

Part 2

Planning, Running, and Analyzing Controlled Experiments on the Web

Roger Longbotham,
Mgr Analytics, Experimentation Platform, Microsoft

Slides available at <http://exp-platform.com>

Planning and Analysis of Online Experiments

- What to measure
- How to compare Treatment to Control
- How long to run test
- Start up options
- Good test design
- Data validation and cleansing
- Before your first experiment
- Common errors
- MultiVariable Tests

What to Measure

- Start with objective
 - Of the site (content, ecommerce, marketing, help/support,...)
 - Of the experiment
- What can you measure to tell you if you met your objective?
 - **Content site:** clicks/user, pageviews/user, time on site
 - **Ecommerce:** rev/visitor, units purchased/visitor, cart-adds/visitor
 - **Marketing:** referrals/visitor, time on site
 - **Help/support:** Pct of users engaged, Pct of users who print, email or download content, time on site

What to Measure

- Measures of user behavior
 - Number of events (clicks, pageviews, scrolls, downloads, etc)
 - Time (minutes per session, total time on site, time to load page)
 - Value (revenue, units purchased)
- Experimental units
 - Per user (e.g. clicks per user)
 - Per session (e.g. minutes per session)
 - Per user-day (e.g. pageviews per day)
 - Per experiment (e.g. clicks per pageview)

Overall Evaluation Criterion

- It is very helpful to have a single metric that summarizes whether the Treatment is successful or not – the Overall Evaluation Criterion, or OEC
- Examples:
 - **Content site:** OEC could be clicks/user or time on site
 - **Ecommerce:** rev/user or lifetime value
 - **Help/support site:** Survey responses
- OEC could also capture monetary value of the site, aka ROI (return on investment)

Comparing Treatment to Control

- Single Treatment

- Two-sample t test works well
 - Large samples sizes => Normal distribution for means
 - Calculate 95% Confidence Interval for difference in two means

$$(\bar{X}_T - \bar{X}_C) \pm 1.96 * S_{\bar{X}_T - \bar{X}_C}$$

if zero not in the interval conclude Treatment mean different from Control

- May have many tests, OEC critical

- Multiple Treatments

- Multiple applications of two-sample t test
- Analysis of Variance

Sample UI for test results

Experiment - MSN Homepage Experiment 5 (Headline Ordering) - 1

ID: msnhp_experiment_5a
Environment: Offline Analysis (Sprint 23)

Details

		Expected Split	Actual Traffic	Description:
Control:	Control	50.00 %	1,938,480 unique users	Re-ordered headlines in the Video module.
Treatment:	Treatment	50.00 %	1,934,921 unique users	

Metrics

 Refresh

Improvement of: T1 over C

Last refresh: 10/19/2007 6:00:00 PM

Name	T1		C		P Value	% Change	Significant
	Observations	Avg	Observations	Avg			
Clicks - News - By Session	179,102		180,545		0.592	0.12 %	No
Clicks - Search - By Session	1,046,728		1,050,701		0.112	-0.18 %	No
Clicks - Sports - By Session	178,261		178,805		0.495	-0.16 %	No
Clicks - Stocks - By Session	53,204		52,752		0.240	-0.94 %	No
Clicks - Today - By Session	89,650		89,637		0.587	-0.10 %	No
Clicks - Video - By Session	56,275		56,671		0.023	-0.73 %	Yes
Clicks - Whole Page - By Session	4,236,134		4,250,427		0.065	-0.18 %	No
CTR - Entertainment - By Session	271,081		272,110		0.235	-0.55 %	No
CTR - Infopane - By Session	332,812		333,514		0.894	-0.04 %	No
CTR - Money - By Session	76,837		77,158		0.145	-1.07 %	No
CTR - Navigation - By Session	1,474,608		1,480,126		0.677	0.10 %	No
CTR - News - By Session	177,528		178,850		0.559	-0.30 %	No
CTR - Search - By Session	1,036,185		1,040,149		0.824	0.06 %	No
CTR - Sports - By Session	176,609		177,129		0.366	0.52 %	No
CTR - Stocks - By Session	52,569		52,161		0.047	3.30 %	Yes
CTR - Today - By Session	88,746		88,760		0.008	1.08 %	Yes
CTR - Video - By Session	55,772		56,172		0.420	0.54 %	No

103 metrics

More Details

Clicks - News - By Session

Absolute change: **0.00142**

Confidence interval:
(-0.00377, 0.00661)

Percent change: **0.12 %**

Confidence interval:
(-0.31 %, 0.55 %)

T1

Unique sessions: **147,749**

STDEV (Mean): **0.00190**

C

Unique sessions: **149,114**

STDEV (Mean): **0.00184**

Note:

Averages for
both variants

P-values

Percent change

Significance

Confidence

Intervals

103 metrics

Comparing Treatment to Control

- P-value is the probability of getting a difference farther from zero than observed under assumption of no difference
- CI for percent effect must use special formulas
- Care must be taken in calculating standard deviations
 - When randomization is by user, any metric that is not per user must take into account non-independence in calculating standard deviation
 - We routinely use bootstrapping to estimate standard deviations

$$n = \frac{16 * r * \sigma^2}{\Delta^2}$$

Power and Sample Size

- The power of a test is the probability of detecting a difference (Δ) of a given size i.e., it is 1-Prob(Type II error)

Power depends on

- The size of effect you want to be able to detect, Δ
- Variability of the metric
- Number of users in each group (T/C)

It is typical to determine the sample size needed to achieve 80% power

Power and Sample Size

- Example: Total number of users needed to achieve 80% power, with equal number of users in Treatment and Control and with standard deviation s is

$$N = \frac{32s^2}{\Delta^2}$$

Ramp up

- Often good practice is to start with small percent in Treatment and increase when you have confidence Treatment is bug-free
- Sample ramp up schedule:
 - 1% in Treatment for 4 hours
 - 5% in Treatment for 4 hours
 - 20% in Treatment for 4 hours
 - 50 % in Treatment for 14 days

Simpson's Paradox

- Beware Simpson's paradox when percent in Treatment changes
- Example: data generated for 7day test with Treatment mean 5% higher than Control mean each day.
First 5 days Treatment had 20% of visitors/day (100,000 visitors/day)
Last 2 days Treatment had 50% of visitors (100,000/day)
(Assume last two days were weekend and averages dropped from about 1.7 to 1.2)

