!mpactmakers

A/B = A Bad Idea?

Improved Insights with Design & Analysis of Experiments in R

David Moxley

11 March 2020

Intro

!m

impactmakers

To Make a Lasting Impact

David Moxley

Senior Consultant

Impact Makers

Founded in **2006**

$20+ million revenue

80 employees

Certified B Corp

Certified B Corporation

Motivation

- Lack of emphasis on experimentation in the Richmond market
- Prevalence in a variety of industries and business units
- Popularity of A/B design
- Ignored subtleties of A/B tests

Agenda

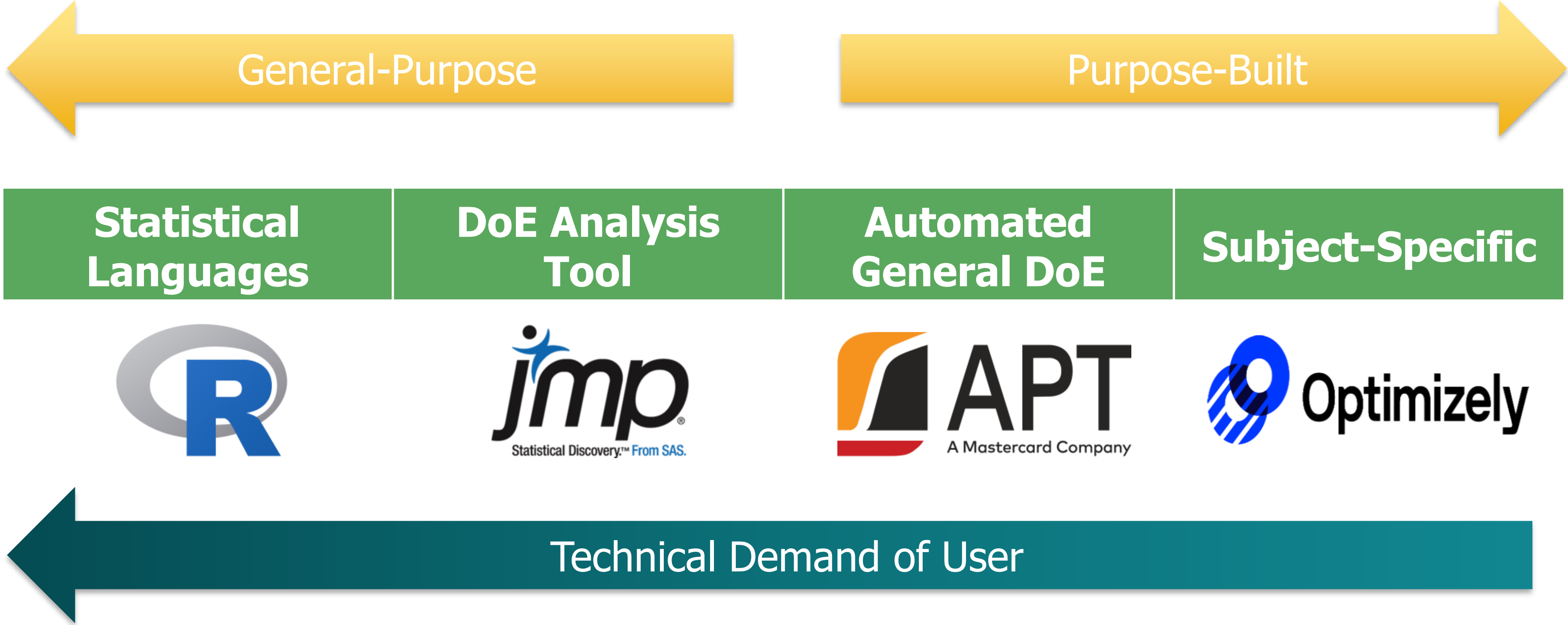
- Why do we experiment & why analyze in R?
- What’s an A/B experiment?
- Case Studies
- Alternative Design of Experiments
- Analyzing in R
- Developing a broader “Test and Learn” culture

01

Why R?

Tools

The Marketplace



General-Purpose

Purpose-Built

Statistical Languages

DoE Analysis Tool

Automated General DoE

Subject-Specific

Technical Demand of User

Why analyze experiments in R?

