Microsoft

Testing with Real Users

User Interaction and Beyond, with Online Experimentation

Seth Eliot, Senior Test Manager
Experimentation Platform (ExP)

Better Software Conference - June 9, 2010

Introduction

- What is Online Controlled Experimentation?
- Employing Online Experimentation
- Data Driven Decision Making
- How does this apply to SQA?
  - Rapid Prototyping
  - Exposure Control
  - Monitoring & Measurement
  - Testing in Production (TiP)
  - Services TiP with Online Experimentation
  - Services TiP with Shadowing
  - Complex Measurements

Latest version of this slide deck can be found at:
http://exp-platform.com/bsc2010.aspx

Who am I?

Software QA Manager

Amazon Digital Media

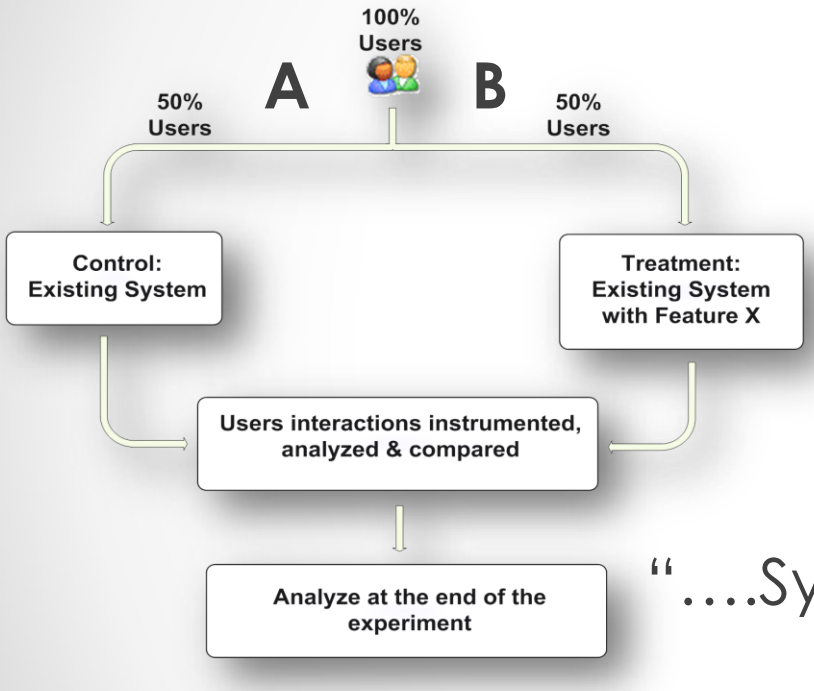
Microsoft Experimentation Platform

Culture Shift

- Services
- Data Driven

What is Online Controlled Experimentation?

Online Controlled Experimentation, Simple Example



This is an "**A/B**" test

…the simplest example

- A and B are **Variants**
  - A is **Control**
  - B is **Treatment**

"….System with Feature X"

can be

"….Website with Different UX"

…and What it's Not.

User KNOWS he is in an experiment

Result is which one he THINKS he likes better

Opt-in (biased population)

User Tries ALL the variants

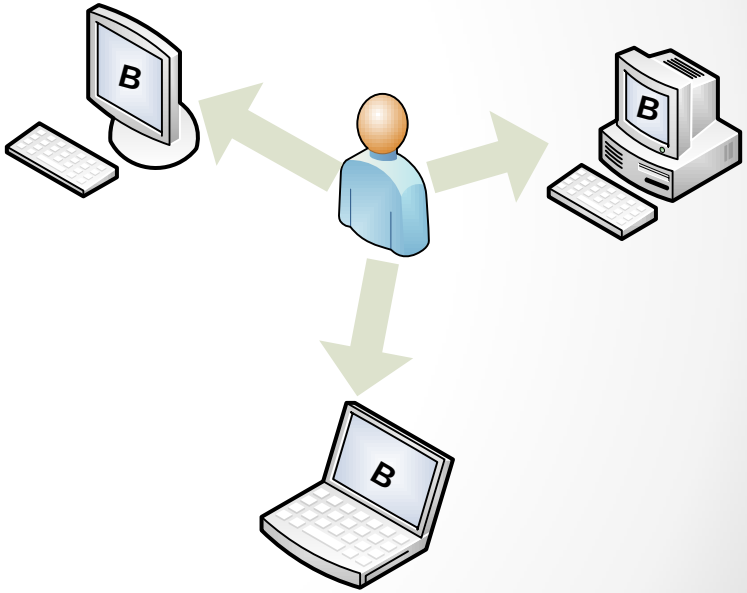
User's goal IS the experiment

What makes a "controlled" experiment?

