
<td>13.7%</td>
</tr>
</tbody>
</table>

**Combination results**

- 5,000 Sends
- 45.7% Open rate
- 22.8% Click rate
Treatment Ramp Up

- Vary the amount of people you expose to different version of the product
- Very good approach to scaling your systems
- Automation and optimization of features is less costly
- Must be supported by good randomization algorithms
- No bias in any particular variant of the experiment
Auto abort

- Reduces the percentage of users assigned to the underperforming Treatment to zero.
- Test on more people
- Reduces risk of exposing more users to some error
- "Move fast and break things"
Ramp Up and Auto Abort

When running an A/B test, what fraction of users should be in the manipulation condition at the start? What fraction should you ramp up to? Why does the author recommend doing this?
“The increase ratio sounds rather arbitrary to me, but the benefit of ramping-up in such a fashion is that egregious errors can be caught early-on; because the sensitivity is squared in the formula for sample size, when it is relaxed to, say, supporting detecting a 20% change instead of 5% (thereby detecting bigger changes in behavior, which are most likely caused by egregious errors), the sample size needed decreases by a factor of $4^2 = 16$.”

- Hyeonsu
Discuss with your research partner!

How will you apply A/B testing to your research project? Discuss with your partner for 1 minute.
Academia vs Industry Perspectives
Experimental evidence of massive-scale emotional contagion through social networks

Kohavi, Longbotham, Sommerfield, Henne
There are criticisms of existing methods.

- Social Interaction vs Emotion
- Non Verbal vs Verbal cues
- Facebook undertook a massive study (~700,000 people) with Cornell University to answer some of these questions.

- They wanted to know if people's exposure to emotions on their news feed affected their emotion and thereby the content they posted. How did it affect it?
Two Parallel Experiments:

- Exposure to friends' positive emotional content was reduced.
- Exposure to friends' negative emotional content was reduced.
- There was also a control condition in which a similar proportion of posts in their news feed was omitted at random.
Posts were determined as positive or negative if they contained at least ONE positive or negative word as defined by the Linguistic Inquiry and Word Count (LIWC) software.
LIWC is a Natural language processing tool.
Sentiment analysis of a Twitter account using LIWC
Sentiment analysis of Scott's Twitter account using LIWC

Scott Klemmer
@DesignAtLarge
Associate Professor @DesignLabUCSD; Coursera #ixDOnline.
La Jolla, San Diego
d.ucsd.edu/srk
 Joined August 2013
They measured two variables:

- Percentage of all words produced by person that were positive

- Percentage of all words produced by person that were negative
Results

1) Positive Posts Reduced Condition:
   (a) % of positive words decreased by 0.1%
   (b) % of negative words increased by 0.04%

2) Negative Posts Reduced Condition:
   (a) % of negative words decreased by 0.07%
   (b) % of positive words increased by 0.06%

Results proved emotional contagion exists. Emotions expressed by our friends are transferred over to us over large scale social networks. Also proved that nonverbal behavior is not necessary for emotional contagion.
Turns out people didn't care about the results as much as the ethical dilemma surrounding the study. Discuss with your partner for 1 minute about how ethical/unethical this was. If it was unethical, what part of this process would you mitigate for it to be ethical?