As a practitioner…

- Continuity with other workflows
- Transparency
- Powerful visualization packages
- Instructional value*

As a business leader…

- Extensible framework
- Transparency
- Learning curve for analytical teams
- Cost

02 | Why do we experiment?

Why do we experiment?

- “You break you buy it”

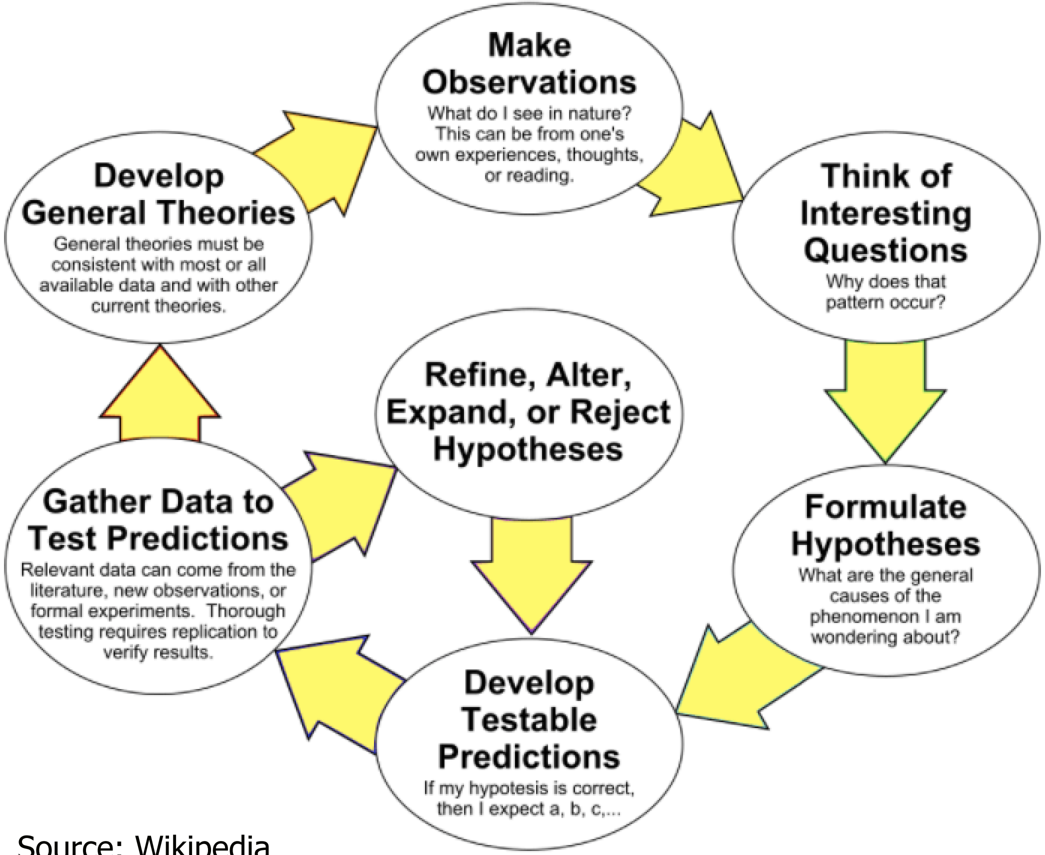
Why do we experiment?

- ~~"You break you buy it"~~
- Break it before you buy it
- Foster a learning culture
- Tease out causality
- Provide direction to the business

CartoonStock.com

The Goal

- Minimize the impact on the business*
- Offer a simple design
- Eliminate systematic error
- Understand the range of validity
- Offer a precise estimate
- Convey uncertainty
- Iterate!


Make Observations
What do I see in nature? This can be from one's own experiences, thoughts, or reading.
Think of Interesting Questions
Why does that pattern occur?
Formulate Hypotheses
What are the general causes of the phenomenon I am wondering about?
Develop Testable Predictions
If my hypotesis is correct, then I expect a, b, c,...
Gather Data to Test Predictions
Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.
Refine, Alter, Expand, or Reject Hypotheses
Develop General Theories
General theories must be consistent with most or all available data and with other current theories.

Source: Wikipedia

Types of Experiments

Types	Example	Strengths	Potential Issues
Laboratory	• Survey Research	• High internal validity • Ease of replicability	• Lack of Realism • Poor Generalizability
Field	• Tele-marketing	• Strong internal validity • Very Realistic	• Generalizability (?) • Potential selection bias
Natural	• TV Ad test	• Highly Realistic • Generalizability (?)	• Causal Inference (?) • Data Collection

What's an A/B Experiment?