Simpson's Paradox example

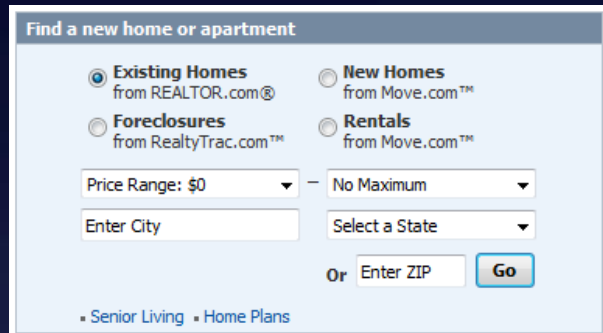
Metric is number of sessions per user per day

Day	T mean	C mean	% in T	users	T users	C users	T total	C total	% effect
M	1.89	1.8	20%	100000	20000	80000	37800	144000	5.00%
T	1.995	1.9	20%	100000	20000	80000	39900	152000	5.00%
W	1.89	1.8	20%	100000	20000	80000	37800	144000	5.00%
Th	1.785	1.7	20%	100000	20000	80000	35700	136000	5.00%
F	1.785	1.7	20%	100000	20000	80000	35700	136000	5.00%
Sa	1.155	1.1	50%	100000	50000	50000	57750	55000	5.00%
Su	1.26	1.2	50%	100000	50000	50000	63000	60000	5.00%
						totals	307650	827000	
						Averages	1.53825	1.654	-7.00%

The Treatment effect each day is +5% but estimated cumulative effect is -7%

MultiTreatment Tests

- Example: Real Estate widget design
 - Test five alternatives to the current design
 - OEC: clicks to links weighted by revenue per click



Find a new home or apartment

☒ Existing Homes from REALTOR.com®
☐ New Homes from Move.com™
☐ Foreclosures from RealtyTrac.com™
☐ Rentals from Move.com™

Price Range: \$0 - No Maximum
Enter City Select a State
Or Enter ZIP Go

Senior Living Home Plans

Control



Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

 Enter City State
or
Enter Zip
Find homes

T1

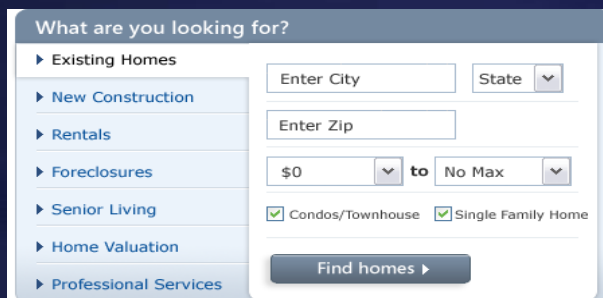


Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

 Enter City State
or
Enter Zip
Find homes

T2



What are you looking for?

Existing Homes New Construction Rentals Foreclosures Senior Living Home Valuation Professional Services

Enter City State
Enter Zip
\$0 to No Max
Condos/Townhouse Single Family Home
Find homes

T3



Find a new Home or Apartment

Enter Zip or Enter City State Search listings

T4



Find Your Dream Home or Apartment

City, State or ZIP

☒ Existing homes ☐ New construction
☐ Foreclosures ☐ Rentals
Search listings

T5

Real Estate Widget

- The widget that performed the best was the simplest



Find Your Dream Home or Apartment

City, State or ZIP

☒ Existing homes ☐ New construction
☐ Foreclosures ☐ Rentals

Search listings ▶

- Revenue increase over control: +9.7%

Note Ronny's example earlier compared the best Treatment to another Treatment, not the Control

BREAK

Design of Experiments

- Triggering
- Blocking
- Measuring non-test factors
- Randomization

Triggering

Only allow users into your experiment if they “trigger” the experiment. i.e. a user’s data should only be used in the analysis of the experiment if they saw one of the variants

Example: MSN UK Hotmail experiment

Which users do you want to track as part of your experiment?

Blocking non-test Factors

- Factor is controlled such that it affects both treatment and control equally, hence not affecting the estimate of the effect
- Blocking on a factor is more common than keeping it fixed (keeping it constant throughout the experiment)
- Advantages to blocking
 - Can estimate the effect of the different levels of the factor, e.g. what is the effect on weekends/weekdays
 - Can make inference to a broader population

Examples of Blocking

- Time (time of day, day of week, etc.)

Bad test design => run control at 100% M-W
then treatment at 100% Th-Sa

Always run treatment and control concurrently in online experiments

- Content

Ex: If content of a site changes during the experiment it must be the same for both Treatment and Control at all times

Design Principle

The Treatment and Control groups should be as alike as possible except for application of the treatment

- Who is in the experiment
- What is done during the experiment
- etc.

Updates to the site during the test must be applied to all variants in the test

Design Principle

Example: One partner was conducting an A/A test (same as an A/B test but no real change is made) What would you expect?

