

Estimating causal effects in large scale online experiments and designing automated A/B testing platforms for machine learning.

Zuzanna Klyszejko, MongoDB

Jerry Chen, Wayfair



Running experiments in an academic lab is different than running online experiments for a company

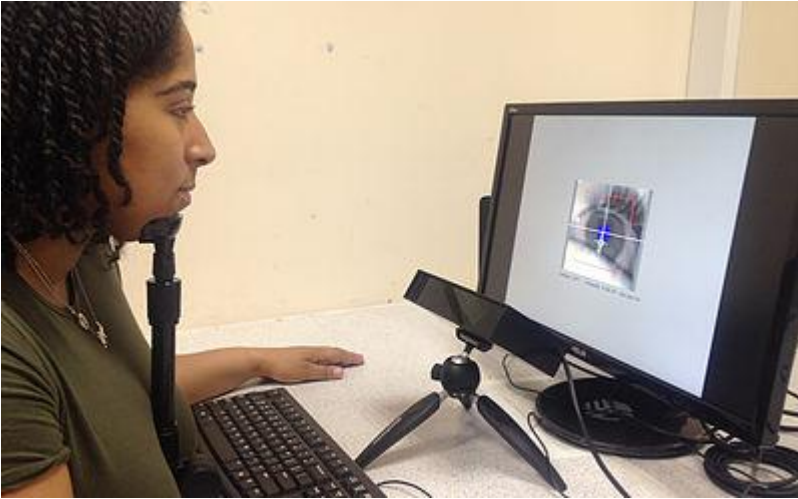



image credit: MAGIC lab

VS



Running experiments in an academic lab is different than running online experiments for a company ... or is it?

- ✓ Replicability across multiple studies
- ✓ Consistency of effects across samples (with variation)
- ✓ Delivery of meaningful results at lowest possible cost
- ✓ Implementation / publication of incremental findings

... and industry moves at a much faster pace 

formulating your hypothesis

"We don't want to prove alternative hypothesis to be right. In fact, we would prefer to prove that null hypothesis is correct."

sample selection

"We don't have sufficient power to run a test"

"Despite full randomization we see bias in our sample."

execution and monitoring

"Users participating in our experiment exhibit unusual behavior."

"Somehow users from Group A ended up doing B"

"There was a change in the backend and our experiment stopped unexpectedly"

scaling up your system

"We have multiple experiments running concurrently (within data science, and also for pricing, product, etc). How do we make sure they don't interfere with each other?"

Modeling and data analysis

"Model assumptions are almost always violated - what do we use to analyze our data?"

formulating your hypothesis

"We don't want to prove alternative hypothesis to be right. In fact, we would prefer to prove that null hypothesis is correct."

sample selection

"We don't have sufficient power to run a test"

"Despite full randomization we see bias in our sample."

execution and monitoring

"Users participating in our experiment exhibit unusual behavior."

"Somehow users from Group A ended up doing B"

"There was a change in the backend and our experiment stopped unexpectedly"

scaling up your system

"We have multiple experiments running concurrently (within data science, and also for pricing, product, etc). How do we make sure they don't interfere with each other?"

Modeling and data analysis

"Model assumptions are almost always violated - what do we use to analyze our data?"

Challenge #1:

*We don't have enough
power to run a test*

Quick recap about hypothesis testing, significance and power

$$H_0 = \mu_1 - \mu_2 = 0 \quad \leftarrow \text{null hypothesis}$$

$$H_a = \mu_1 - \mu_2 \neq 0 \quad \leftarrow \text{alternative hypothesis}$$

		Population	
		H_0 true	H_a true
Experimental sample	reject H_a		Type II error (probability = β)
	reject H_0	Type I error (probability = α)	

false positive - significance (points to Type I error)

false negative - power (points to Type II error)

Solution #1: In order to correctly assess power it's good to understand what influences its value and know your base rates

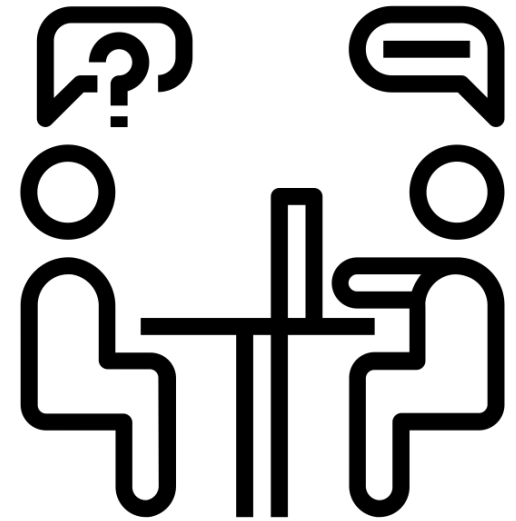
- You can either increase sample size or reduce variance of treatment effect to increase the power.
- Dig into your data: maybe there is a systematic measurement error that affects the variance?

