Human-Computer Interaction Design

COGS120/CSE170 - “Intro. HCI”

Instructor: Philip Guo

Week 7 - Running Experiments (2016-11-08)

some slides adapted from Scott Klemmer’s Intro. HCI course

Learning Objective

to measure the usability of your app by planning, running, and analyzing data from experiments

Outline

- User testing vs. controlled experiments
- Planning and running experiments
- Analyzing data from experiments: chi-squared test

So far in this class we've done a lot of design and a bit of engineering ...

Now it's time to do some ***SCIENCE!!!***

But first let’s talk briefly about user testing …

User testing = getting users to use and test your app (yep!)

User testing consists of …

PLANNING: develop a written test protocol ("script") for consistency, *pilot* (practice) your protocol on friends to fix obvious bugs so that you don't waste time with real users.

RUNNING*:** get informed consent (verbal/written), have one person be facilitator and rest of team be observers. tell the tester ***"we are testing our app, not your skills! any mistakes are our fault." maybe encourage tester to "think aloud" (but may slow them down or alter their behavior).

ANALYZING collected data (e.g., written notes, audio, video, usage logs) to find UI failures and ways to improve

User testing is vital, but sometimes we need to be more formal … we may instead want to run *controlled experiments*.

User testing: “Let’s find people to use our app and we’ll hopefully get some feedback about how to improve it.”

Controlled experiment: “We want to see whether users of our app do task X *faster/more often/with fewer errors/etc.* than users of our competitor’s app.”

So far in this class we've done a lot of design and a bit of engineering ...

Now it's time to do some ***SCIENCE!!!***

How to do science in HCI/Design

Come up with a *hypothesis* related to some part of your app, design an experiment to *test* that hypothesis, then collect and analyze data to *statistically argue* whether your hypothesis is likely to be true.

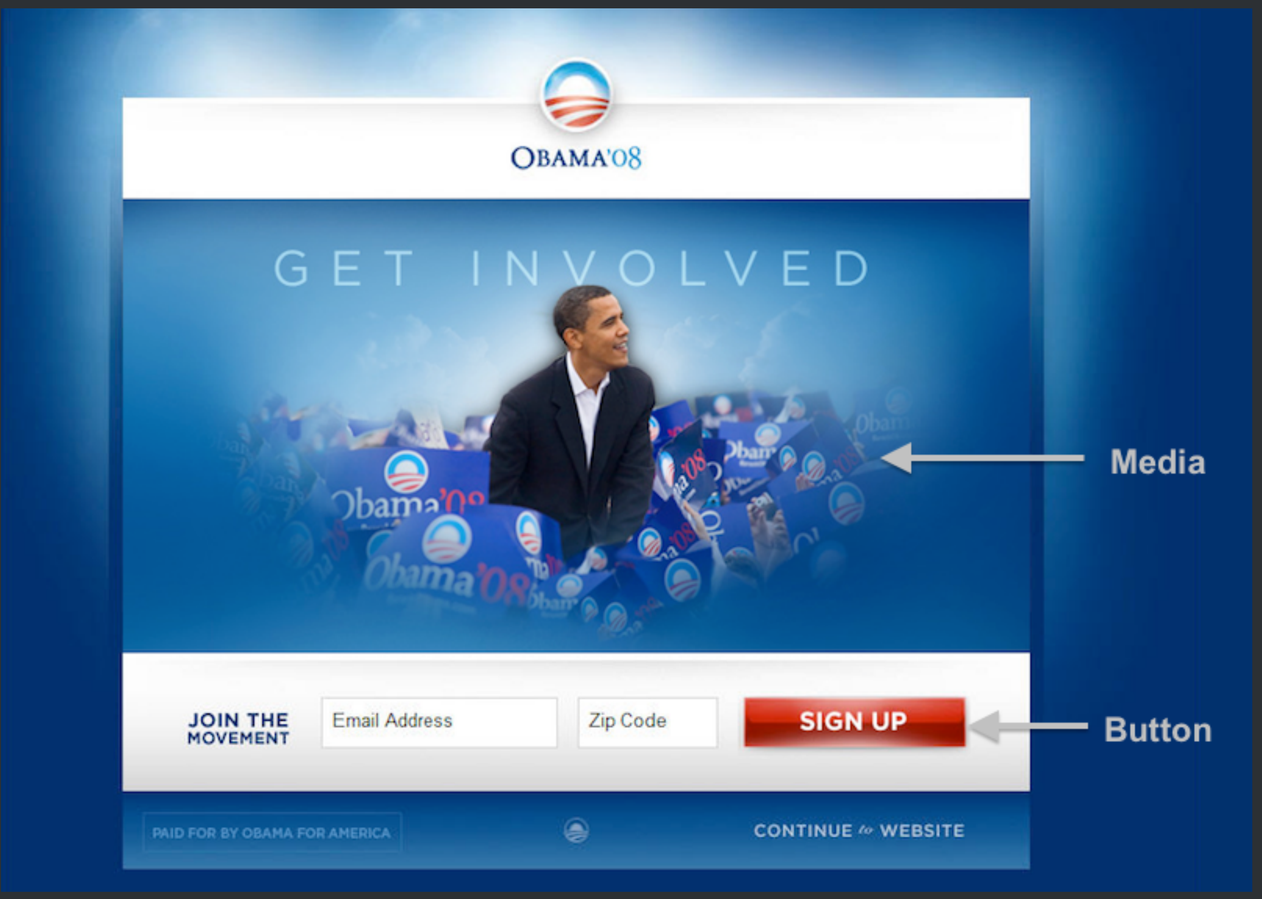
Is this “real science”? Absolutely!!!

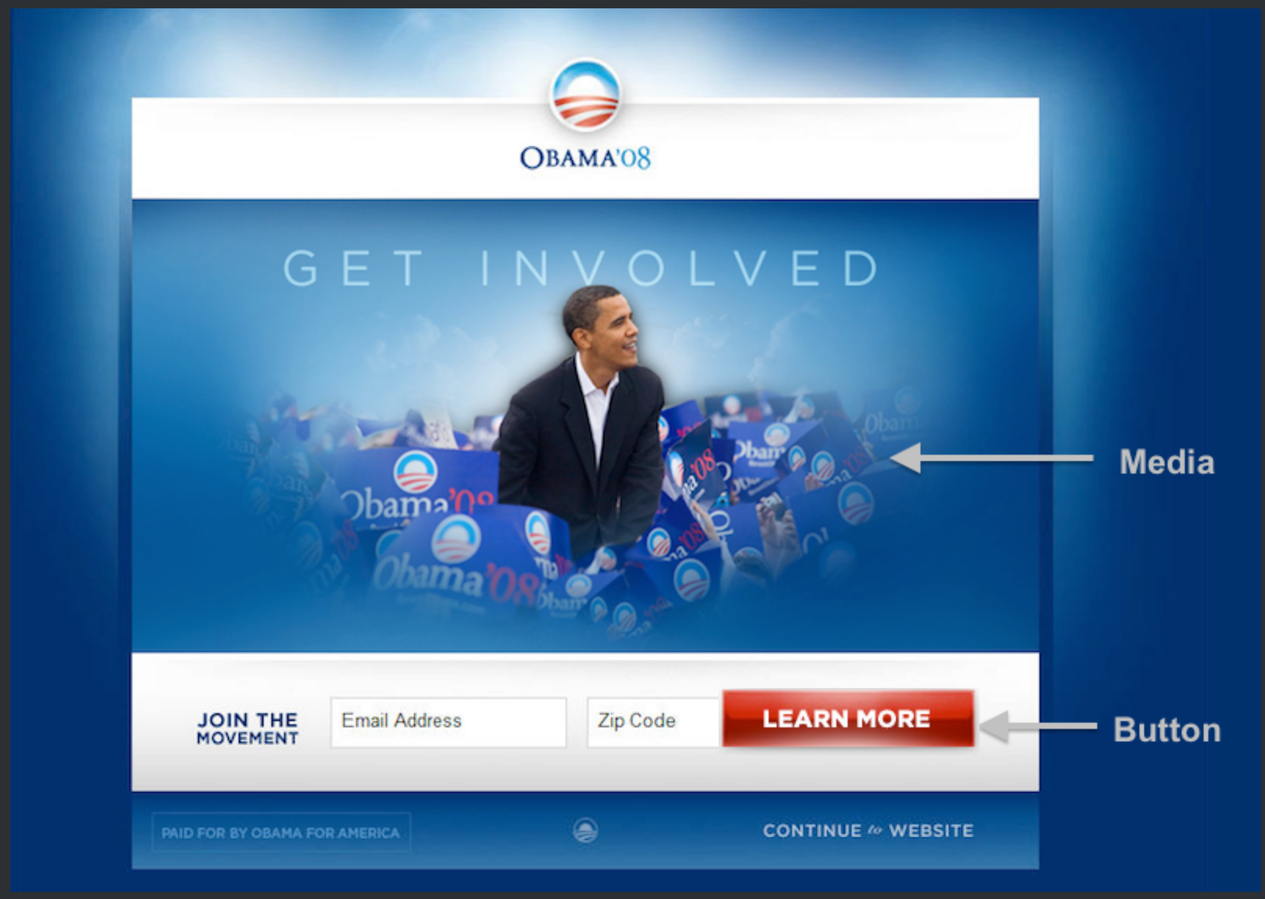
Come up with a *hypothesis* related to some _____, design an experiment to *test* that hypothesis, then collect and analyze data to *statistically argue* whether your hypothesis is likely to be true.