- Randomized Controlled Trial
    - Split-run testing
- Random assignment to two or more variants, A, B,…*n*
- Widely used for testing Machine Learning models
- Popular in digital space, UX research, etc.

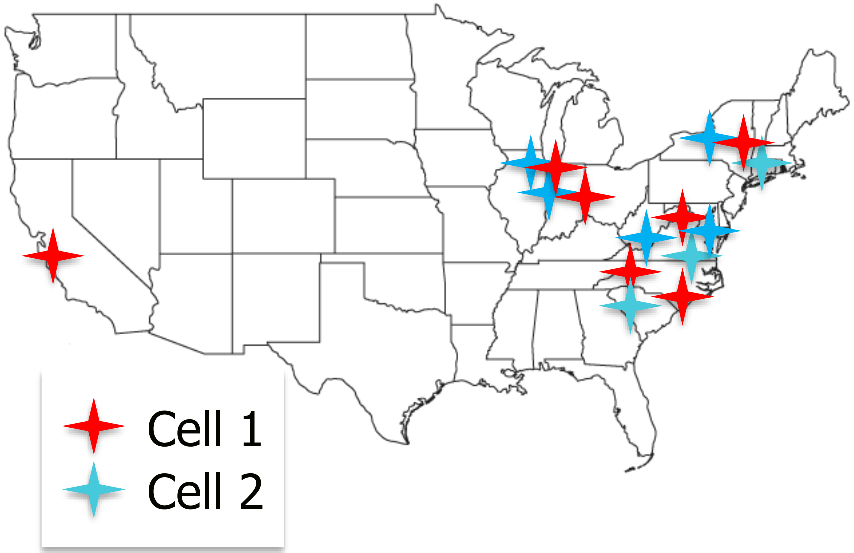
03

Naïve A/B Designs

Case Studies

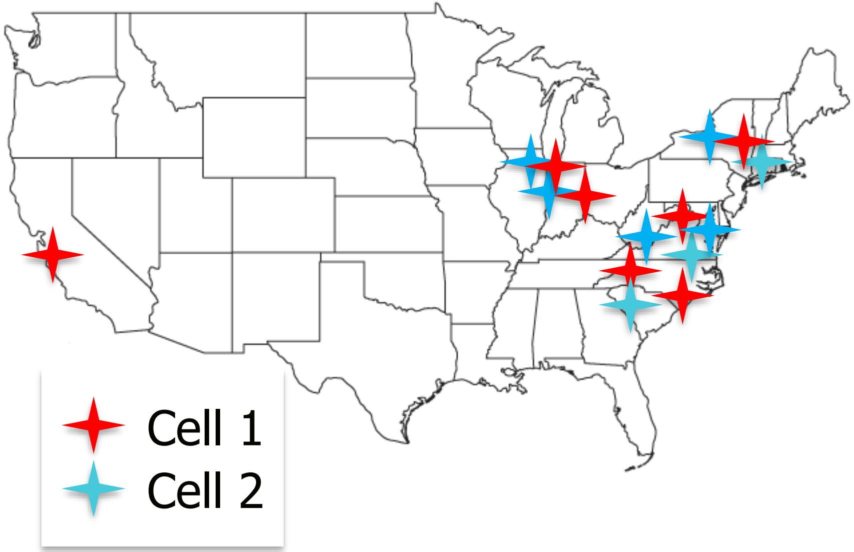
Case Study: Price Elasticity Test

- B2B retailer tested price changes for a set of SKUs
    - Cell 1: 5% increase, free shipping
    - Cell 2: 10% increase, free shipping
- Salespeople/Accounts pseudo-randomly assigned to cells
    - Cells were balanced for total sales
- Total Sales for the Cells were measured
- Analysis via Difference in Differences
- Test continually monitored until predetermined alpha of .1 was reached