Nothing but the variants should influence the results

- Variants run simultaneously
- Users do not know they are in an experiment
- User assignment is random and unbiased

….and Sticky



Why are controlled experiments trustworthy?

- Best scientific way to prove causality
  - changes in metrics are caused by changes introduced in the treatment(s)

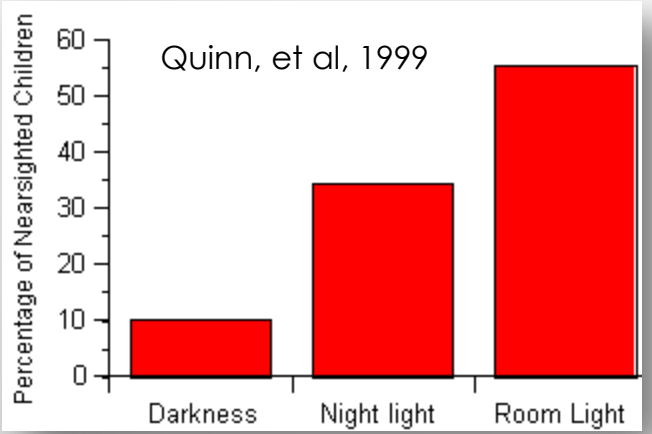
Why are controlled experiments trustworthy?

- Best scientific way to prove causality
  - changes in metrics are caused by changes introduced in the treatment(s)

Correlation Does not Imply Causation

Higher Kindle Sales correlate with deployment of B

Did Website B cause the sales increase?



Do night-lights cause near-sightedness in children?

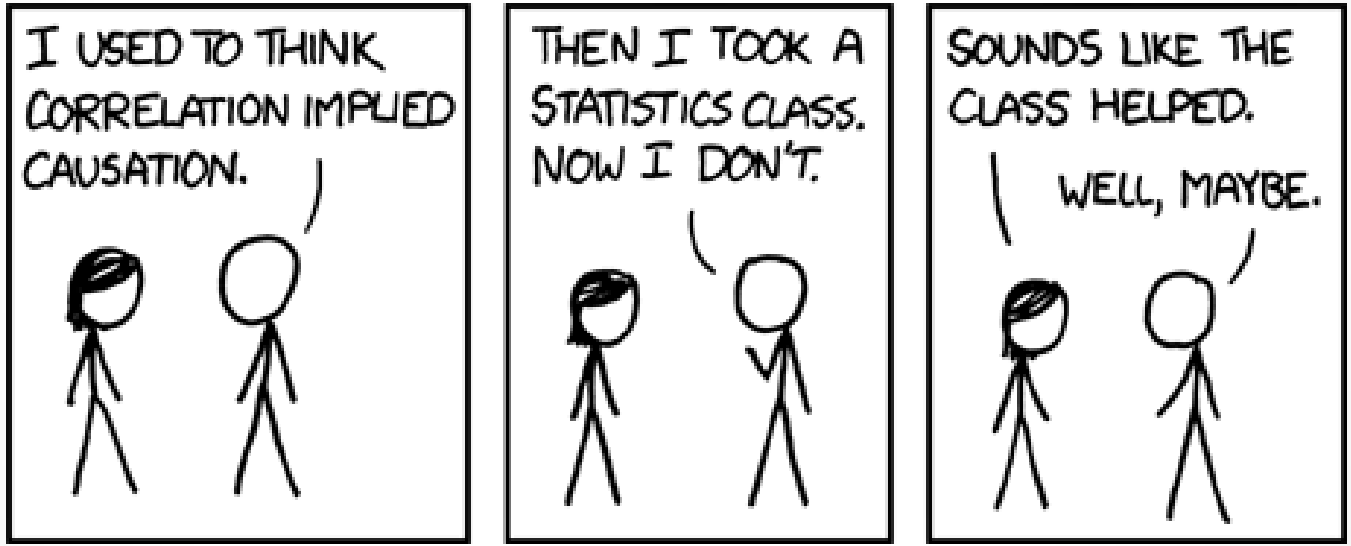
Nope. Near-sighted parents do

[Zadnik, et al, 2000]

xkcd

A WEBCOMIC OF ROMANCE, SARCASM, MATH, AND LANGUAGE.