User testing (anecdotal, observation-driven): "Let's find people to use our app and we'll hopefully get some feedback about how to improve it."

Controlled experiment (scientific, hypothesis-driven): "We want to see whether users of our app do task X *faster/more often/with fewer errors/etc*. than users of our competitor's app."

Let’s run through a simple yet realistic experiment related to web design …



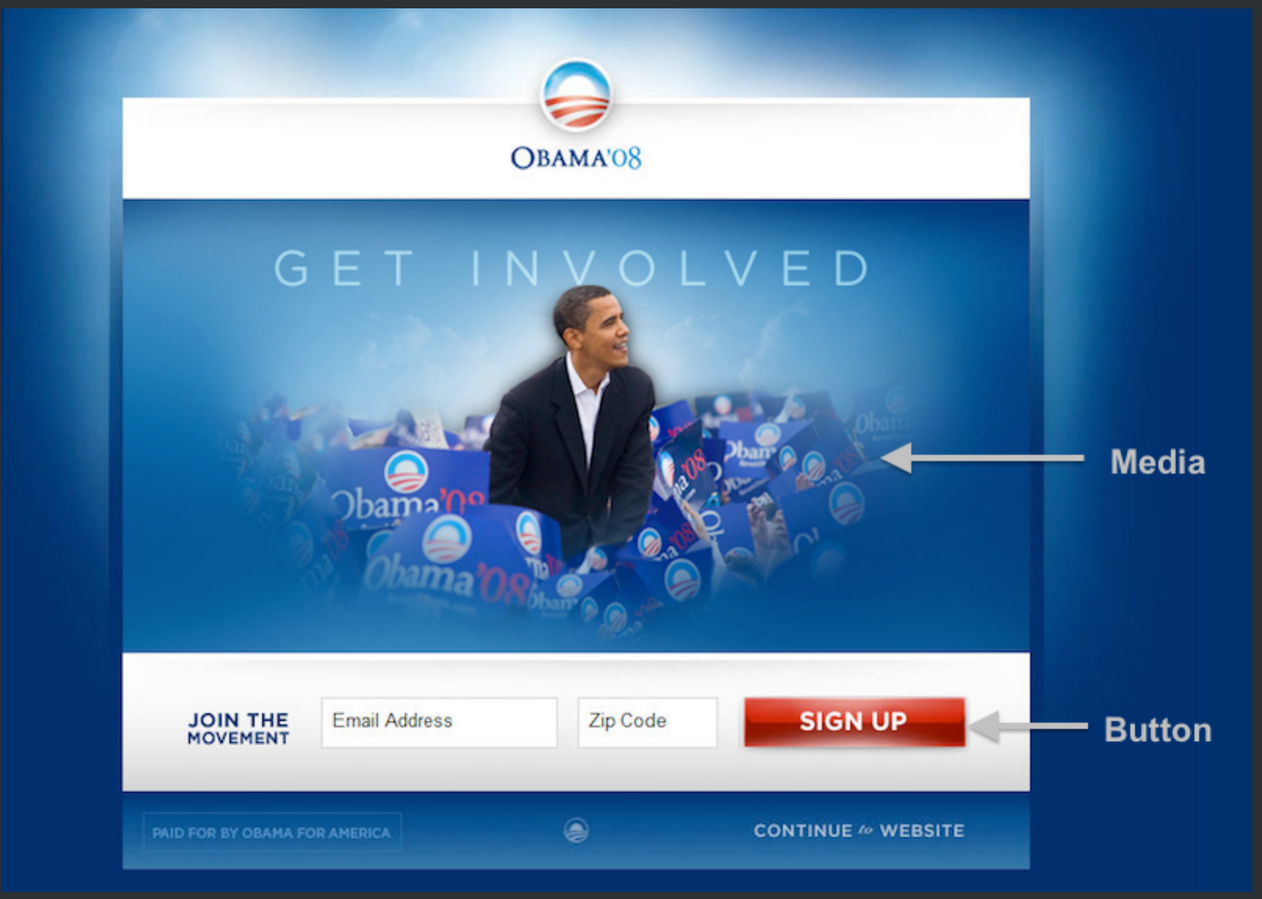


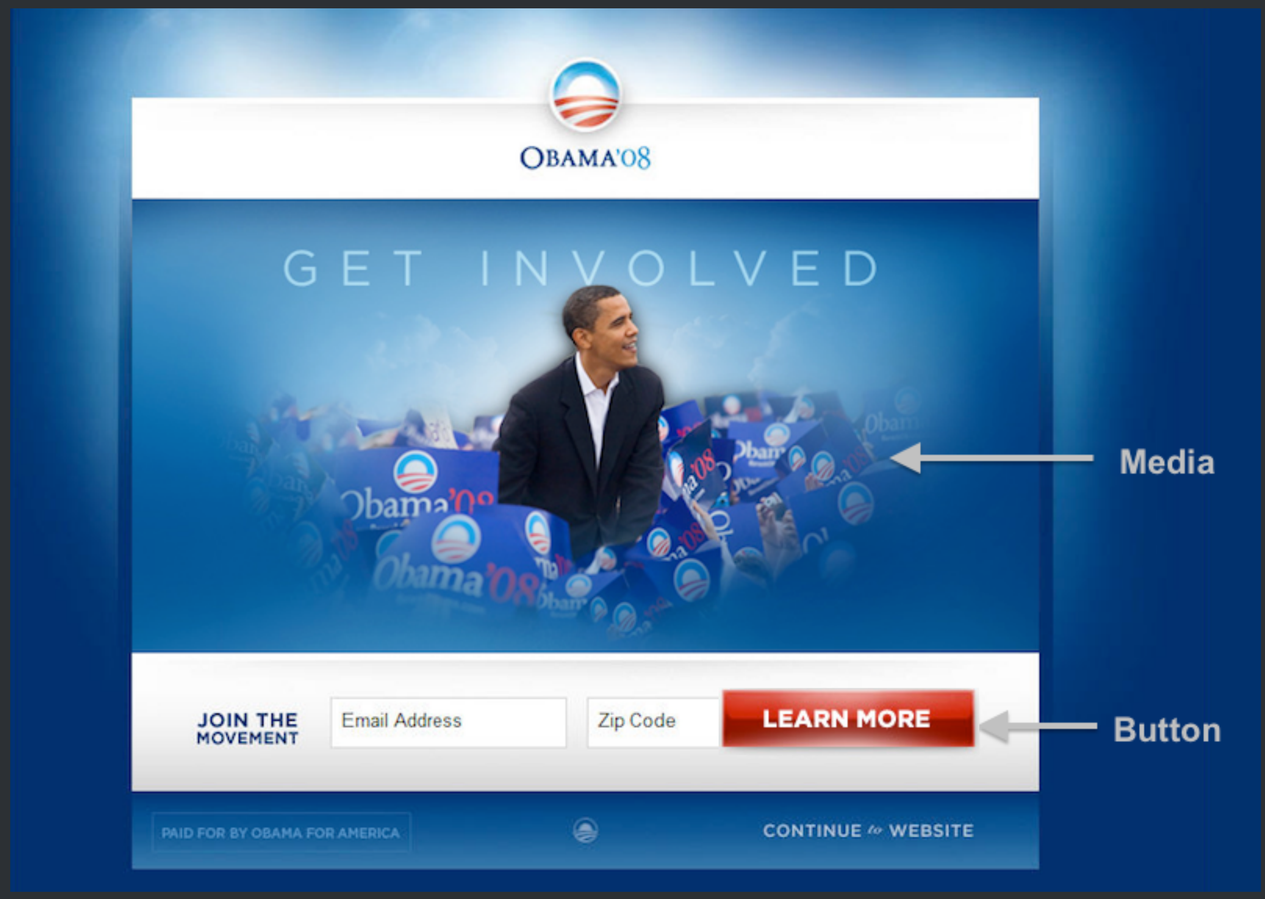
Source: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

Hypothesis

The “Learn More” button will lead to significantly more people signing up to donate to Obama’s campaign versus the original “Sign Up” button.