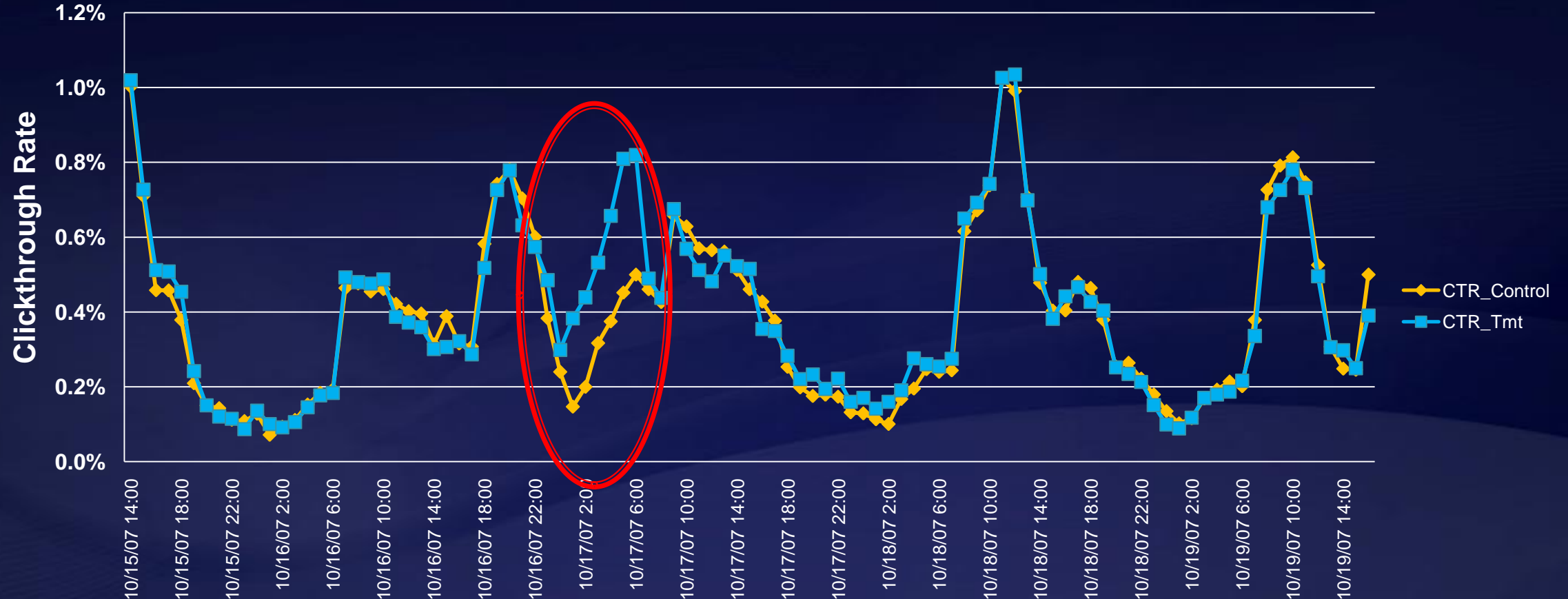
Results: T very significant (much more than it should be)
Why?

Found out another group was using their Treatment group to test something so there really was a difference between T and C

Design Principle

- Ex: A site was testing a change to the layout of their page
 - Content to T and C was not the same for a 7 hour period

Hourly Clickthrough Rate for Treatment and Control for Module



Measure non-test Factors

Measuring the value of non-test factors allows you to

- Delve into why the treatment had the effect it did (e.g. more PVs are correlated with faster load time which explains almost all the effect of T)
- Determine if subpopulations behave the same (e.g. did the treatment have the same effect for new users as for experienced users?)

Randomize

Why randomize?

So that those factors you can't control (or don't know about) don't bias your results



“Randomization is too important to be left to chance”
Robert Coveyou, ORNL

Randomize

How to randomize? (online tests)

Randomly assign T or C to user (alternately could use user-session, search query, page view or product/SKU)

Usually best by user (store UserID in cookie)

How persistent is the UID?

Ideally user always gets same treatment group

Limitations:

- Clearing cookies => can change treatment
- Different browser => may get different treatment
- Can't allow opt-in or opt-out

Representative Test

Make sure users and conditions are as representative of launch environment as possible

- Time period: not holiday (unless holiday factor), pre-holiday, complete cycle (day, week)
- Users: all users who would see T in the future, not robots, not internal testers, outliers(?)
- Not during special events

Robot Detection and Removal

- Remove robots (web crawlers, spiders, etc.) from analysis
 - They can generate many pageviews or clicks in Treatment or Control skewing the results
 - Remove robots with known identifiers (found in the user agent)
 - Develop heuristics to identify robots with many clicks or pageviews in short period of time
 - Other patterns may be used to identify robots as well, such as very regular activity

Data Validation checks

- Carry out checks to make sure data is not affected by some unknown factor
 - Check that percentage of users in each variant is not different from planned (statistical test)
 - Check that number of users in the experiment is approximately what was expected (and doesn't change too much during experiment)
 - Check that the Treatment effect does not change too much during experiment
 - Check that means for primary metrics do not change unexpectedly

Before Your First Experiment

- Conduct logging audit
 - Compare data collected for experiment to system of record
 - Should have approximately same number of users, clicks, pageviews, orders, etc.
- Conduct A/A test
 - Split users into two groups that get same experience
 - Should have about 5% of tests significant
 - p-values should have $U(0,1)$ distribution
 - No p-values should be extremely small (say $<.001$)

Common Errors

- Not conducting logging or A/A tests
 - Find caching issues, UID reassignment
- Not keeping all factors constant or blocking
 - Content changes to site
 - Redirect for Treatment but not for Control
- Sample size too small
- Not measuring correct metric for OEC
 - Measure clicks to buy button (instead of revenue)
 - Clicks to download button (instead of completed downloads)

MultiVariable Tests (MVTs)

- Several factors/variables, each of which has two or more levels (C/T1/T2/...)
- *Main effects*: Comparison of Treatments to Control for each variable (i.e. compare means for T and C same as before)
- *Interactions*: Determine if combinations of variables have different effect than adding main effects

Example: MultiVariable Test on MSN HP

The screenshot shows the MSN homepage with several sections highlighted by colored boxes and labels:

- Top Navigation:** Includes links for Web, MSN, Images, Video, News, and Maps. A search bar is present with the text "Search: Lincoln speech | Avril Lavigne video | Salmonella".
- Left Sidebar:** Contains links for Hotmail, Messenger, My MSN, and MSN Directory. Below these are sections for "Sign out", "Make MSN your homepage", and "Customize your Page".
- Main Content Area:**
 - Top Left:** A video player titled "The Life of Lovebirds" with a "click to play" button.
 - Top Center:** A news article titled "50 Dead as Plane Slams Into Home" with a large image of a plane crash.
 - Top Right:** A T-Mobile advertisement featuring a flip phone and the text "Right now, get T-Mobile's best rates, with or without the long-term contract." This section is labeled **F1**.
 - Middle Left:** A section titled "Today's Picks" with a list of articles.
 - Middle Center:** A section titled "Angels & Demons" with a video player and a list of articles.
 - Middle Right:** A section titled "MSNBC News" with a list of articles. This section is labeled **F2**.
 - Bottom Left:** A section titled "Custom MSN Content" with a list of articles.
 - Bottom Center:** A section titled "Entertainment" with a list of articles.
 - Bottom Right:** A section titled "FOX Sports" with a list of articles. This section is labeled **F3**.
- Right Sidebar:** Contains links for Tech & Gadgets, TV, Weather, White Pages, and Yellow Pages.