Correlation

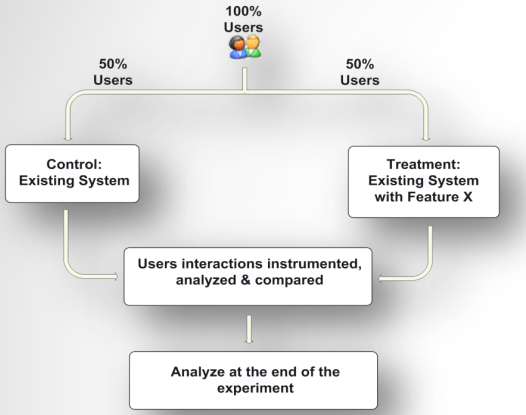


http://xkcd.com/552/

XKCD

Employing Online Experimentation

Where can Online Experimentation be used?



"….System with Feature X"

can

"…V

System

- Website
- Service

Feature X

- Different UX
- Different functionality
- Vcurr/Vnext
- Platform Change/Upgrade

Platform for Online Experimentation

Platforms used Internally

amazon.com WebLab

Microsoft EXP

"design philosophy was governed by data and data exclusively" – Douglas Bowman, Visual Design Lead [Goodbye, Google, Mar 2009]

Google

Public Platforms

Google Website Optimizer

OMNITURE Test & Target Campaign

Adobe

webtrends

widemile

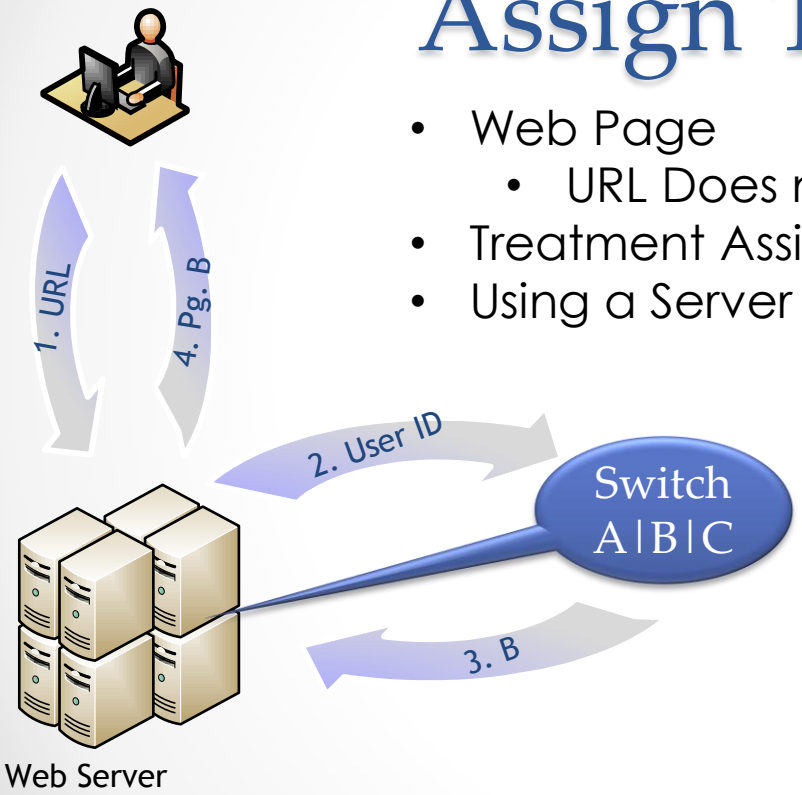
Nuts and Bolts of Online Experimentation