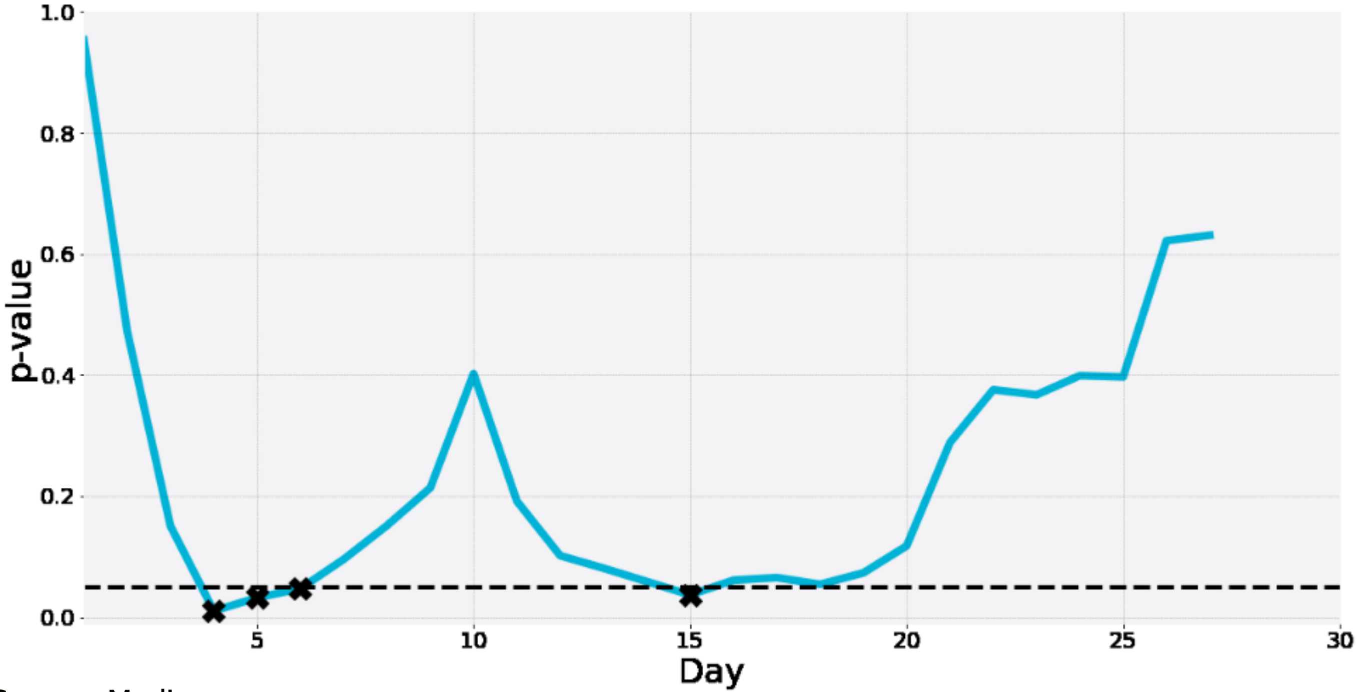


Case Study: Potential Drawbacks

- No treatment for confounding factors
- Sample bias invalidated Difference in Differences analysis
- Entangle effects through overly simple design
    - No understanding of interactions
- “Peeking” violated an underlying assumption of statistical inference



What is “Peeking”?



Source: Medium.com

The Real Issue

"To consult the statistician[/data scientist/consultant] after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of."

-R.A. Fisher, 1938

04

Design of Experiments

Overview

Keys to a Good Design

1. Replication
2. Randomization
3. Control

Types of Designs

- Comparative studies
- Single Factor
- Blocking Designs
  - Randomized Complete Block
  - Balanced Incomplete Block
- Factorial
- Fractional Factorial
- Response Surface Designs

Types of Designs

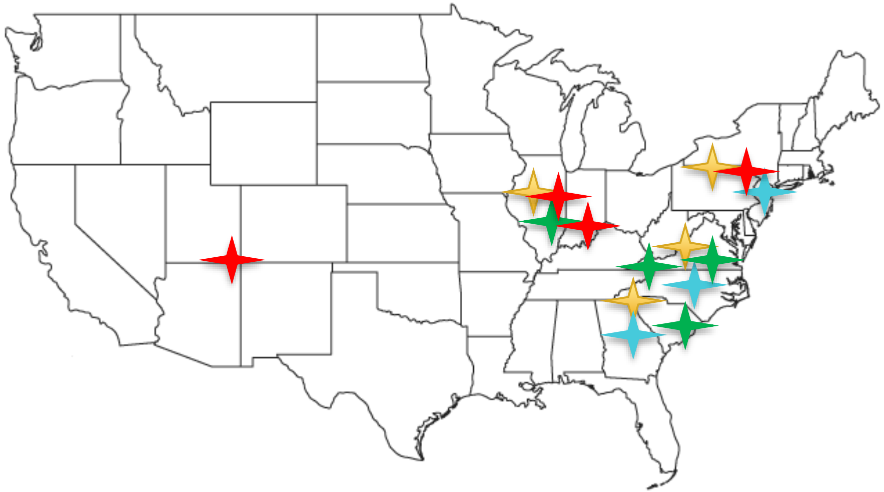
- Comparative studies
- Single Factor
- Blocking Designs
  - Randomized Complete Block
  - Balanced Incomplete Block
- Factorial
- Fractional Factorial
- Response Surface Designs