Factors/variables

F1: Size of Right col ad

C = current size

T1 = 10% larger

T2 = 10% smaller

F2: MSNBC news stories

C = Top international

T = Specific to country ID'd

F3: Sports/Money placement

C = Sports above Money

T = Money above Sports

OEC: Clicks per User

Other metrics: PVs, CTR

(This is for illustration purposes only, it does not reflect any previous or planned test on MSN HP)

Multivariable Tests

Advantages:

- Can test many things at once, accelerating innovation
- Can estimate interactions between factors

Disadvantages

- Some combinations of factors may give negative customer experience
- Analysis and interpretation is more difficult
- May take longer to set up test

Designs for Multivariable Tests

On-line experiments can simply run overlapping, concurrent, independently randomized experiments

Example: Test 7 factors each at 2 levels

Set up 7 separate experiments to run at the same time with the same users. Get all 128 combinations in the results.

Advantages:

- Easier to implement
- Can turn off one experiment if negative
- Get all interactions

Analysis for Interactions

Procedure for analyzing an MVT for interactions

1. Since there are potentially a vary large number of interactions among the variables being tested, restrict the ones you will look at to a few you suspect may be present. (If 7 factors, 21 two-factor interactions, 35 three-factor interactions, etc.)
2. Conduct the test to determine if the interaction between two factors is present or not
3. **If interaction is not significant, stop!**
If the interaction IS significant, look at the graphical output to interpret.

Analysis for Interactions

Example: Factors from MSN HP illustration

F2: MSNBC news stories

C = Top international

T = Specific to country ID'd

F3 Sports/Money placement

C = same order every day

T = Sports higher on wkends
and Money higher wkdays

Hypothesis tests for interactions similar to main effects
(details omitted)

Example: MVT Experiment on MSN HP

Rachael's flat stomach secret: **OBEY** Friday, February 13, 2009 RSS | Español

Web | MSN | Images | Video | News | Maps

Search: Lincoln speech | Avril Lavigne video | Salmonella

Hotmail | Messenger | My MSN | MSN Directory

Air Tickets/Travel | Food & Entertainment | Investing | Music | Tech & Gadgets

Autos | Games | Lifestyle | News | TV

Careers & Jobs | Green Living | Maps & Directions | Real Estate/Rentals | Weather

City Guides | Health & Fitness | Money | Shopping | White Pages

Dating & Personals | Horoscopes | Movies | Sports | Yellow Pages

Sign out | Make MSN your homepage | Customize your Page

Hotmail | Inbox (409) | Windows Live | Compose | Contacts | Show Mail

Video Highlights

The Life of Lovebirds | Courtship rituals in the wild | Jonas Brothers meet Letterman | The many names of Diddy | How to dress for dating success | Valentine's Day on '30 Rock' | Summers on stimulus plan | View more MSN videos

Custom MSN Content | Select & refine your content | Local News | Stocks | Horoscopes | Weather

Local News | Customize this module on your page to get today's latest news. | Simply enter a city or ZIP code. | Find news by city or ZIP | Go

50 Dead as Plane Slams Into Home | Crash near Buffalo, N.Y., kills all people on flight, one person on ground | 9/11 widow among victims | Video: Cockpit audio released | Photos from the crash scene

Today's Picks | Drew Peterson's fiancée breaks silence: 'He's nice' | Examining 'baby addiction' | Quick & easy cover letter tips

Dreamgirls | Beyoncé & Hudson win Image Awards

msn 'Angels & Demons' | Play 'Path of Illumination': | Puzzle: Time to conquer 'Air' | Help stop an unthinkable crime | What is 'Path of Illumination'? | Photos from 'Angels & Demons' | Win a trip to the movie premiere

Also on MSN | How to know if refinancing makes sense | 4 polar-opposite couples who made it work | 10 chic & affordable designer lines | Will you get spot at Business Fantasy Camp? | Shop for last-minute Valentine's Day flowers

Entertainment | Rocky career moments of the stars | Beyoncé & Hudson win Image Awards | Review: Does 'The International' pack thrills? | Nude photo of Madonna snags \$37,500 | Addison returns to Seattle Grace on 'Grey's'

Find movies, actors and actresses | Go

A-list Searches

Right now, get T-Mobile's best rates, with or without the long-term contract. | learn more | T-Mobile | stick together

Advertisement | Ad feedback

Get the latest security updates

MSNBC News | Recession leaves many working in limbo | Live video: Buffalo TV reports on crash | Congress may vote on stimulus today | Iraq bomber targets Shiite women, kids | Former teen sex slave describes ordeal

F2

Box Sports | Kriegel: Put an * on Steelers' six titles | 10 favorites to win Daytona 500 | Lineup | Selig blasts A-Rod | Time to move on, Bud | Reports: 3 Pro Bowl WRs want to be traded | Most explosive sports couples break-ups

Money | Get a reprieve on your credit cards | Stocks flat as investors await stimulus | 5 stocks for the (eventual) recovery | How 5 traders survived...and 1 thrived | Hoofy & Boo: Juicing your portfolio

Get quote | Go

F3

Dow | 7,872.79 | -59.97 (-0.76%)

Factors/variables

F2: MSNBC news stories

C = Top international

T = Specific to country ID'd

F3: Sports/Money placement

C = Sports above Money

T = Money above Sports

OEC: Clicks per User

Other metrics: PVs, CTR

(This is for illustration purposes only, it does not reflect any previous or planned test on MSN HP)

Graphical Analysis of Interactions

- If hypothesis test for interaction is not significant
 - Assume no interaction present
 - Interaction graph would show lines approximately parallel
- If interaction is statistically significant
 - Plot interaction to interpret

Graphical Analysis of Interactions

Case 1: No Interaction (parallel lines)