1. **Assign Treatment**

2. **Record Observation(s)**

3. **Analyze and Compare**

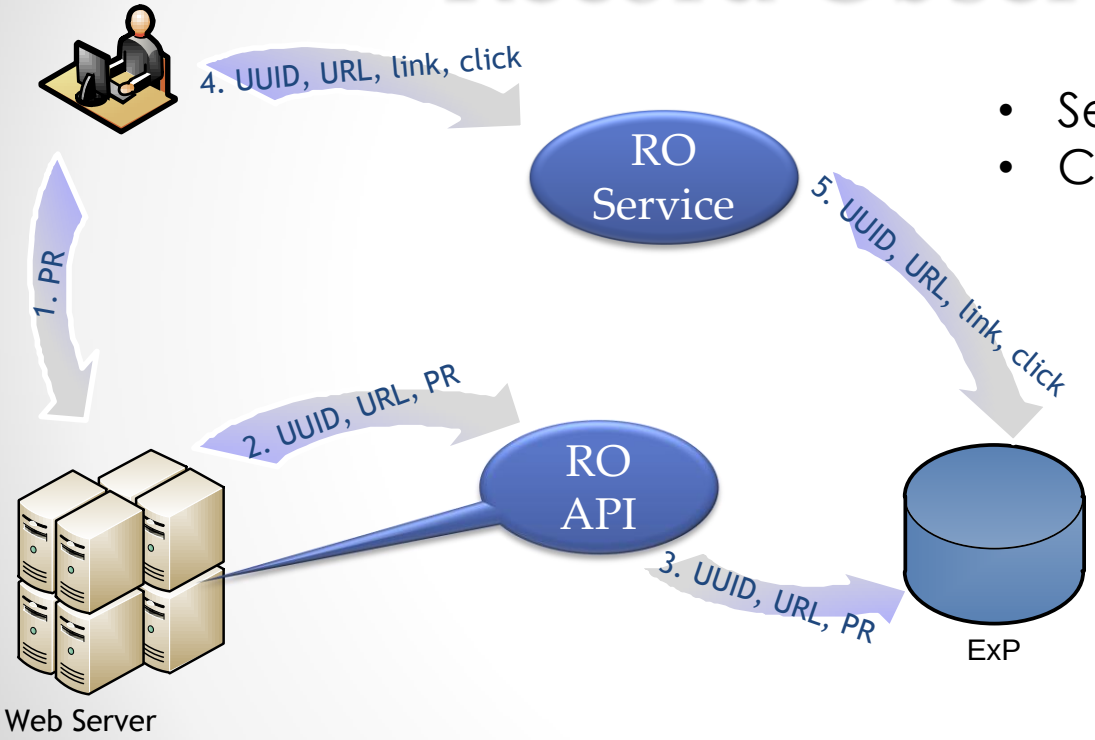
An Experiment Architecture: Assign Treatment



- Web Page
  - URL Does not change
- Treatment Assignment
- Using a Server Side Switch

- Instead of a Web Page could be
  - Code Path
  - Service Selection
  - V-curr / V-next

An Experiment Architecture: Record Observation



- Server-side Observations
- Client-side Observations
  - Require Instrumented Page

PR = Page Request
UUID = Unique User ID
RO = Record Observation

Analyze & Compare

Details

Refresh | Edit | Copy

		Expected Split	Actual Traffic
C:	C: Control	12.50 %	17,420,461 unique users
T1:	T1	12.50 %	17,422,129 unique users
T2:	T2	12.50 %	17,414,591 unique users
T3:	T3	12.50 %	17,414,324 unique users
T4:	T4	12.50 %	17,418,765 unique users
T5:	T5	12.50 %	17,416,762 unique users
T6:	T6	12.50 %	17,417,993 unique users
T7:	T7	12.50 %	17,407,350 unique users

Description:

Test re-ordering of Nav menu items on MSN HP

Metrics

Export | View All Metrics | Exclude Dates

Report: Final Report

Improvement of: T2 over C

Analysis for: 8/4/2009 thru 8/31/2009

Excludes: 8/7/2009-8/10/2009 , 8/25/2009-8/26/2009 , 8/28/2009-8/29/2009

Last refresh: **9/1/2009 1:00:00 AM**

Name	T2		C		P Value	% Change	Signif.
	Total	Avg	Total	Avg			
Bing Searches - Clicks (%) per User	746,787	4.628 %	715,680	4.477 %	< 0.001	3.39 %	Yes
Bing Searches - Clicks (all) per User	746,787	0.100	715,680	0.096	< 0.001	4.34 %	Yes
Entertainment - Clicks (%) per User	1,176,979	5.020 %	1,162,892	4.945 %	< 0.001	1.53 %	Yes
Entertainment - Clicks (all) per User	1,176,979	0.157	1,162,892	0.155	0.007	1.20 %	Yes
Food - Clicks (all) per User	53	0.000	15,013	0.002	< 0.001	-99.65 %	Yes
Footer Search - Clicks (%) per User	14,475	0.087 %	13,103	0.072 %	< 0.001	20.61 %	Yes
Footer Search - Clicks (all) per User	14,475	0.002	13,103	0.002	0.073	10.46 %	No

More Details

Bing Searches - Clicks (%) per User

Absolute change: **0.152%**

Confidence interval: **(0.131%, 0.173%)**

Percent change: **3.39 %**

Confidence interval: **(2.91 %, 3.87 %)**

Analyze & Compare

Data Driven Decision Making

Example: Amazon Shopping Cart Recs

- Amazon.com engineer had the idea of showing recommendations based on cart items [Greg Linden, Apr 2006]
  - Pro: cross-sell more items (increase average basket size)
  - Con: distract people from checking out (reduce conversion)
- A marketing senior vice-president was dead set against it.
- Ran an Experiment…