05 | Redesign Our Case Study

Case Study: Two-Factor Factorial in a Randomized Complete Block

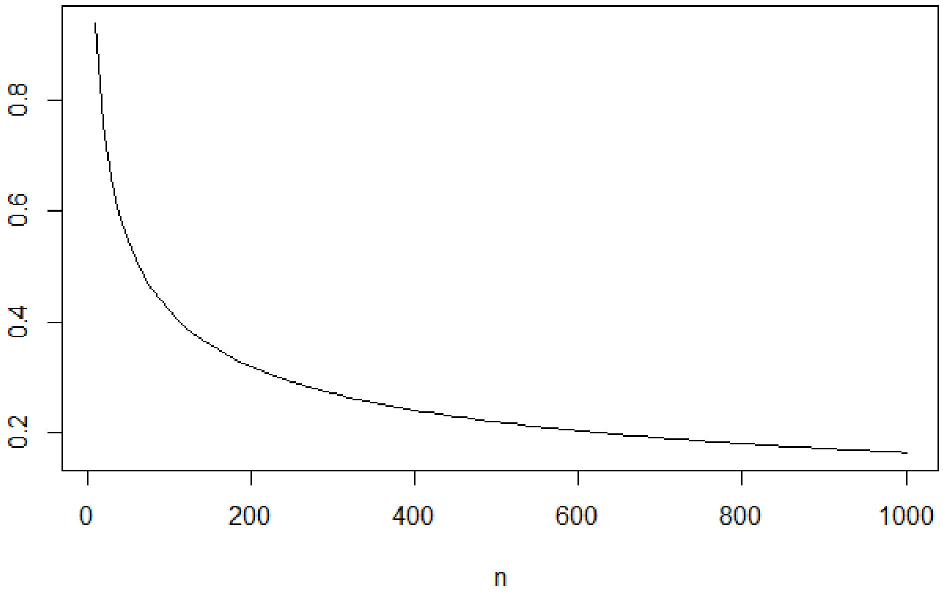
Sales Person	Price % Increase	Shipping Fee
1	5	Yes
1	10	No
1	10	Yes
1	5	No
…	…	
5	5	Yes
5	10	No
5	10	Yes
5	5	No

$$y_{ijk} = \mu + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \beta_k + \epsilon \begin{cases} i = 1 \; to \; 2 \\ j = 1 \; to \; 2 \\ k = 1 \; to \; 5 \end{cases}$$



Case Study: Sequential Sampling Procedure

- Frequentist hypothesis testing assumes **fixed sample**
    - Central Limit Theorem
- Business pressure, ethical considerations, user negligence can lead to a desire to monitor, "peek" at test results
- Larger conversation intersects with Theory of Optimal Stopping
    - The "Secretary Problem"



Case Study: Sequential Sampling Procedure

- Wald's Sequential Probability Ratio Test (SPRT)
- Optimizely's Mixture Sequential Probability Test (mSPRT)
- Multi-Armed Bandit
- Bayesian Methods
- Evan Miller's Sequential Procedure with Stopping Metric
    1. Choose a target sample size, n, at the outset
    2. Assign to treatments with equal probabilities
    3. Track incoming successes for each treatment cell, A, B,..*i*
    4. If A – B = $2\sqrt{n}$, declare A the winner, B-A = $2\sqrt{n}$, B is the winner
    5. If T + C = N, there is no winner, fail to reject null hypothesis