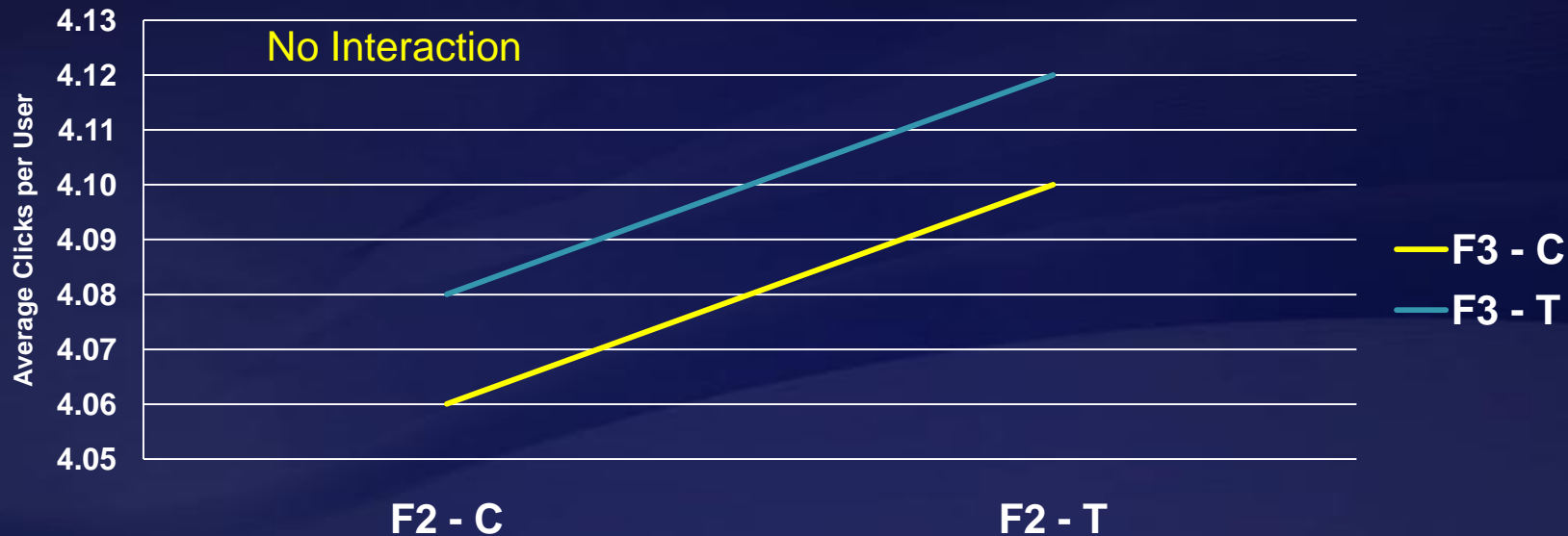
Data Table

	F2 - C	F2 - T
F3 - C	4.06	4.10
F3 - T	4.08	4.12

Main Effects Results

	Pct Effect	p-value
Effect(F2)	0.98%	<.001
Effect(F3)	0.49%	0.032

F2xF3 Interaction



Graphical Analysis of Interactions

- When interaction is statistically significant

Two types of interactions:

- **Synergistic** – when the presence of both is **more** than the sum of the individual treatments
- **Antagonistic** – when the presence of both is **less** than the sum of the individuals