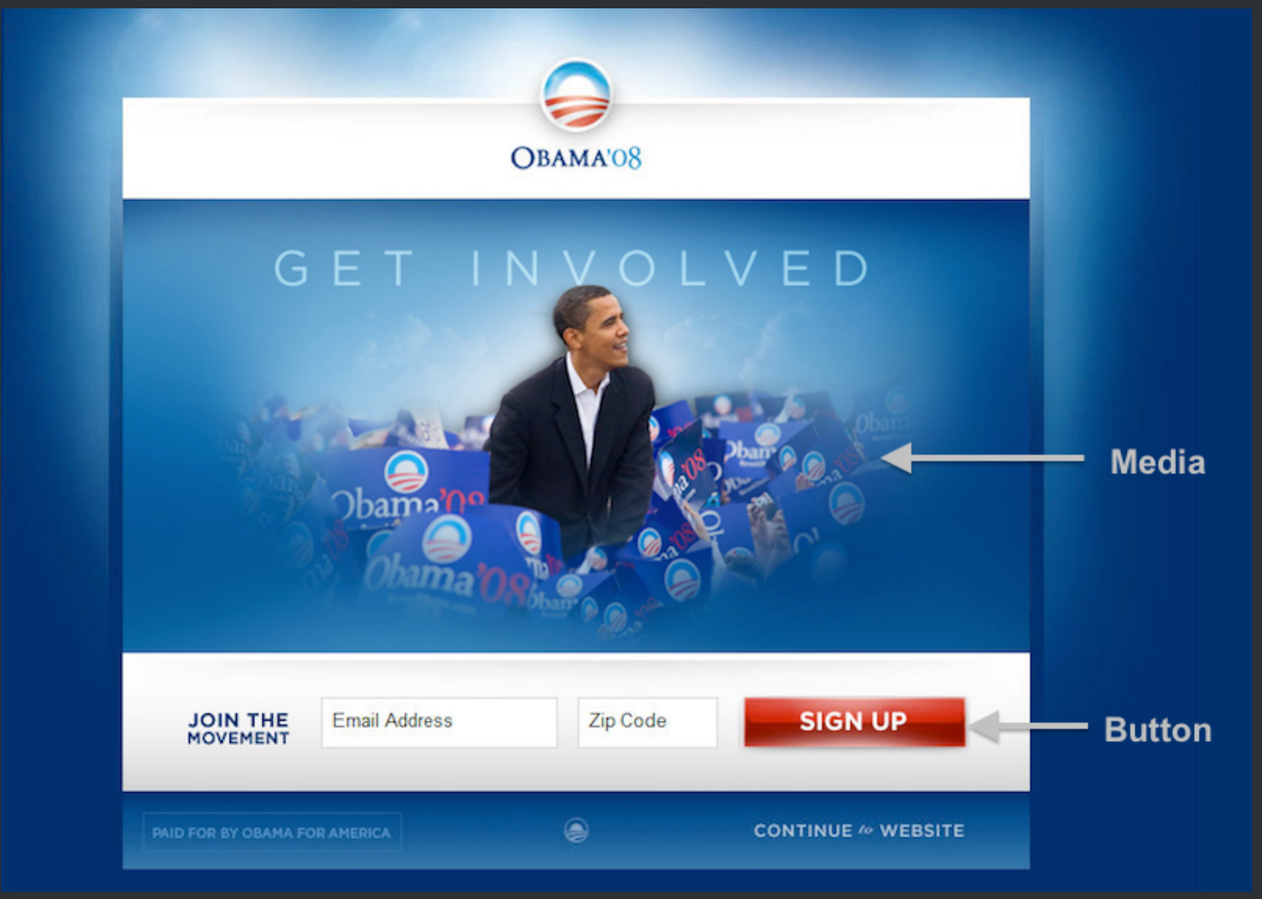
Experiment design: online A/B test … *randomly* show each visitor one of two versions of the home page:

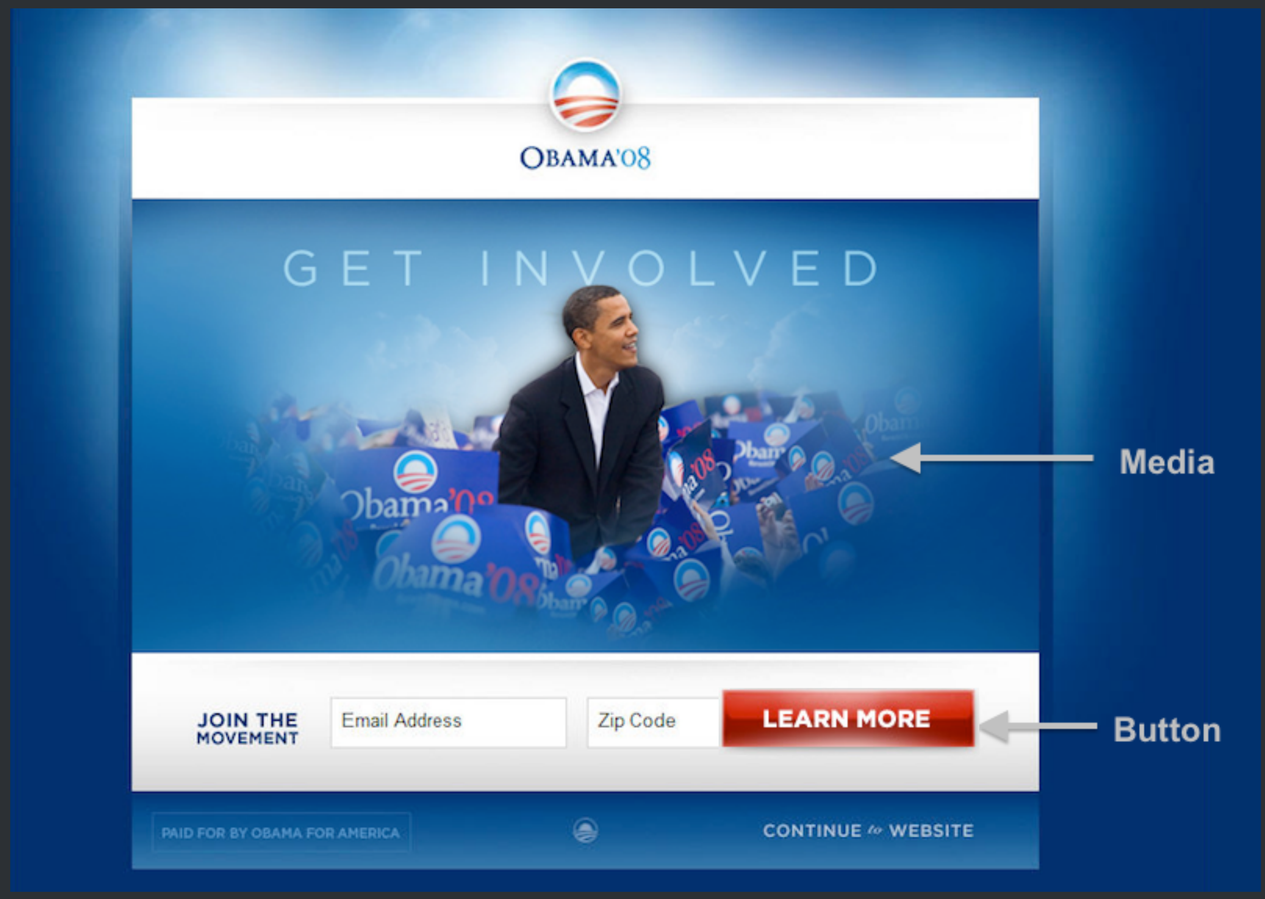




Source: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

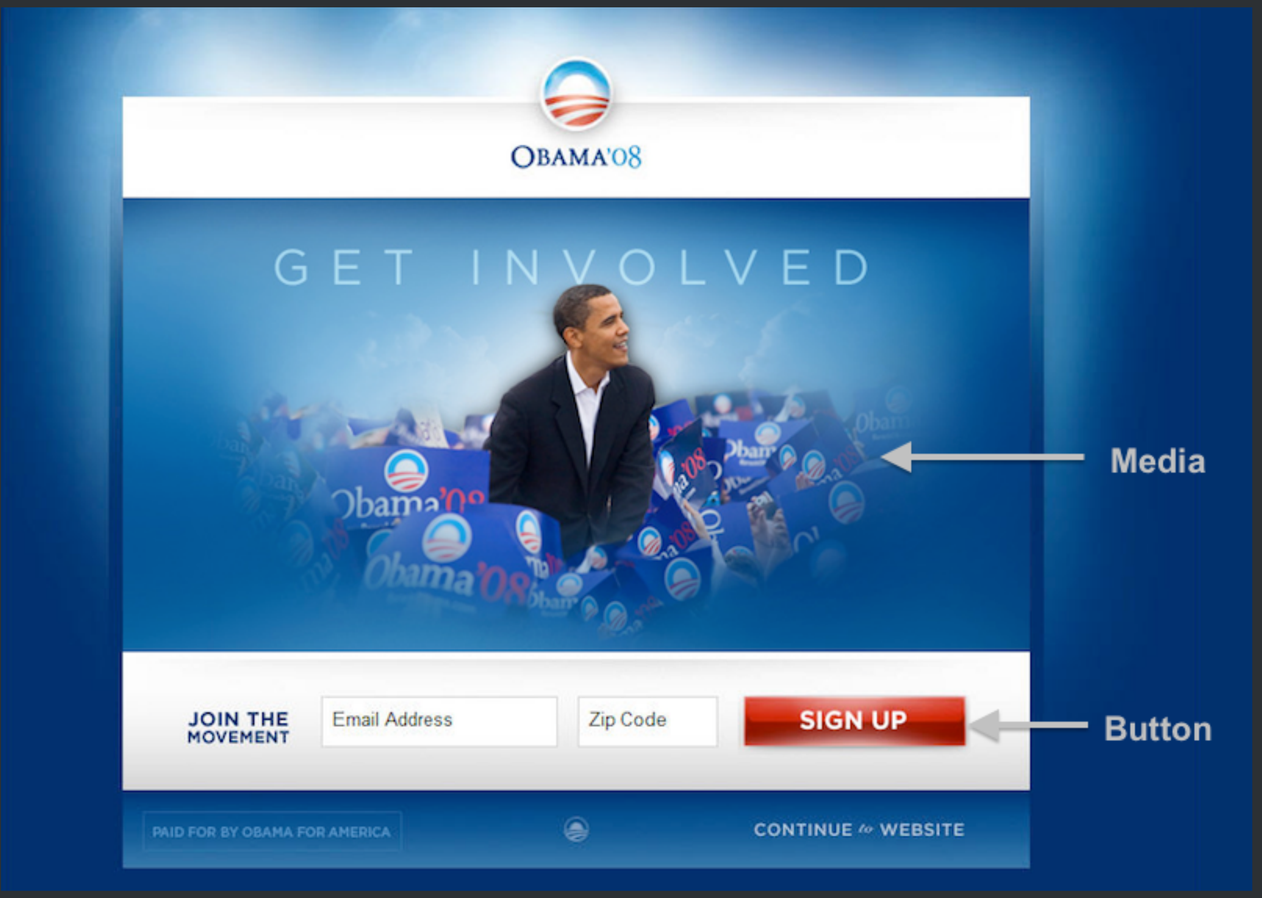
What are we measuring? *The sign-up rate.*
How many people actually sign up on the site.