Introducing the HiPPO

- **Hi**ghest **P**aid **P**erson's **O**pinion
- A marketing person's opinion was dead set against it.
- Highest Paid Person's Opinion

"A scientific man ought to have no wishes, no affections, - a mere heart of stone." - Charles Darwin

Data Trumps Intuition

- Based on experiments with ExP at Microsoft

1/3	1/3	1/3
Positive Ideas	No Statistical Difference	Negative Ideas

- Our intuition is poor:
  - 2/3rd of ideas do not improve the metric(s) they were designed to improve

"It's amazing what you can see when you look"
Yogi Berra

A Different Way of Thinking

- Avoid the temptation to try and build optimal features through extensive planning without early testing.

- Try radical ideas. You may be surprised, especially if "cheap"

  i.e. Amazon.com shopping cart recs

Example: Microsoft Xbox Live

Goal: Sign More People up for Gold Subscriptions

A

B

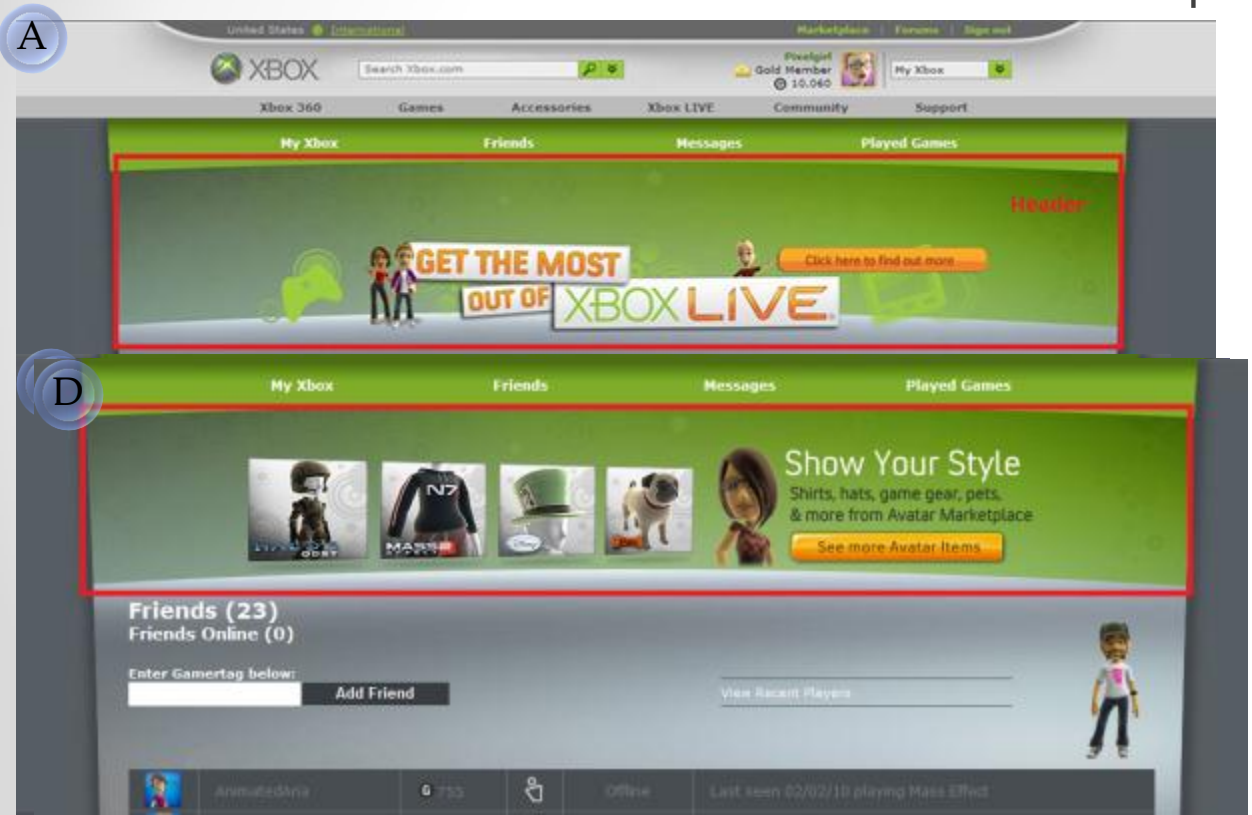
http://www.xbox.com/en-US/live/joinlive.htm

Which has higher Gold Sign-up…???