Graphical Analysis of Interactions

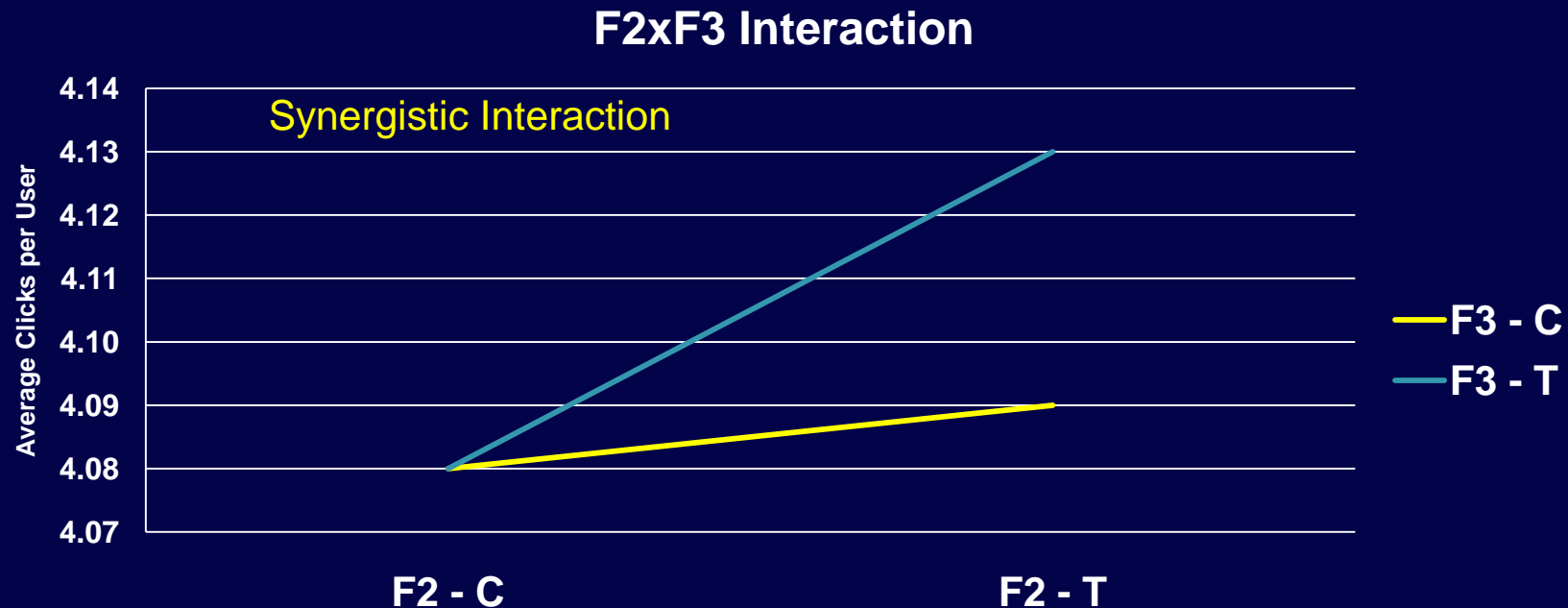
Case 2: Synergistic Interaction

Data Table

	F2 - C	F2 - T
F3 - C	4.08	4.09
F3 - T	4.08	4.13

Main Effects Results

	Pct Effect	p-value
Effect(F2)	0.74%	0.008
Effect(F3)	0.49%	0.032



Graphical Analysis of Interactions

Case 3: Antagonistic Interaction

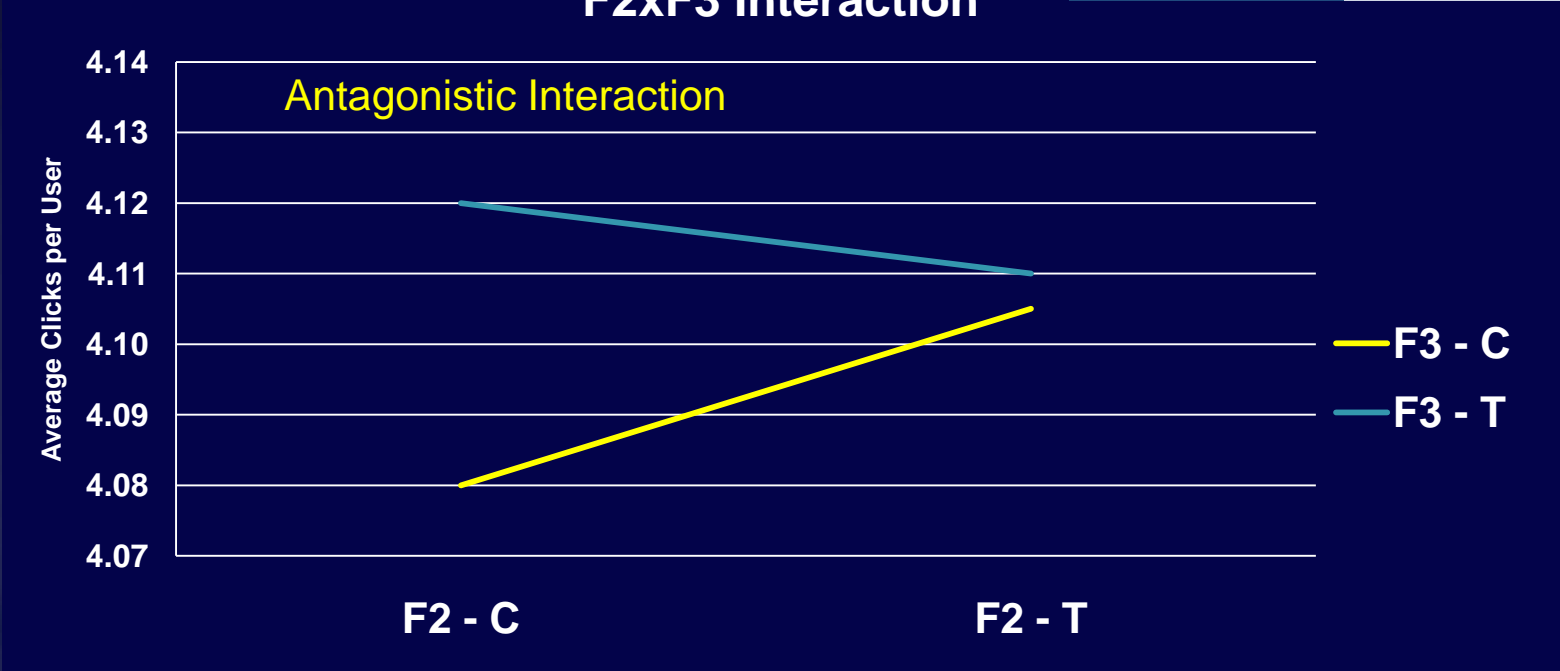
Data Table

	F2 - C	F2 - T
F3 - C	4.08	4.11
F3 - T	4.12	4.11

Main Effects Results

	Pct Effect	p-value
Effect(F2)	0.18%	0.396
Effect(F3)	0.55%	0.028

F2xF3 Interaction



Case Study: EVS Experiment

Current Model

- Pre-roll ad played before first content stream
- Don't disturb users by playing ad when a content stream is playing
- Ad stream played before the content stream when content streams played for more than 180 seconds continuously

Case Study: EVS Experiment (cont.)

Business Questions

- Could removing pro-roll ad stream attract more returning users?
- Could shortening the minimum time between two ad streams attract more returning users?
- Would ad stream gain from returning users offset the loss of not playing pre-roll or playing ad less frequently?

Case Study: EVS Experiment (cont.)

Experiment Design

- Factor 1: Play (Control) or Do Not Play pre-roll
- Factor 2: 5 levels of minimum time between two ad streams
 - 90, 120, 180 (Control), 300, 900 seconds
- Users who received treatments in two week observation window continued to receive treatments and were monitored for the following six weeks for their return rate

Case Study: EVS Experiment (cont.)

Assuming the Overall Evaluation Criterion (OEC) is *Percent of Returning Users*

Vote for result on Factor 1:

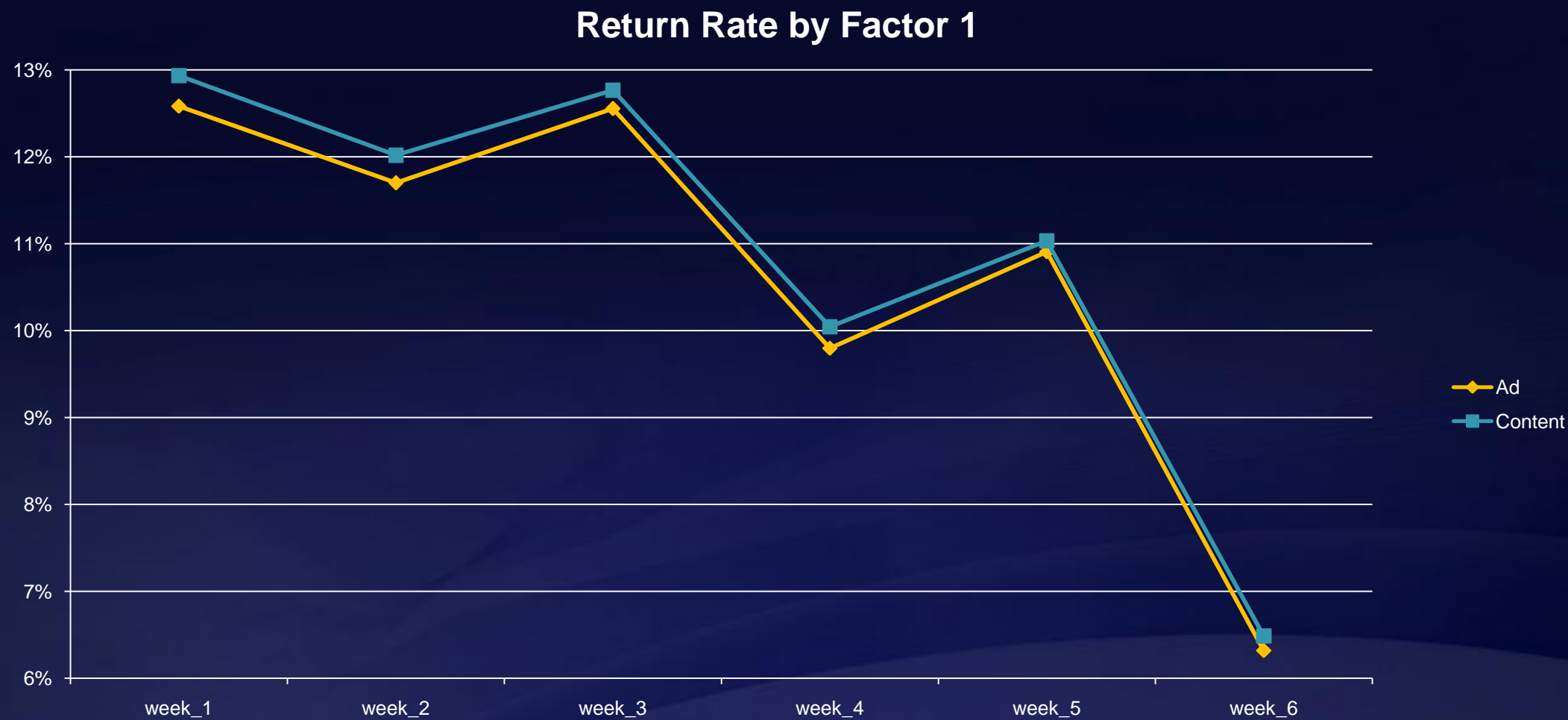
1. Playing pre-roll is statistically significantly better
2. Flat (no statistical difference)
3. Playing pre-roll is statistically significantly worse

Case Study: EVS Experiment (cont.)

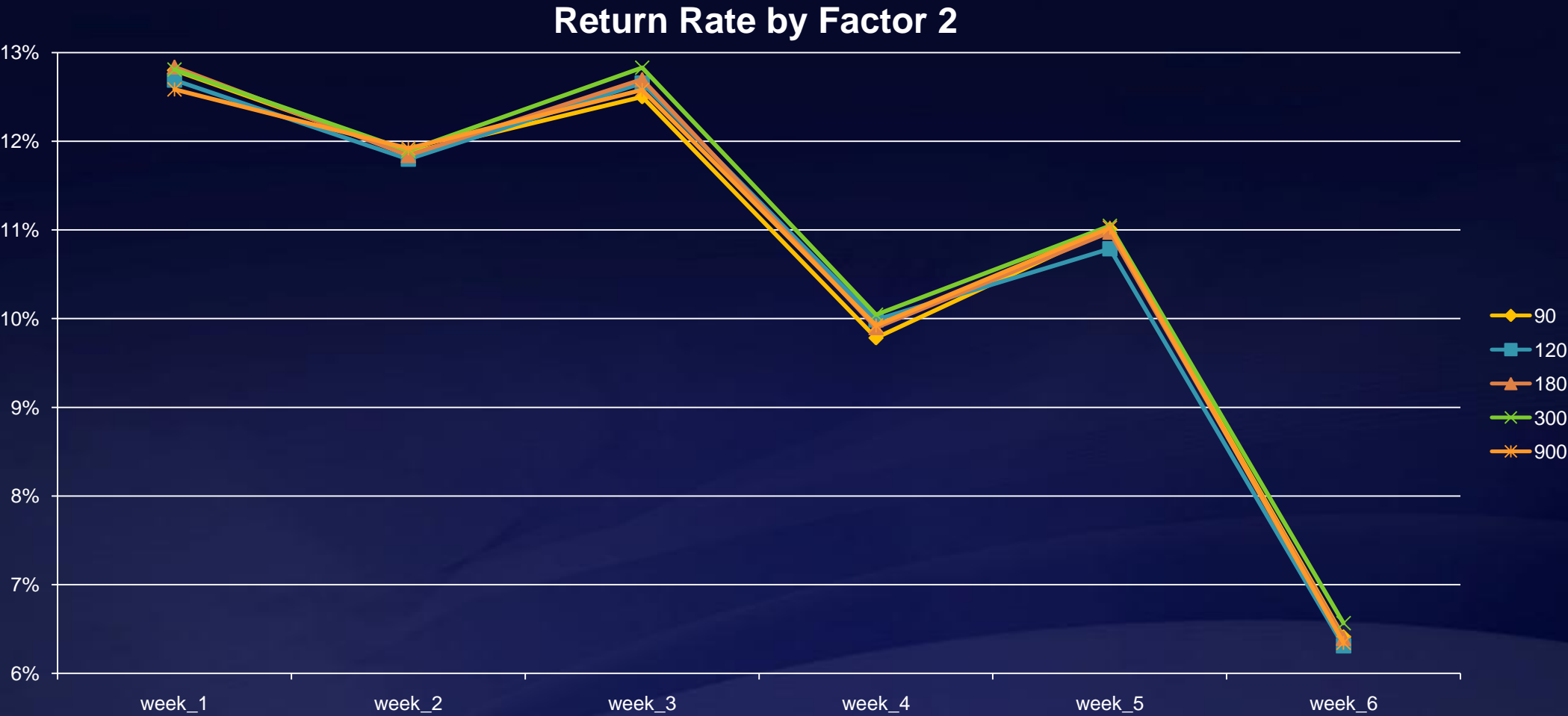
Vote for result on Factor 2: which of the following attract statistically significantly more returning users