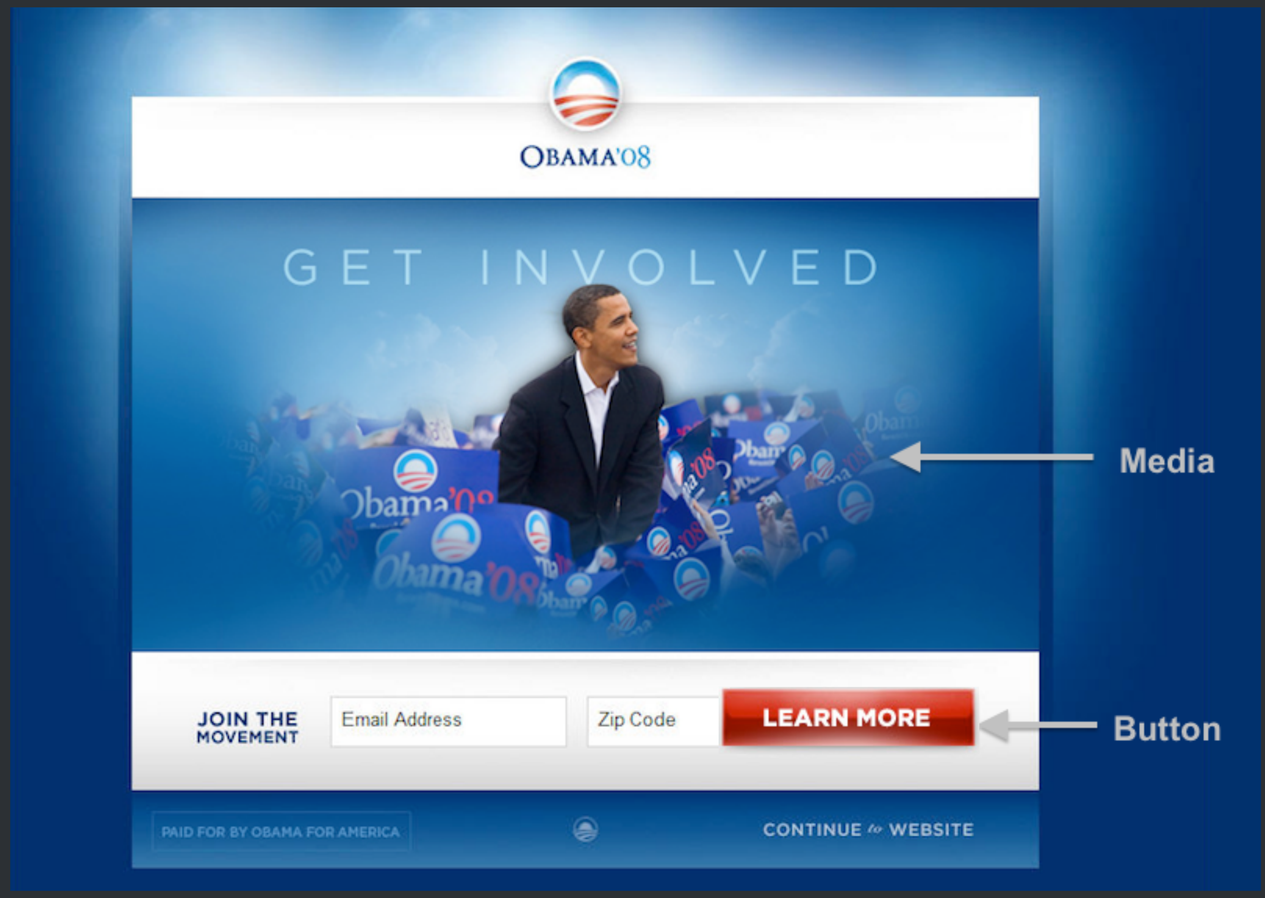




Source: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

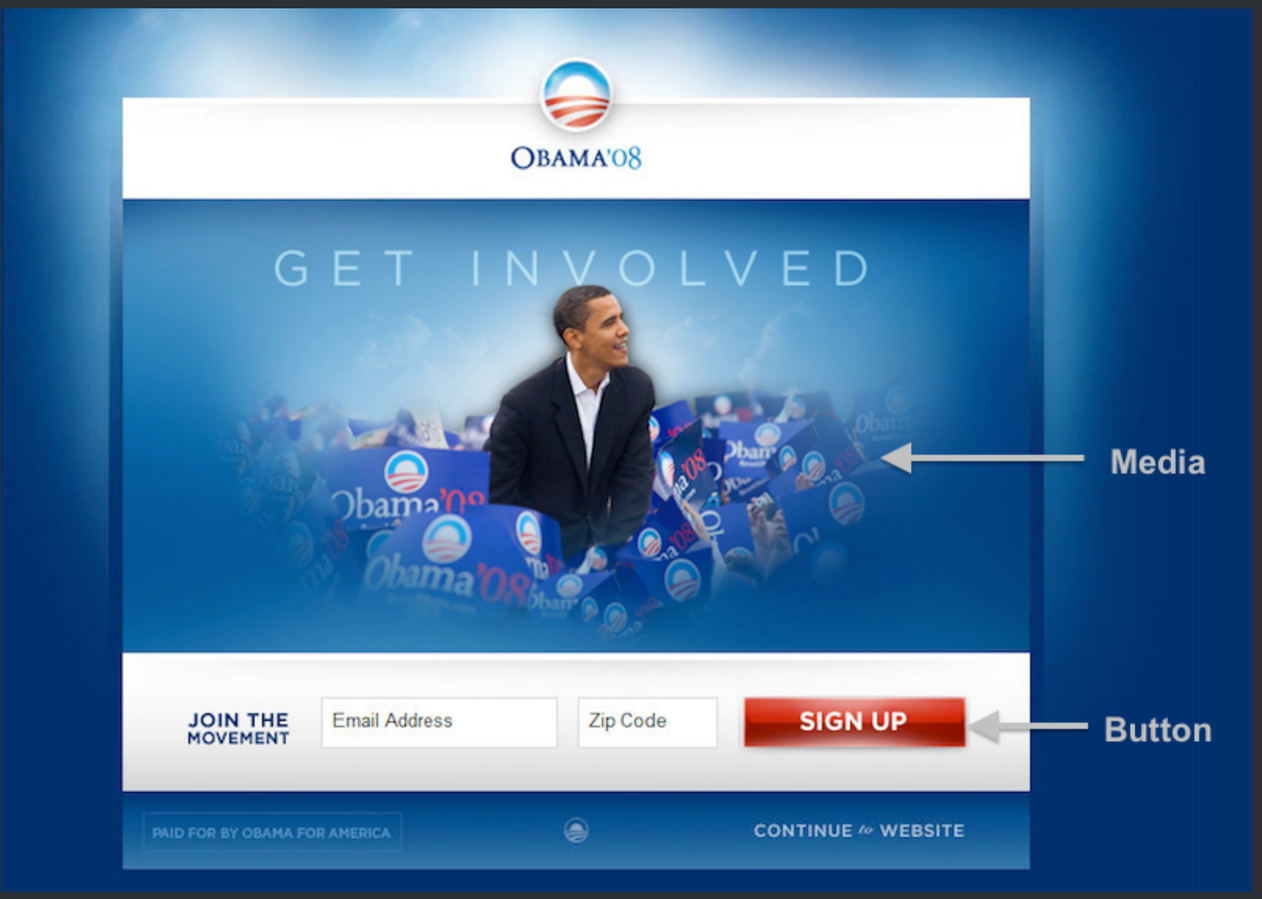
Why is randomness crucial? To eliminate selection bias.
(e.g., bad if you showed one version only to devout supporters at an Obama rally since they're more likely to sign up anyways)