A. Control
B. Treatment – up 29.9%
C. Neither

Example: Microsoft Xbox Marketplace

Goal: Increase Total Points Spent per User



http://marketplace.xbox.com/en-US

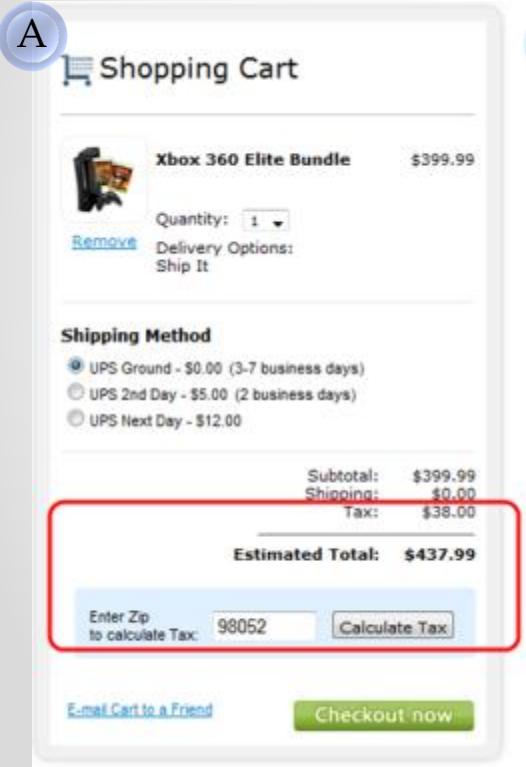
Which has higher Points Spent…???

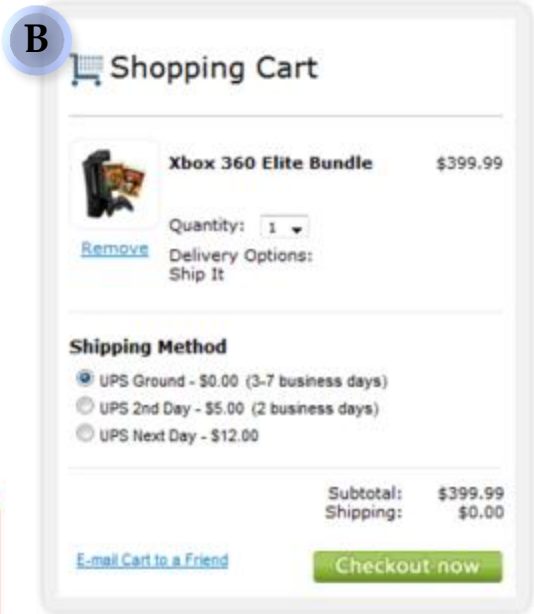
A. Control
B. T1: Game Add-Ons
C. T2: Game Demo
D. T3: Avatar Gear
E. None

Promoted content up, but at expense of others

Example: Microsoft Store

Goal: Increase Average Revenue per User





http://store.microsoft.com/home.aspx

Which increased revenue…?

A. Control

B. Treatment – up 3.3%

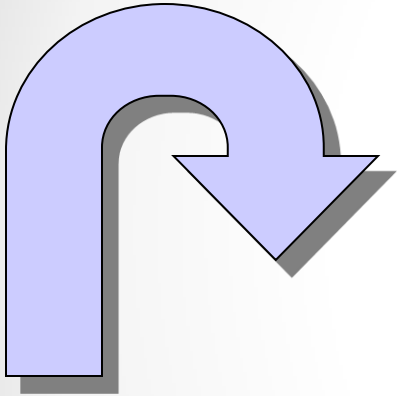
C. Neither

How Does This Apply to SQA?

...

Online Experimentation Used for SQA…

…or more specifically, Software Testing



- Meeting Business Requirements = Quality?
  - Sure, But QA not often involved in User Experience testing
- Experimentation Platform enables Testing in Production (TiP)
  - Yes, I mean Software QA Testing

How Does This Apply to SQA?

