A/B Testing

Crowdsourcing and Human Computation
Lecture 19
Instructor: Chris Callison-Burch
TA: Ellie Pavlick

Website: crowdsourcing-class.org

Active versus Passive Crowdsourcing

- So far we have mainly looked at *active* crowdsourcing, where we explicitly solicit help from the crowd
- Many applications of crowdsourcing rely on *passive* information collection from multitudes of individual

Example: Apple Maps

- iOS 7 allows users to “help improve maps” by enabling a feature called “frequent locations”
- Frequent locations gives Apple a method to verify business locations and other destinations by tracking user movements in the aggregate
- Participation also transmits drive and other travel time data to Apple

A/B Testing

- A/B Split Testing is a mechanism for passive crowdsourcing that allows web developers to empirically optimize the design of their sites
- Splits web users into two groups and shows them slightly different versions of the site
- Measures the behavior of the groups in aggregate and calculates whether one design leads to a better measurable outcome

A/B Testing

- A/B Split Testing is a mechanism for passive crowdsourcing that allows web developers to empirically optimize the design of their sites
- Splits web users into two groups and shows them slightly different versions of the site
- Measures the behavior of the groups in aggregate and calculates whether one design leads to a better measurable outcome

Why A/B tests?

- Lets us evaluate the goodness of alternate designs, instead of relying on our intuitions
- A typical web site may convert only 2% of its visitors into customers
- Small changes can have a big impact
- Google uses A/B testing all the time, and makes it available through Google Analytics

What sorts of things can you optimize with A/B tests?

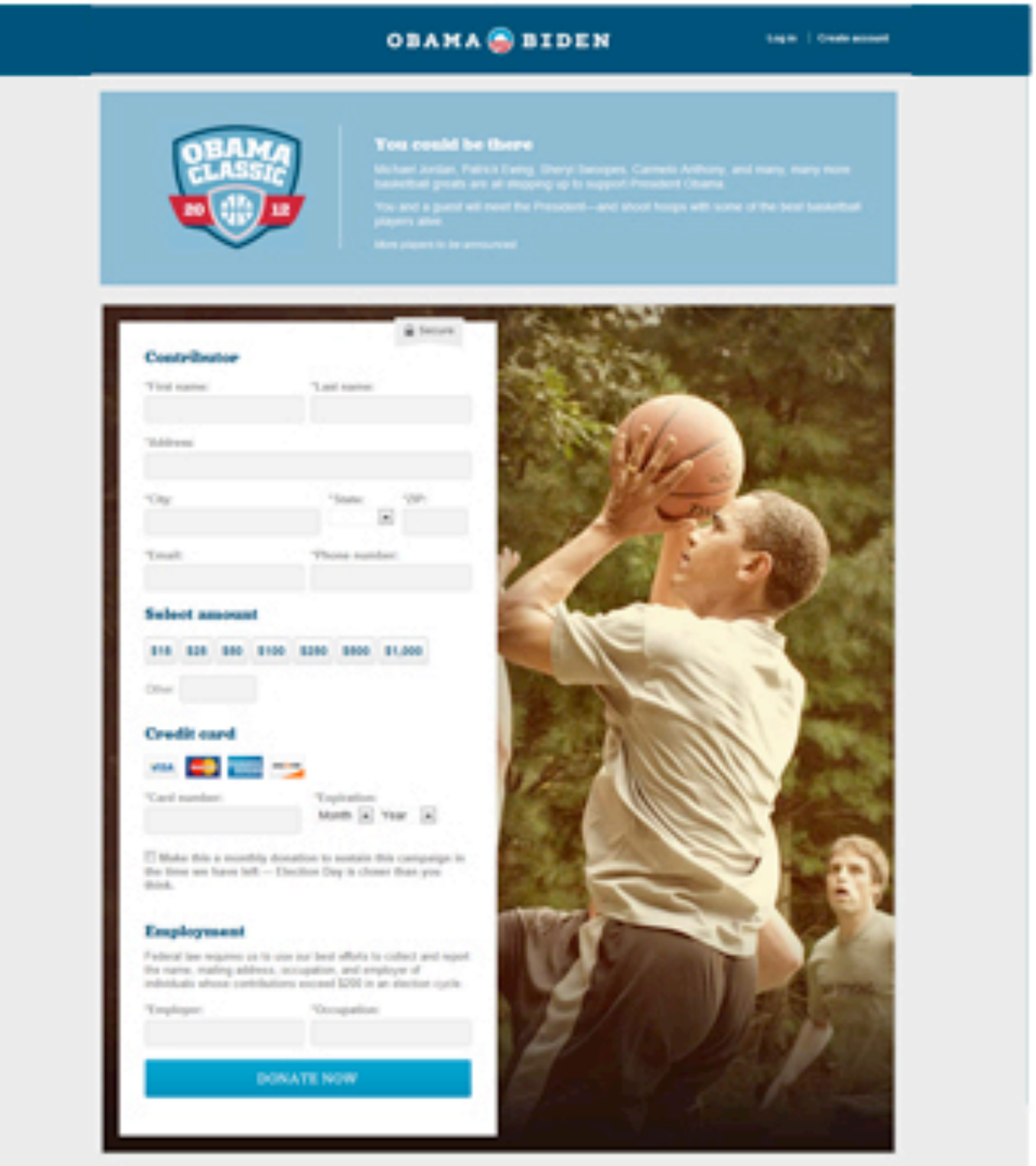
- Whether changing the order of collecting form information gets users to stick through to the end
- Whether changing the copywriting on your page improves things
- Whether different images are better at motivating web site visitors to do something that you want them to