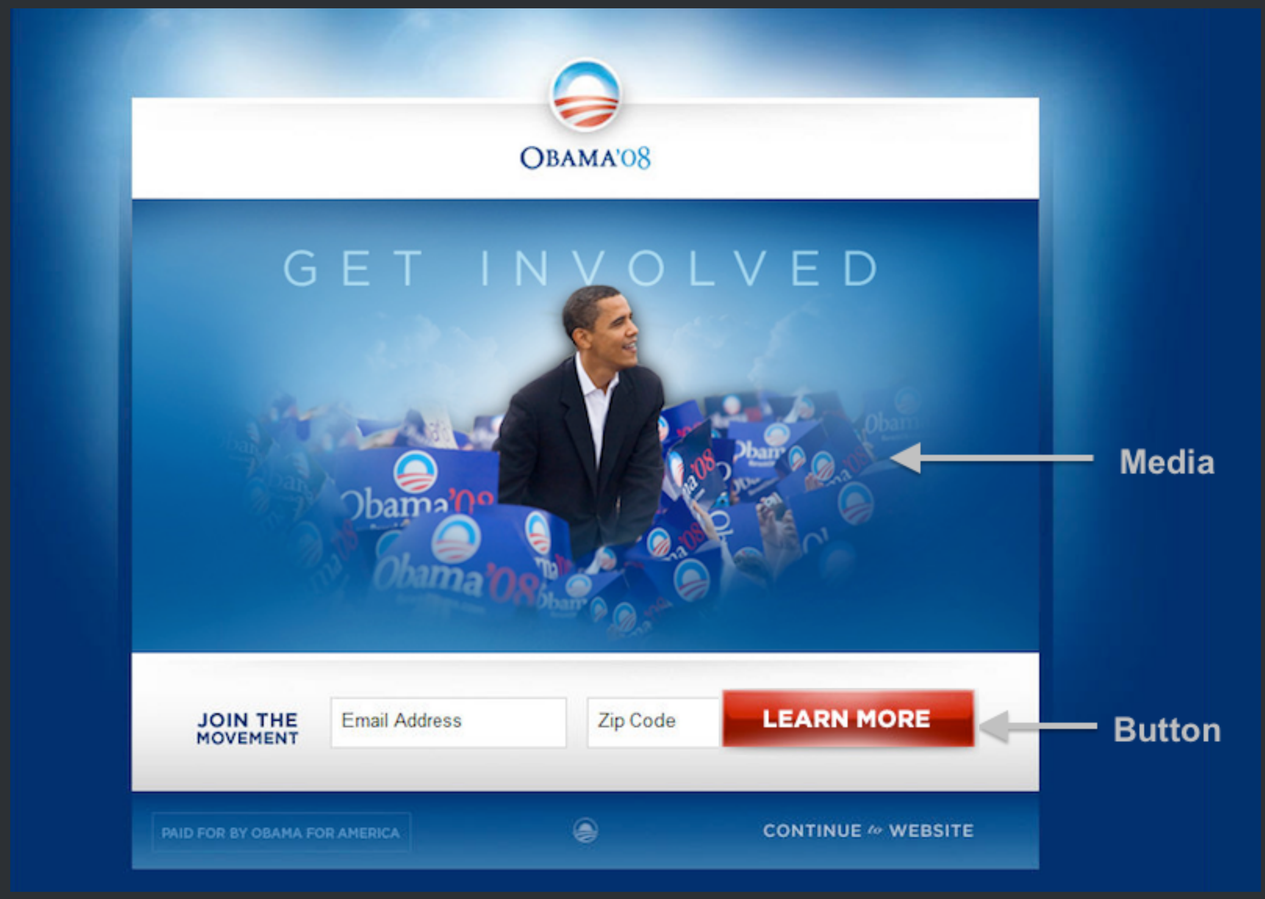




Source: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

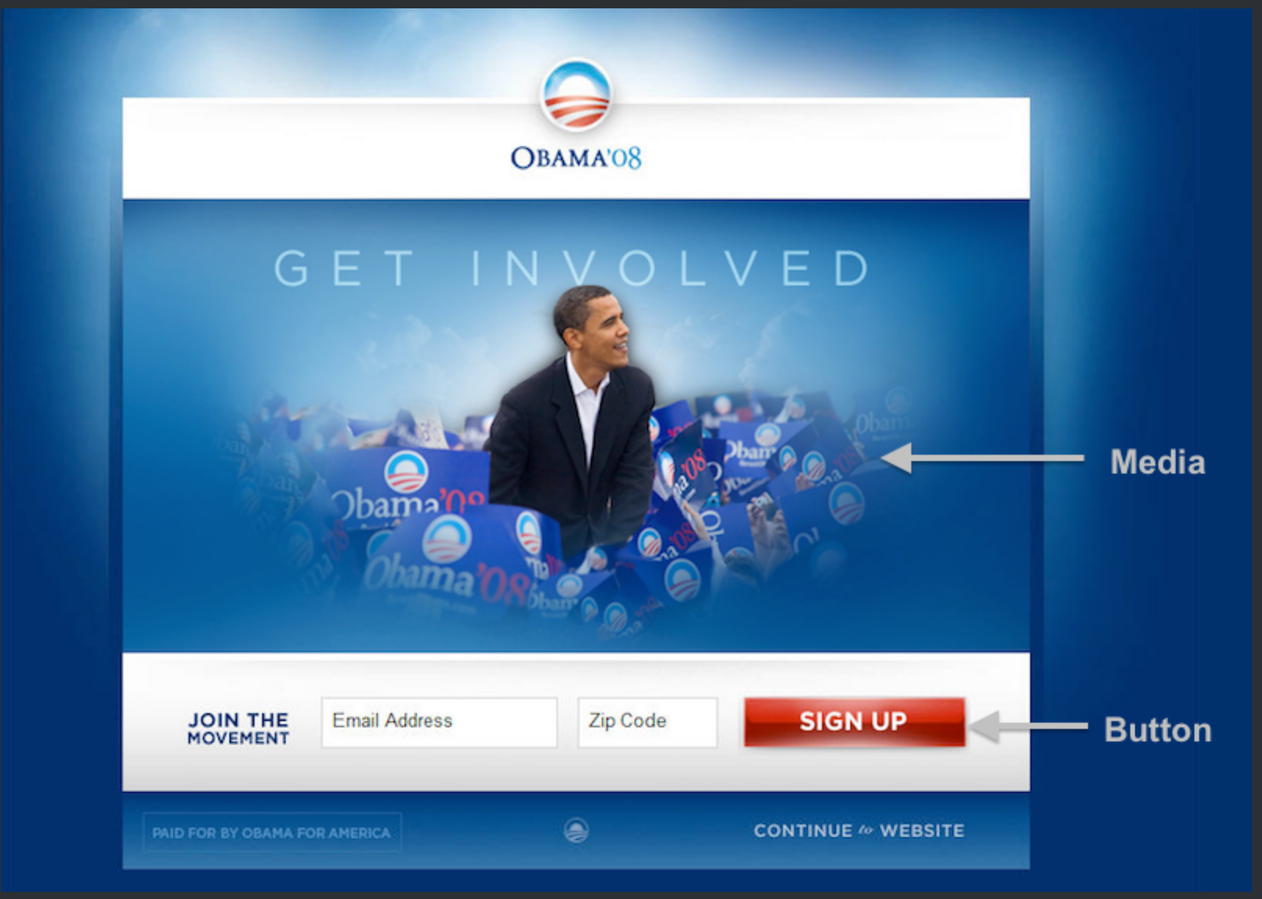
Deploying the experiment: write code to randomly fetch one home page design and to log user sign-up events (you will be using Google Analytics for this class)

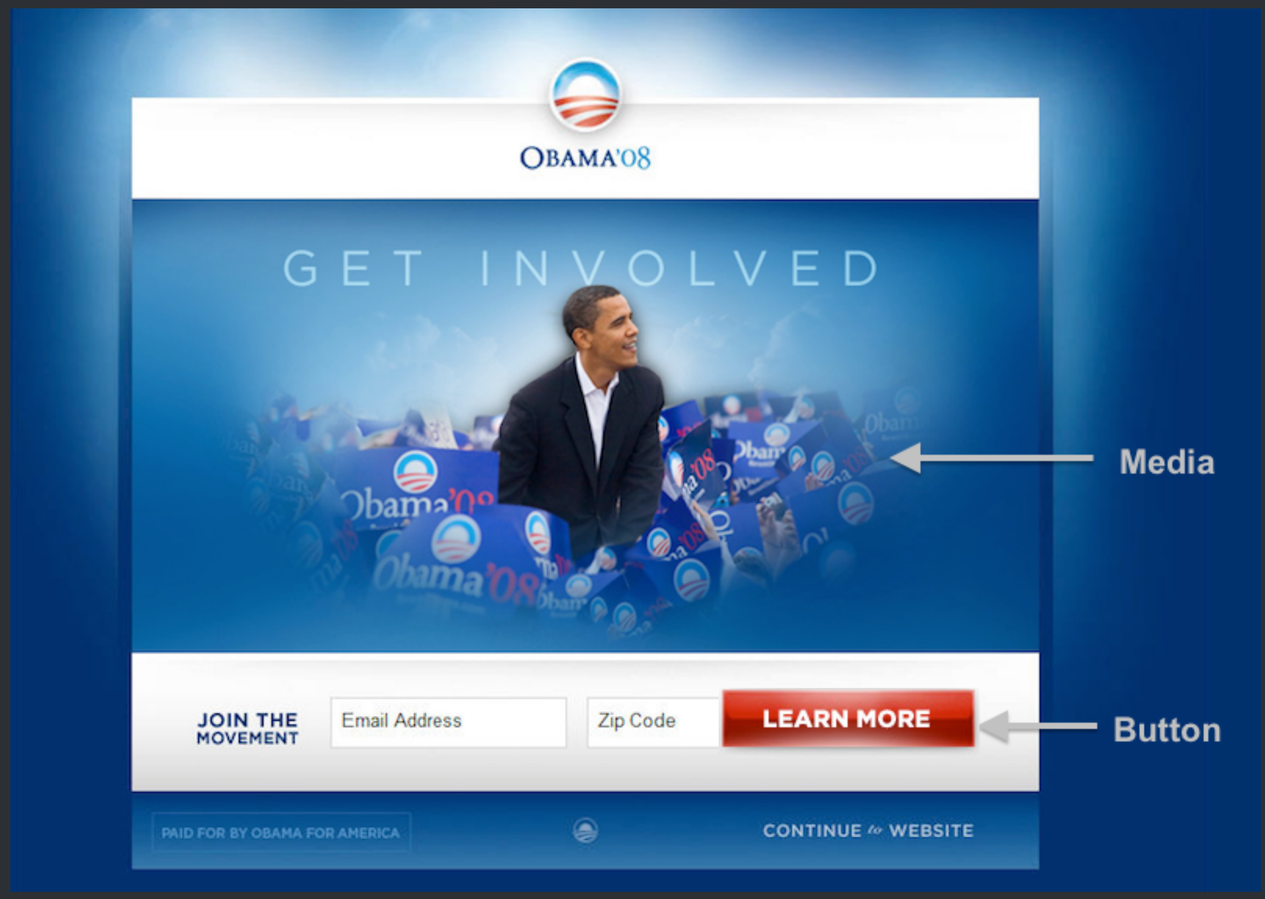




Source: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

Analyzing collected data: use *chi-squared test* to make a statistical argument





Source: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

Analyzing collected data: use *chi-squared test* to make a statistical argument

	Button on home page	
	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn't sign up	75	100

This data is made-up and matches the example from this lecture video, so follow along there: https://www.coursera.org/learn/design-principles/lecture/tOvhD

Analyzing collected data: use *chi-squared test* to make a statistical argument

	Button on home page	
	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn't sign up	75	100
Total visitors	100	120

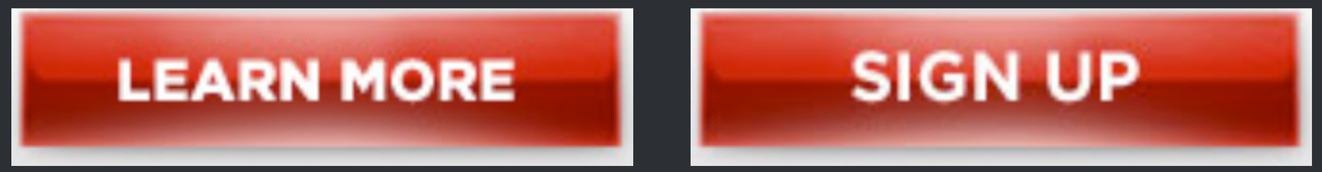
This data is made-up and matches the example from this lecture video, so follow along there: https://www.coursera.org/learn/design-principles/lecture/tOvhD

Analyzing collected data: use *chi-squared test* to make a statistical argument

	Button on home page	
	LEARN MORE	SIGN UP
Visitor did sign up	25 (**25%**)	20 (**17%**)
Didn't sign up	75 (75%)	100 (83%)
Total visitors	100	120

Analyzing collected data: use *chi-squared test* to make a statistical argument

Button on home page



Visitor did sign up	25 (**25%**)	20 (**17%**)

25% sign-ups is higher than 17%, so we're done, right?!? "Learn More" is clearly better!