…

Rapid Prototyping

Test Early, Test Often

"To have a great idea, have a lot of them" -- Thomas Edison

"If you have to kiss a lot of frogs to find a prince, find more frogs and kiss them faster and faster" -- Mike Moran, Do it Wrong Quickly

- Replace BUFT (Big UpFront Test) with "Smaller" Testing and TiP
- …and Iteration

Rapid Prototyping to Reduce Test Cost

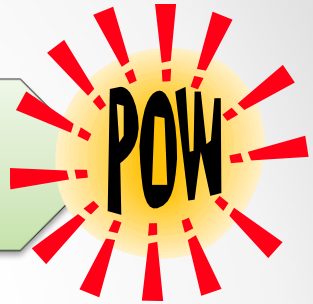
- UpFront Test your web application or site for only a subset (or one) browser
- Release to only that subset of browsers
- Evaluate results with real users
- Adjust and Add another browser

  or
- Abort

Enabled by ExP

Rapid Prototyping

Big Scary New Code → BUFT → Release "Safe" for Everyone → oops a bug… Scramble!



Limit impact of potential problems

Big Scary New Code → Small UFT → Release to segment of users → Monitor & Fix → Ramp to 100%

Saves you from having to BUFT if product is a dud

Big Scary New Code → Small UFT → Release to segment of users → Bad Idea → Move on to a new idea

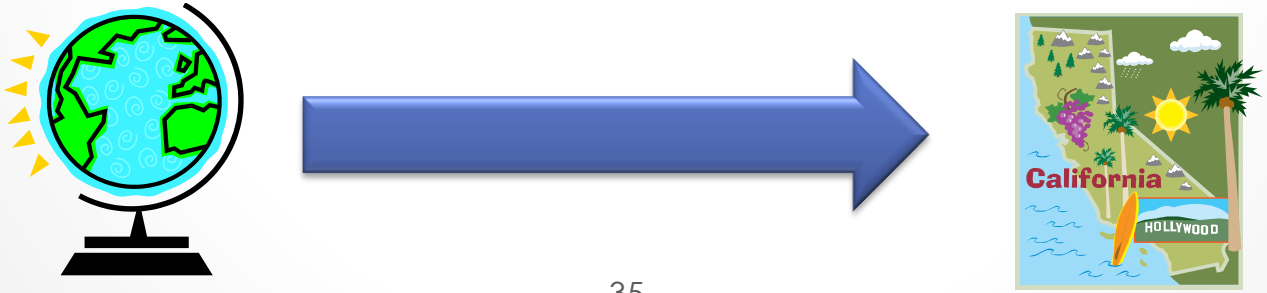
How Does This Apply to SQA?

…

Exposure Control

Rapid Prototyping utilizes Exposure Control

…to limit the **Diversity** of Users exposed to the code

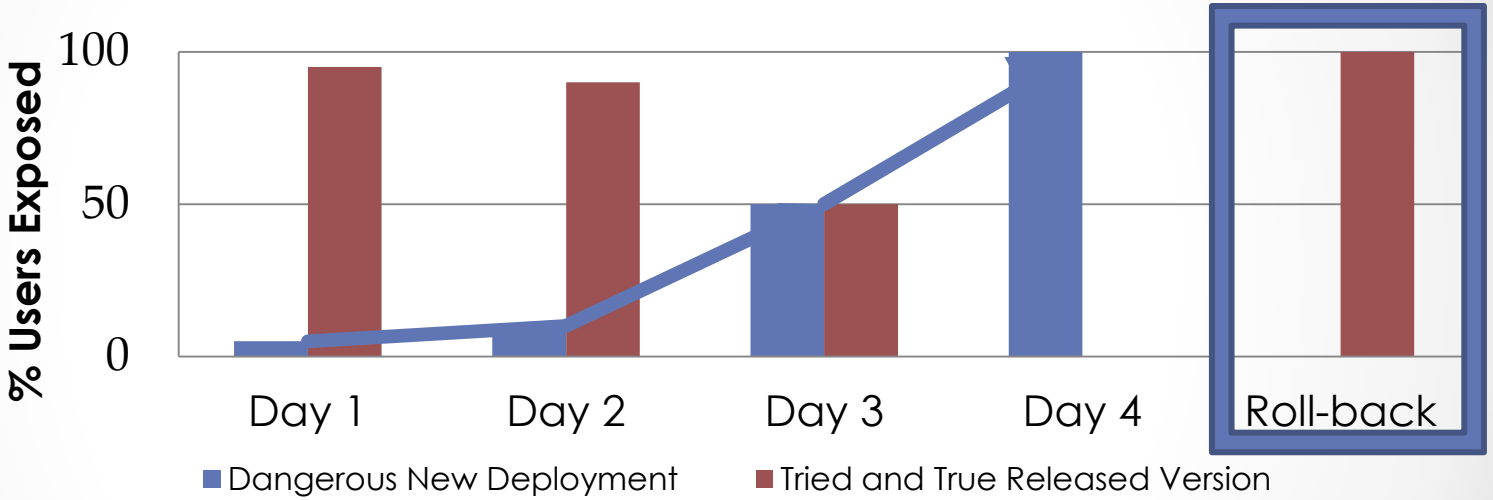


Exposure Control to limit Diversity

- Other filters also
  - ExP can do this.
    - Location based on IP
    - Time of day
  - Amazon can do this
    - Corporate affiliation based on IP