What outcomes could you measure?

A/B testing was used to optimize the Obama Campaign

- Kyle Rush was the deputy director of frontend web development at Obama for America
- Managed online fundraising totaling $690 million in 20 months
- Conducted 500+ A/B tests, which increased the donation conversion rate by 49% and the email acquisition conversion rate by 161%

CONTROL



"SEQUENTIAL"

+5%

http://by.lerush.

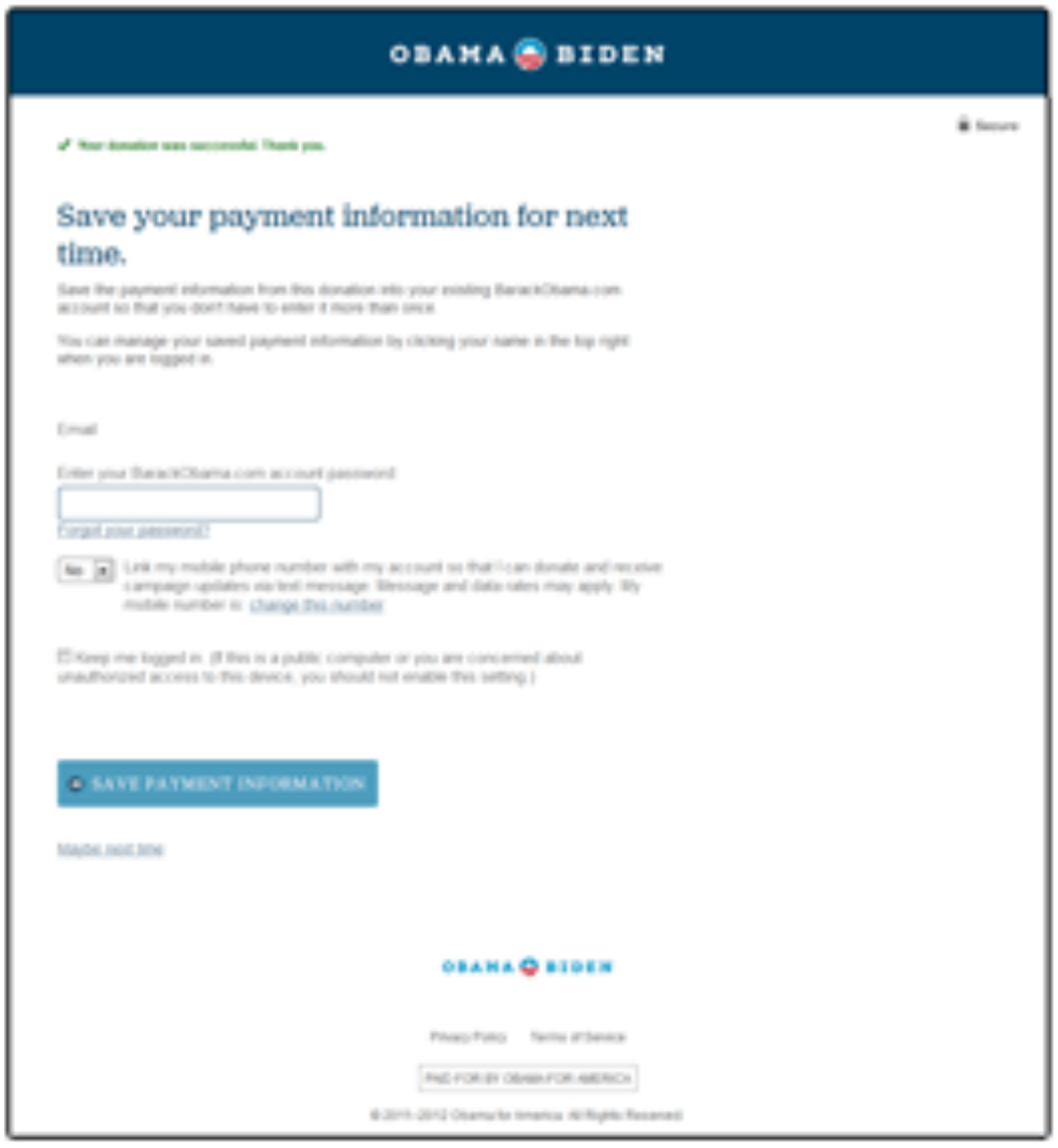
CONTROL

IMAGE VARIATION

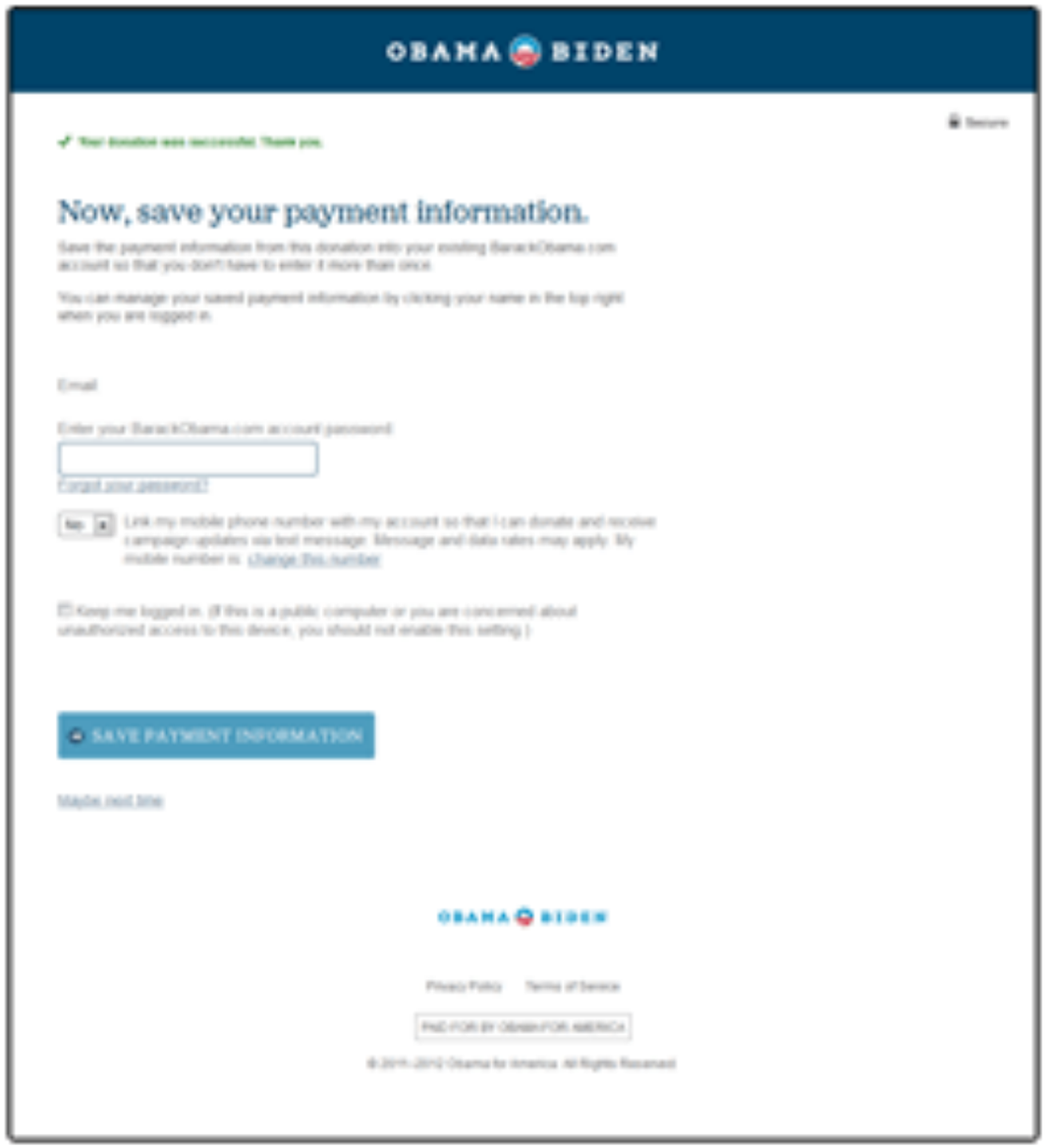
+19%

http://hyperurl.net

CONTROL



SEGUE COPY



+21%

http://ly lerus

Statistics behind A/B testing

- One commonly used method is the G-test
- Invented by Karl Pearson in 1900
- Provides a method for comparing 2 data sets, or answering a yes/no question
- Good tutorial by Ben Tilly on “Effective A/B Testing”

What to measure?