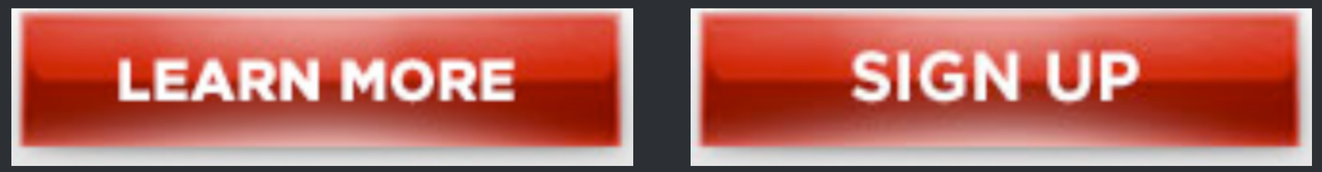
Analyzing collected data: use *chi-squared test* to make a statistical argument

	Button on home page	
	LEARN MORE	SIGN UP
Visitor did sign up	25 (**25%**)	20 (**17%**)

Not so fast! What if this happened just by chance since we had so few visitors?

Analyzing collected data: use *chi-squared test* to make a statistical argument

Button on home page



Visitor did sign up	25 (**25%**)	20 (**17%**)

The *chi-squared test* will tell us whether this particular 25% vs. 17% sign-up rate difference is *statistically significant*

Analyzing collected data: use *chi-squared test* to make a statistical argument

	Button on home page	
	LEARN MORE	SIGN UP
Visitor did sign up	25 (**25%**)	20 (**17%**)

In general, the *chi-squared test* is used to compare two (or more) sets of rates ("% occurrences") to tell whether the percentage differences are significant.

Hypothesis

The “Learn More” button will lead to significantly more people signing up to donate to Obama’s campaign versus the original “Sign Up” button.

Null Hypothesis

The “Learn More” button will lead to no significant change in the number of people signing up to donate to Obama’s campaign versus the original “Sign Up” button.

Statistical tests usually provide the likelihood that *null hypothesis* is true

The “Learn More” button will lead to no significant change in the number of people signing up to donate to Obama’s campaign versus the original “Sign Up” button.

Button on home page

	LEARN MORE	SIGN UP	Totals
Visitor did sign up	25	20	45
Didn’t sign up	75	100	175
Total visitors	100	120	220

Observed in experiment

	LEARN MORE	SIGN UP	Totals
Visitor did sign up	25	20	45
Didn't sign up	75	100	175
Total visitors	100	120	220

Expected if null hypothesis is true

	LEARN MORE	SIGN UP	Totals
Visitor did sign up	(45/220)*100 =**20.5**	(45/220)*120 =**24.5**	45
Didn't sign up	(175/220)*100 =**79.5**	(175/220)*120 =**95.5**	175
Total visitors	100	120	220

Observed in experiment

	LEARN MORE	SIGN UP	Totals
Visitor did sign up	25	20	45
Didn't sign up	75	100	175
Total visitors	100	120	220

Expected if null hypothesis is true

	LEARN MORE	SIGN UP	Totals
Visitor did sign up	20.5	24.5	45
Didn't sign up	79.5	95.5	175
Total visitors	100	120	220

Observed in experiment

	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn't sign up	75	100

Expected if null hypothesis is true

	LEARN MORE	SIGN UP
Visitor did sign up	20.5	24.5
Didn't sign up	79.5	95.5

To calculate the *chi-squared value, called* χ^2, sum up all pairs of:

$(\text{observed} - \text{expected})^2 / \text{expected}$

Observed in experiment

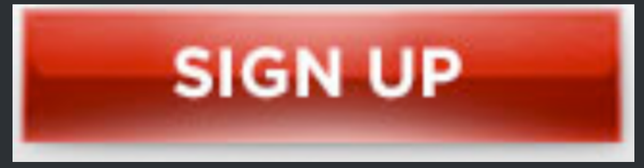
	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn't sign up	75	100

Expected if null hypothesis is true

	LEARN MORE	SIGN UP
Visitor did sign up	20.5	24.5
Didn't sign up	79.5	95.5

$$(25-20.5)^2 / 20.5 + (20-24.5)^2 / 24.5 + (75-79.5)^2 / 79.5 + (100-95.5)^2 / 95.5 = \mathbf{2.28}$$

Now we have chi-squared value = 2.28. We need one more magic number, called *degrees of freedom,* which represents how many entries in table need to be filled before all other entries are known. In this case, it's only 1 entry since we have a 2x2 table, so only 1 entry is needed to fill the table ...

	Button on home page		
			Totals
Visitor did sign up	25	20	45
Didn't sign up	75	100	175
Total visitors	100	120	220

Now we have chi-squared value = 2.28 and degrees of freedom df=1. Look up probabilities in a table:

https://people.richland.edu/james/lecture/m170/tbl-chi.html

Table: Chi-Square Probabilities

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860

$p \approx 0.10$, so there is a reasonable chance that the null hypothesis is true. We usually reject the null hypothesis at $p < 0.05$ or even $p < 0.01$.

Observed in experiment

	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn't sign up	75	100

Expected if null hypothesis is true

	LEARN MORE	SIGN UP
Visitor did sign up	20.5	24.5
Didn't sign up	79.5	95.5

The sad ending: this experiment was inconclusive in showing that the two different buttons differed in sign-up rates :(

Observed in experiment

	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn't sign up	75	100

Expected if null hypothesis is true

	LEARN MORE	SIGN UP
Visitor did sign up	20.5	24.5
Didn't sign up	79.5	95.5

In statistical terms, we calculated the chi-squared test $p \approx 0.10$, which means we cannot reject the null hypothesis …

Is the Null Hypothesis true? We still don't know, but we can't reject it yet.

The "Learn More" button will lead to no significant change in the number of people signing up to donate to Obama's campaign versus the original "Sign Up" button.

A small chi-squared value means that observed rates are very close to the expected rates, so there's a high probability that the null hypothesis is true (e.g., p=0.975)

https://people.richland.edu/james/lecture/m170/tbl-chi.html

Table: Chi-Square Probabilities

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860

A large chi-squared value means that observed rates are very far away from expected rates, so there's a low probability that the null hypothesis is true (e.g., p =0.005)

Observed in experiment

	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn't sign up	75	100

Expected if null hypothesis is true

	LEARN MORE	SIGN UP
Visitor did sign up	20.5	24.5
Didn't sign up	79.5	95.5

Observed in experiment

	LEARN MORE	SIGN UP
Visitor did sign up	25	20
Didn’t sign up	75	100

How would we get a larger chi-squared value and hence smaller p-value so that we can reject the null hypothesis with confidence?

One way: if we observed the same proportions but with far more people.

Observed in experiment

	LEARN MORE	SIGN UP
Visitor did sign up	250	200
Didn’t sign up	750	1000

How would we get a larger chi-squared value and hence smaller p-value so that we can reject the null hypothesis with confidence?

One way: if we observed the same proportions but with far more people.