06 | Analysis Procedures

Overview

Stats Refresher

- Degrees of Freedom
- T-test
  - One-Sample: $\frac{\bar{X}-u}{\frac{sd}{\sqrt{n}}}$
  - Unpaired: $\frac{\bar{X}_1-\bar{X}_2}{sd_p\sqrt{\frac{1}{n_1}-\frac{1}{n_2}}}$
  - Paired: $\frac{\bar{X}_d}{\frac{sd_d}{\sqrt{n}}}$
- Sum of Squares: $\sum(x_i - \bar{x})^2$

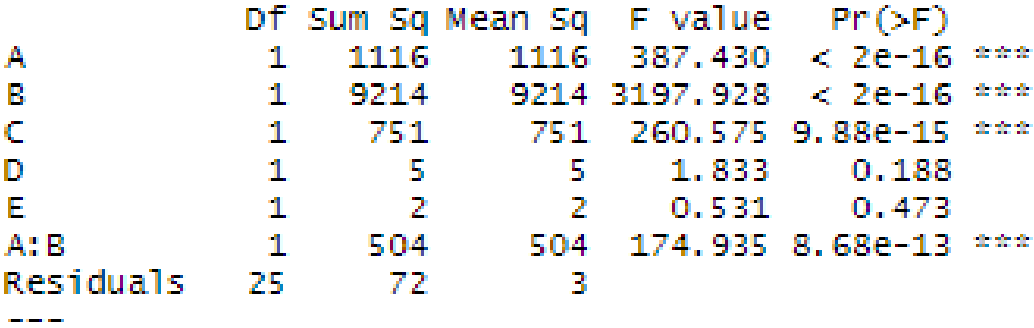
- F-test
  - $\frac{MSE_{Larger\ Sample}}{MSE_{Smaller\ Sample}}$

- p-value: Type I error
- Power: Type II error

Stats Refresher

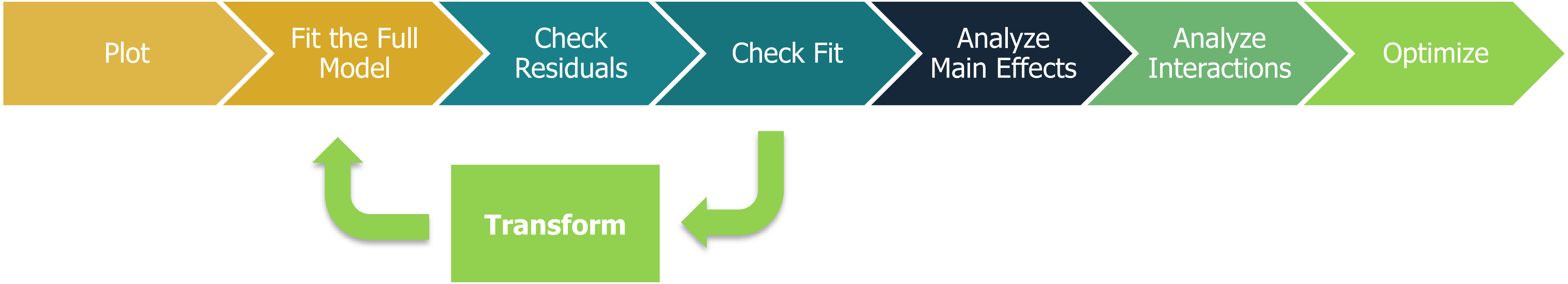
- ANOVA

$$y_{ijk} = \mu + \tau_i + \alpha_j + \beta_k + \epsilon \begin{cases} i = 1\ to\ i \\ j = 1\ to\ j \\ k = 1\ to\ k \end{cases}$$



```
             Df Sum Sq Mean Sq  F value   Pr(>F)    
A             1   1116    1116  387.430  < 2e-16 ***
B             1   9214    9214 3197.928  < 2e-16 ***
C             1    751     751  260.575 9.88e-15 ***
D             1      5       5    1.833    0.188    
E             1      2       2    0.531    0.473    
A:B           1    504     504  174.935 8.68e-13 ***
Residuals    25     72       3                      
---
```