Tip: Understand your base rates. For example, it is absolutely essential you know what is the typical conversion rate of the website experience you want to improve.

—🔗 **Read / code more:** Pettingill, L.M. (2017). 4 Principles for Making Experimentation Count.

Solution #2: If still unable to run a test with sufficient power, be creative

- Can you limit the scope of your experiment?
- Quantitative online experiment is not the only way to learn about your users or establish your base rates. If you find out that required sample to conduct an experiment with sufficient power is unfeasible, think about other ways you can learn about your users. Can you partner with UX researchers?




— **Read more:** Pettingill, L.M. (2017). 4 Principles for Making Experimentation Count.

Challenge #2:
***Despite full randomization we
see bias in our sample***

Solution #1: If not too late, use blocked design which helps you control for factors that influence your outcome variable and infer about causality

- Implement automated blocked design on key variables you want to measure and track backend pipeline for group assignment (for post-mortem).

Tip: If you don't have the capacity to do it yourself use open source tools such as **blockTools** or **agricolae** in R.

—  **Read / code more:** A. Gerber & D. Green “Field Experiments”

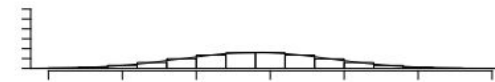
Solution #1: If not too late, use blocked design which helps you control for factors that influence your outcome variable and infer about causality

Assign participants into blocks

# Orders	# Page views	Return rate	Group
16	145	0.04	1
16	130	0.035	2
16	100	0.06	3
15	150	0.06	4
15	150	0.004	5
15	100	0.003	1
14	200	0.03	2
13	100	0.04	3

Intuition behind blocking:

If you choose the right predictors your estimates will have lower variance



fully randomized design



blocked design (strong predictor)



blocked design (weak predictor)

Solution #2: If too late and you collected your data use covariate adjustment

- Adjust your effects by entering those variables as covariates into your model. You can confirm how this works in backtesting (A/A test).

$$Y_i = \alpha + \beta Z_i + \gamma X_i + \epsilon_i$$

where X_i is a list of your covariates

Tip#1: Adjusting for covariates works best when they are predictive of the outcome variable.

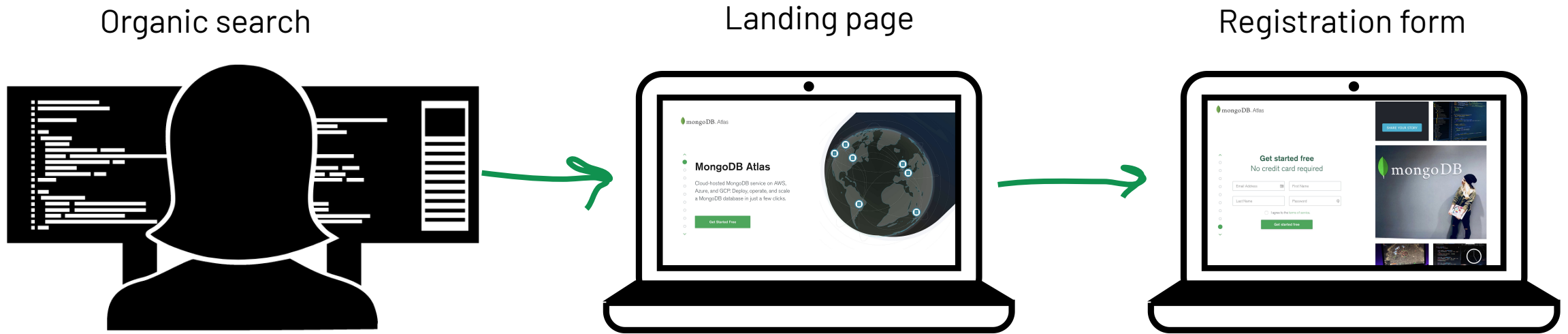
Tip#2: Automate adjusting for covariates when you monitor your tests (run A/A test before turning on A/B test).

🔗 **Read / code more:**
A. Gerber & D. Green “Field Experiments”
EGAP “10 Things to Know About Covariate Adjustment”

Challenge #3:

*We have multiple experiments
running at the same time*

What does it mean to be running multiple experiments at the same time? Contrary to a common belief it applies to teams of all sizes!