Observed in experiment

	LEARN MORE	SIGN UP
Visitor did sign up	2500	2000
Didn't sign up	7500	10000

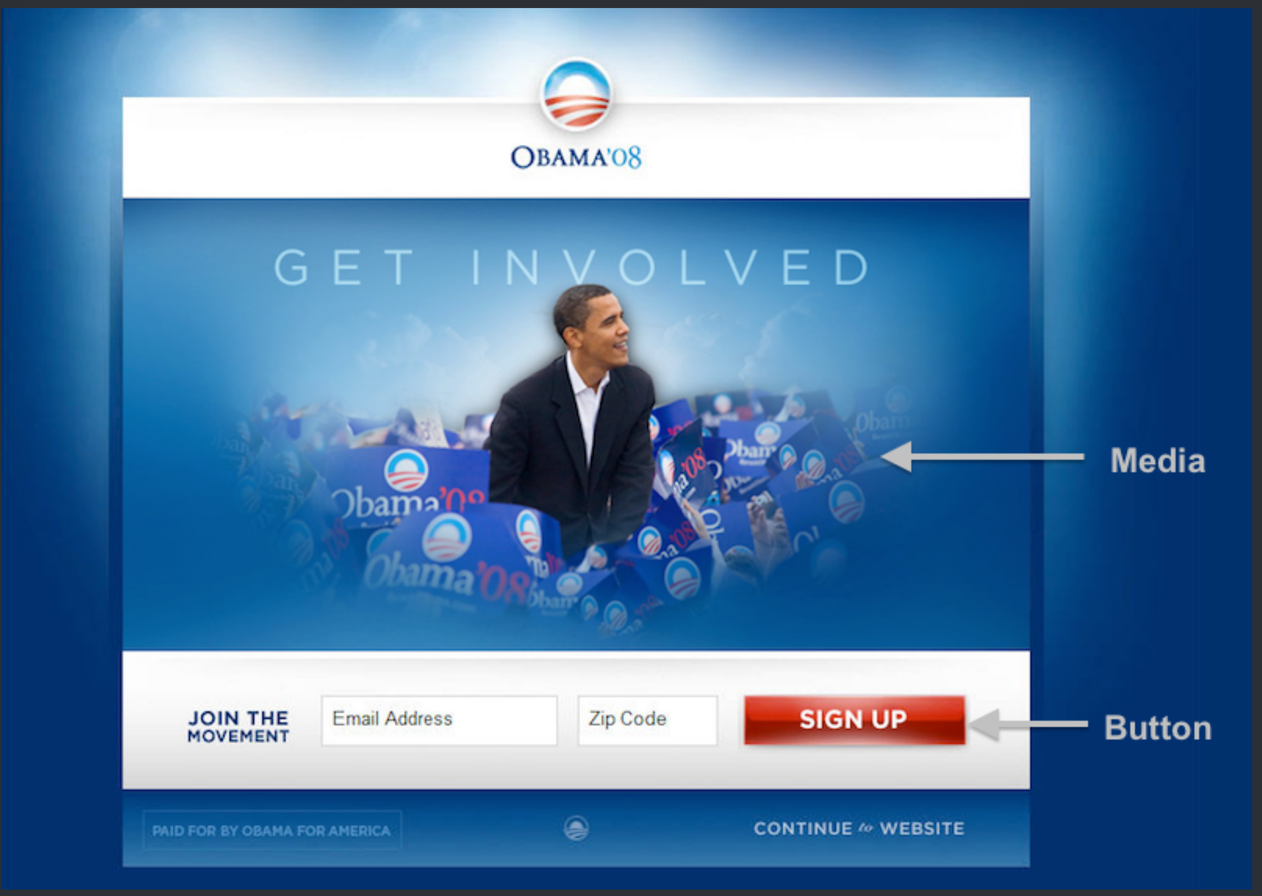
How would we get a larger chi-squared value and hence smaller p-value so that we can reject the null hypothesis with confidence?

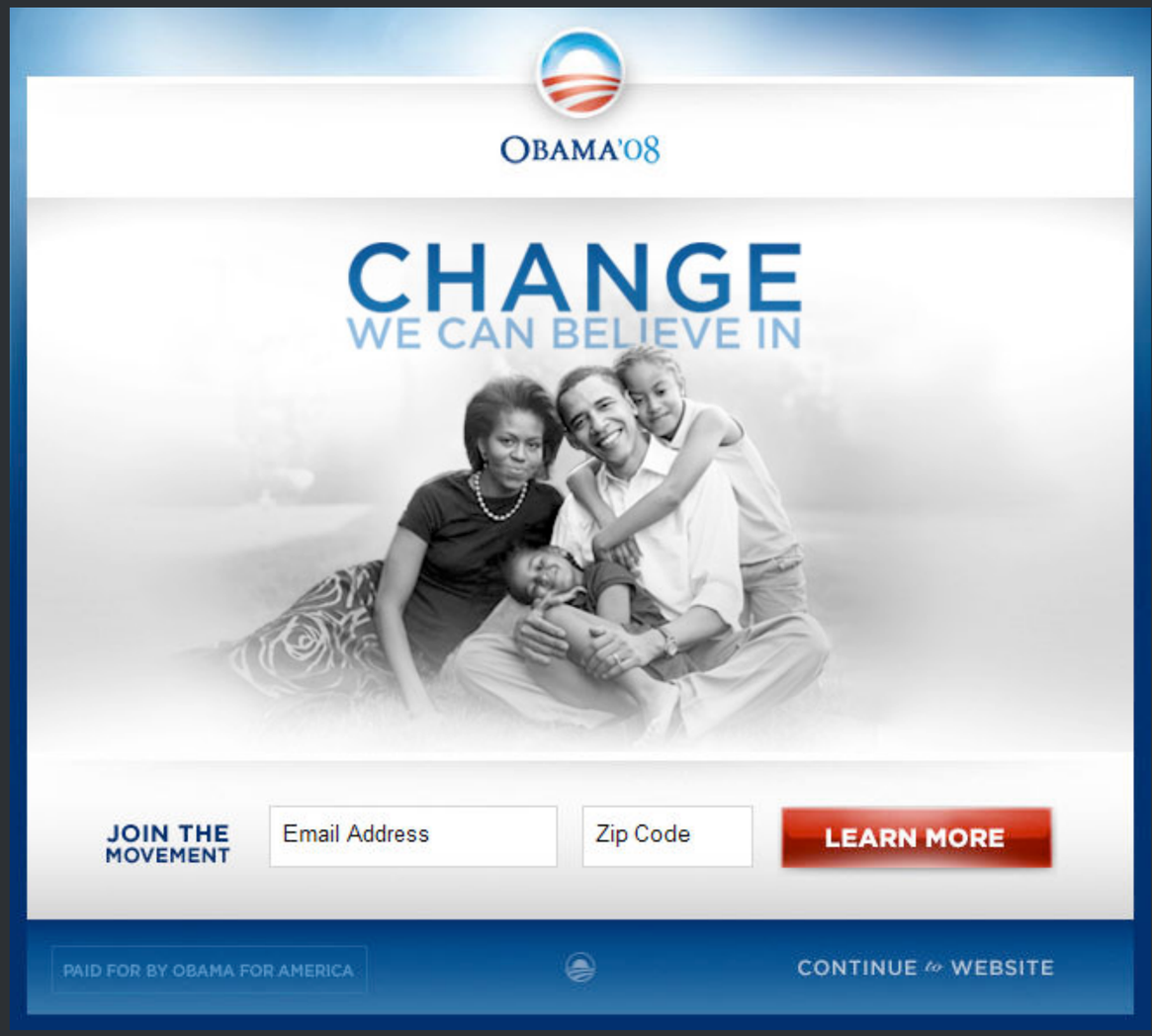
One way: if we observed the same proportions but with far more people.

For your assignment, you should calculate these chi-squared values by hand to show your work, but in the real world, you can use software or online calculators to run the *chi-squared test*. (Beware: there are many variants of this test!)

The real experiment had a happy ending, though!

The winning design, based on online A/B testing, led to an extra **$60 million** in donations





Source: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

Learning Objective

to measure the usability of your app by planning, running, and analyzing data from experiments

TODOs after class

- check Google Spreadsheet grades for accuracy
- lots of coding and user testing coming up!