- Start your A/B test
- Divide your users into groups A and B
- Decide whether each person did what you wanted
- Reduce your results into 4 numbers: $a_yes, $a_no, $b_yes, $b_no

Counts of outcomes

	Yes	**No**	
A	**$a_yes**	**$a_no**	$a
B	**$b_yes**	**$b_no**	$b
	$yes	$no	$total

Totals

	Yes	**No**	
A	$a_yes	$a_no	**$a**
B	$b_yes	$b_no	**$b**
	$yes	**$no**	**$total**

Expectations

	Yes	**No**	
A	**$e_a_yes**	**$e_a_no**	$a
B	**$e_b_yes**	**$e_b_no**	$b
	$yes	$no	$total

$e_a_yes = $a * $yes / $total

$e_a_no = $a * $no / $total

$e_b_yes = $b * $yes / $total

$e_b_no = $b * $no / $total

G-test

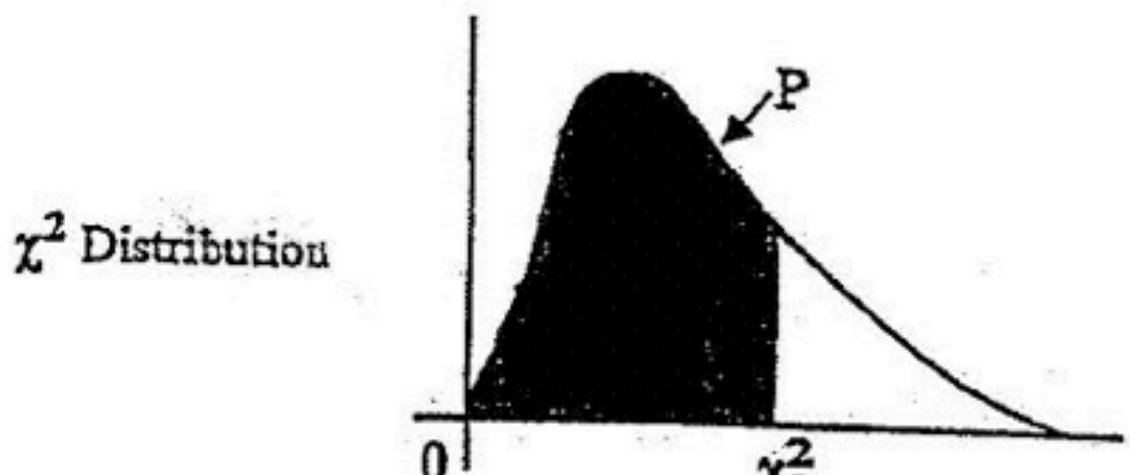
observations

	Yes	**No**
A	**$a_yes**	**$a_no**
B	**$b_yes**	**$b_no**

expectations

Yes	**No**
$e_a_yes	**$e_a_no**
$e_b_yes	**$e_b_no**

$G_test = 2 * ($a_yes * ln($a_yes / $e_a_yes)
+ $a_no * ln($a_no / $e_a_no)
+ $b_yes * ln($b_yes / $b_a_yes)
+ $b_no * ln($b_no / $b_a_no)



The table below gives the value x_0^2 for which $P[x^2 < x_0^2] =$ P for a given number of degrees of freedom and a given value of P.

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	

G-test Calculator

How many versions? 2

Success is measured by

- (•) count of successes
- ( ) percentage of trials that succeeded

Use the Yates' continuity correction? ☑

Test	Trials	Success
A	1000	100
B	500	25

Calculate

The G-test statistic is 11.0928 so version 'A' wins with 99.91% confidence.

http://elem.com/~btilly/effective-ab-testing/g-test-calculator.html

Optimizely

- http://www.optimizely.com
- 4 minute video

A/B Split Testing Protocol

- Identify your initial control web page – this could your current landing page or whatever you want to optimize
- Establish your goals – what is the thing that you want to optimize? Number of people signing up for your service? Revenue generated by a particular ad campaign?

A/B Split Testing Protocol

- Determine how long you need to run the experiment – this depends on how much traffic your web site gets, and what level of statistical significance you want
- Create 1 to 3 significant re-designs – your designers can propose a bunch of different overhauls, use the initial phase to hone in on the best high-level re-design

A/B Split Testing Protocol

- Use A/B testing to choose among the different re-designs. Ideally you can test every pages against every other one, but if that is impractical, you can do a tournament
- Based on the results, choose your true control page – this initial pick will likely generate the lion's share of the improvements

A/B Split Testing Protocol

- Finally, optimize the nitty-gritty elements of the web page using A/B testing
  - Headline
  - Call to Action
  - Page Copy
  - Graphics
  - Color
  - Configuration of Page Elements
  - Etc.

Upcoming Guest Lectures

- Wednesday: Ellie Pavlick will give a lecture on prediction markets
- Next Wednesday: Scott Novotney (JHU) will present his research into crowdsourcing for automatic speech recognition systems