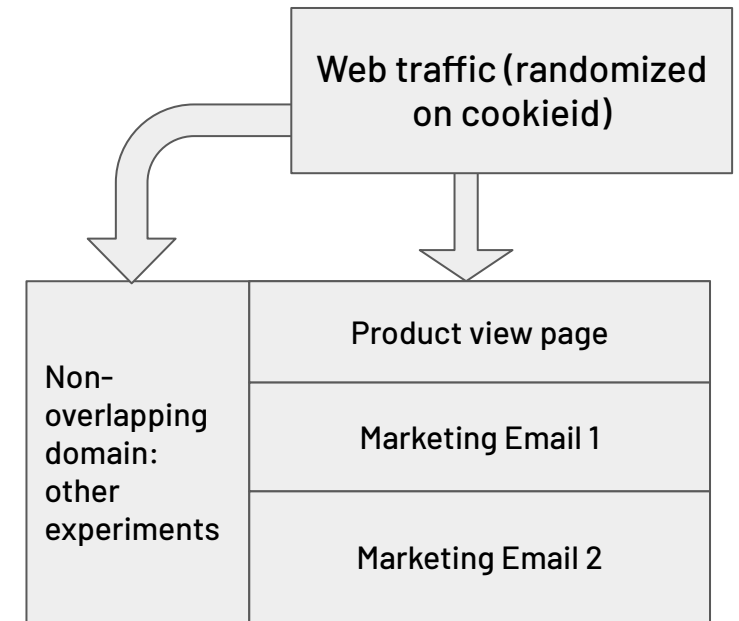
Each of those steps could be managed by a different team running a separate experiment. Without coordination we could a) get confounding results and b) decrease user experience.

Solution: use experimental “layers” which control in experiments each user / cookieid participates in.

Each user / cookieid participates in experiments that affect non-overlapping aspects of experience and within each subset you can manipulate a whole array of experimental parameters (e.g. font color, background color etc.)

- e.g. the banner in product view page will be in an Experiment A, and then after a user adds this product to cart they will receive a marketing email from Experiment B. The

Tip: Re-randomize users often to avoid carryover effects in your experiments.



Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments.

Kohavi, R. & Thomke, S. (2017). The surprising power of online experiments. [and others from Kohavi's group]

Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation.

🔑 Read /
code more:

For randomizing on the level of users / items / cookieids you can start with PlanOut, a free tool with implementations in Java, Python, JavaScript and PHP

```
button_color = uniformChoice(  
choices=['#51c3c9ff',  
        '#13a950ff', '#b33316'],  
unit=cookieid);  
  
button_text = weightedChoice(  
choices=['Sign up', 'Join  
now'], weights=[0.8, 0.2],  
unit=cookieid);
```

Proportional Web traffic split
based on experimental condition :

button_text	<div>“Sign up”, $p = 0.8 \times 0.3$</div>	<div>“Sign up”, $p = 0.8 \times 0.3$</div>	<div>“Sign up”, $p = 0.8 \times 0.3$</div>
	<div>“Join now” $p = 0.2 \times 0.3$</div>	<div>“Join now” $p = 0.2 \times 0.3$</div>	<div>“Join now” $p = 0.2 \times 0.3$</div>
button_color			

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments.

Kohavi, R. & Thomke, S. (2017). The surprising power of online experiments. [and others from Kohavi's group]

Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation.

Read /
code more:

Finally, you need to facilitate knowledge exchange between teams. Team size and business needs determine best strategy.

- if you're at a large company: consider creating a centralized team that owns planning, execution and reporting on experiments (see Kohavi & Thomke for more on this)
- if you're at a small / mid-size company: discuss and implement cross-team strategies to store and share experimental results and agree on the definition of your Key Performance Metrics. This could take a form of an experimental council. Read e.g. Kohavi & Thomke for more on this.

🔑 **Read / code more:** Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments.
Kohavi, R. & Thomke, S. (2017). The surprising power of online experiments. [and others from Kohavi's group]
Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation.

Key Takeaways to improve causal inference from A/B experiments at your company:

- I forgot to mention this – but first thing you should do is to be experimentation advocate and ask annoying questions
- Introduce and eventually automate blocked experimental design, power analysis and covariate adjustment combined with A/A backtesting before launching an actual A/B test
- Coordinate concurrent experiments and start building experimentation culture

“Good judgment comes from experience, and a lot of that comes from bad judgment.”

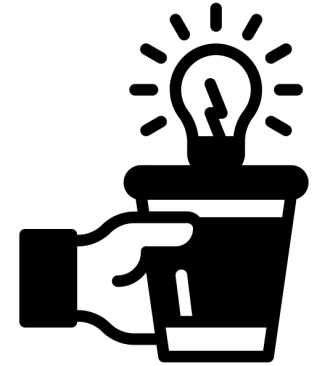
Will Rogers

Thank you for your attention!

Feel free to ask me about life at MongoDB and take a look at openings for internships and full time positions:

twitter.com/MongoDBcareers
mongodb.com/careers

If you want great short reads about experimentation for your coffee break here are my top suggestions:



- “Points of Significance” series by *Nature Methods*
- Method Guides series from EGAP (Evidence in Governance and Politics)
- Stitchfix, Netflix, and Microsoft / Bing blogs
- A. Gerber & D. Green’s “Field Experiments” (because never enough, such a great book)