1. 90 seconds
2. 120 seconds
3. 180 seconds
4. 300 seconds
5. 900 seconds
6. Flat (no difference)

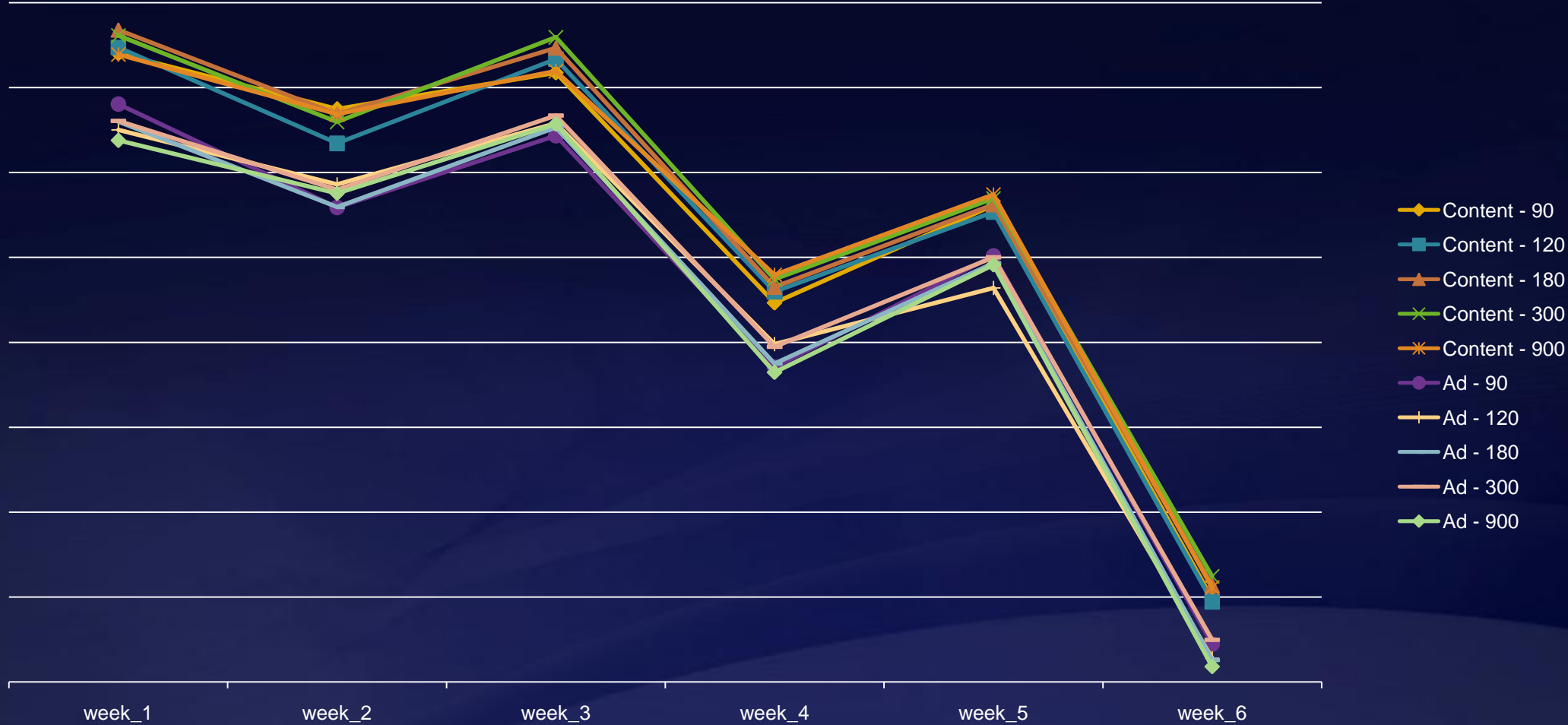
EVS Experiment: Effect on Factor 1



EVS Experiment: Effect on Factor 2



EVS Experiment: Interaction between Factors 1 and 2



Appendix: Challenges and Advanced Statistical Concepts

- Variance calculations for metrics
- Non-parametric alternatives to t-test, ANOVA
- Robot detection
- Automatic detection of interesting population segments
- Experimentation with exploration/exploitation schemes
- Predicting when a metric will be significant

Variance calculations for metrics

- Metrics that are not “per user” currently use bootstrap to estimate variance
 - Can we get a formula to take into account correlation of experimental units?
 - Example: Clickthrough rate (CTR) per experiment

$$\text{CTR} = \frac{\text{Total Clicks}}{\text{Total Impressions}}$$

True variance is much larger than that from Binomial distribution

Non-parametric alternatives to t-test, ANOVA

- Permutation or Mann-Whitney tests are natural
- Pros
 - Can get a p-value
 - May have better power for some metrics
 - Works better for small sample sizes
- Cons
 - Understandability by business managers
 - Can be computationally intensive
 - Confidence intervals for effect not straight-forward

Robot filtering

- What is “best” way to develop heuristics to detect robots?
- What is “best” way to assess how well heuristics are doing?
- How to adjust robot detection parameters based on site in the test?

For example

- Sites with low traffic may need more aggressive robot filtering
- Sites that expect active users (e.g. many clicks per hour) need less aggressive robot filtering
- Sites that have more robot traffic may need more aggressive robot filtering

Automatic detection of interesting population segments

- A population segment is interesting if their response to the Treatment is different from the overall response
- Segments can be defined by a number of variables
 - Browser or operating system
 - Referrer (e.g. from search engine, etc.)
 - Signed-in status
 - Loyalty
 - Demographics
 - Location – country, state, size of city (use IP lookup)
 - Bandwidth

Experimentation with exploration/exploitation schemes

- Want to automatically display best content based on exploration/exploitation strategy
- Is this strategy better than editor-placed content?
- What are the optimal parameter values?
 - Percent in exploration group?
 - How long to test content in exploration group?
 - What level of significance is needed?

Predicting when a metric will be significant

- After experiment has run for some period of time and have estimates of effect and standard deviation can we give a helpful estimate of how long experiment needs to run in order to get a significant result for a particular metric?
 - Statistical philosophical issues
 - Technical issues