General Procedure



07

Analyzing in R

Case Studies

R Script

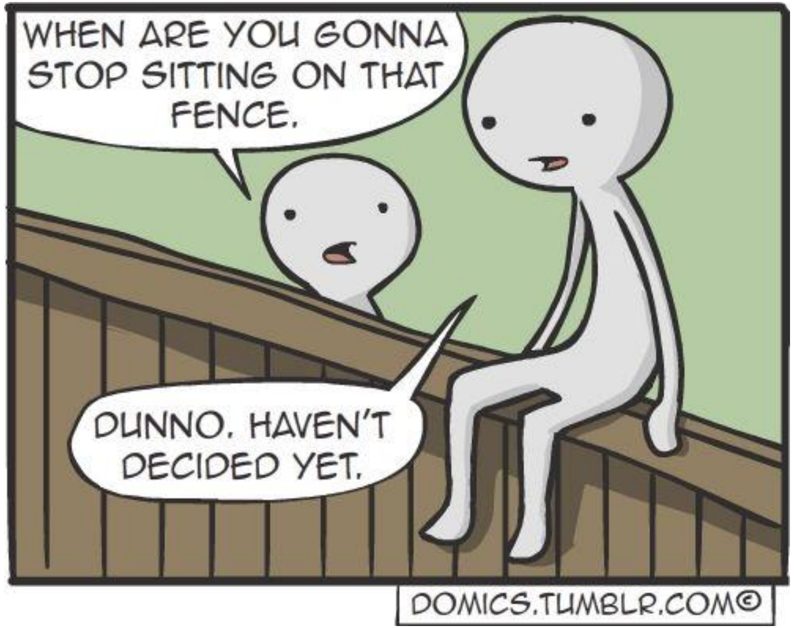
08

A/B[ad] Idea?

Conclusion

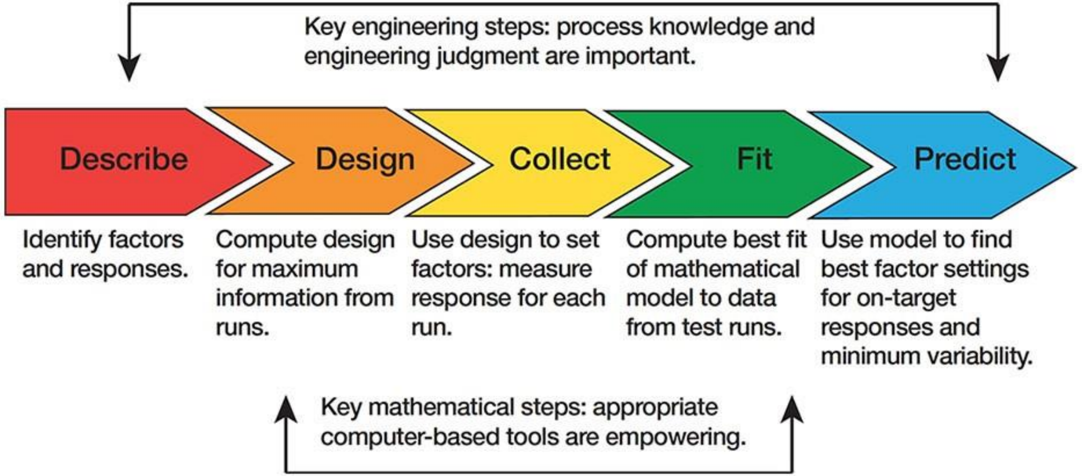
Are A/B Designs a good idea?

Pros	Cons
Simple Design	Entangled Effects
Statistical efficiency with minimal sample	Susceptible to systematic bias
Well suited for digital studies	Potential issues with "peeking"


WHEN ARE YOU GONNA STOP SITTING ON THAT FENCE.
DUNNO. HAVEN'T DECIDED YET.
DOMICS.TUMBLR.COM©

Parting Advice

- Collaborate with the business
- Keep Design and Analysis simple
- Remember statistical vs business significance
- Experimentation is mean to be iterative


Key engineering steps: process knowledge and engineering judgment are important.
Describe
Design
Collect
Fit
Predict
Identify factors and responses.
Compute design for maximum information from runs.
Use design to set factors: measure response for each run.
Compute best fit of mathematical model to data from test runs.
Use model to find best factor settings for on-target responses and minimum variability.
Key mathematical steps: appropriate computer-based tools are empowering.

Source: JMP.com

Thank you!

- Connect: dmoxley@impactmakers.com
- Questions: dmoxley@impactmakers.com
- Data/Slides/R Code: github.com/davidrmoxley/DoE
- Continue the Conversation: impactmakers.lpages.co/advanced-analytics/