- Still random and unbiased.
  - Exposure control only determines in or out.
  - If in the experiment, then still random and unbiased.

Exposure Control to Limit Scale



Control how many users see your new and dangerous code

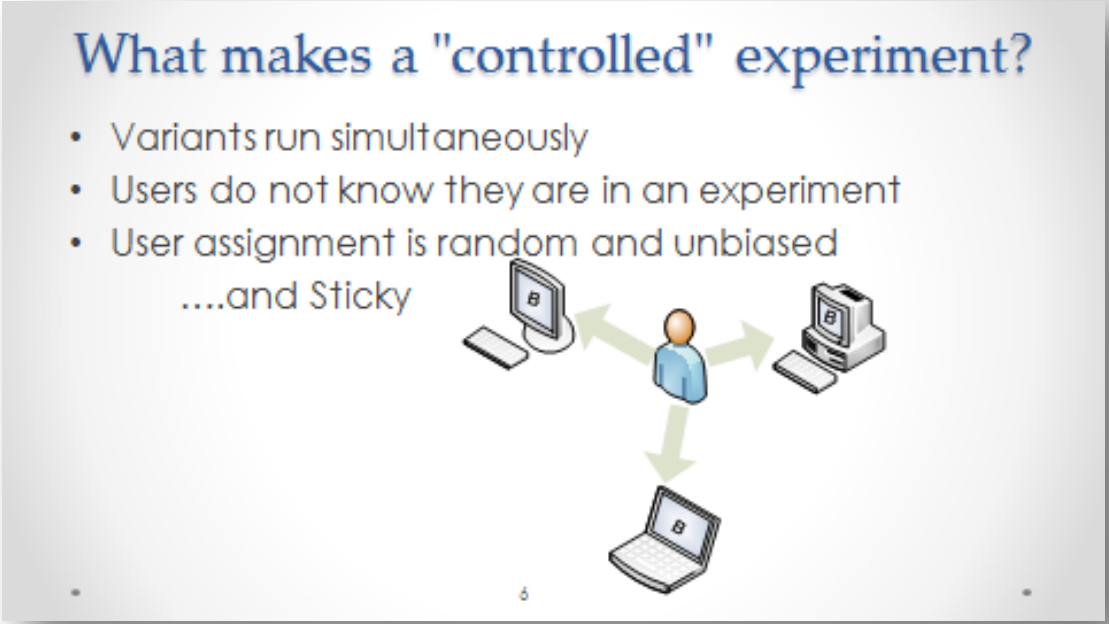
Example: Ramp-up and Deployment: IMVU

"Meet New People in 3-D"

- [v-next is deployed to] a small subset of the machines throwing the code live to its first few customers
- if there has been a statistically significant regression then the revision is automatically rolled back.
- If not, then it gets pushed to 100% of the cluster and monitored in the same way for another five minutes.
- This whole process is simple enough that it's implemented by a handful of shell scripts. [Timothy Fitz, Feb 2009]

Important Properties of Exposure Control

- Easy Ramp-up and Roll-back
- Controlled Experiment



How Does This Apply to SQA?

…

Monitoring and Measurement

Experiment Observations

- Website/UX Observations
  - Client Side: Page View (PV), Click
  - Server Side: Page Request (PR)

- Service Observations
  - Client Side
    - If there is a client, then client side results can indicate server side issues
  - Server Side
    - Service Latency
    - Server performance (CPU, Memory) if variants on different servers
    - Number of requests

Experiment Metrics

Compare means of your variant population

- CTR per user
  - CTR: % Users who Click on monitored link of those who had Page Views (PV) including that link (impression)
- ExP Xbox Gold Membership
  - % of Users with PV on US Xbox JoinLive page who had a PV on Gold "congrats" page.
- ExP Microsoft Store
  - Mean Order Total ($) per User
  - Observations can have data (e.g. Shopping Cart Total $)
- Amazon Shopping Cart Reccomendations [?]
  - % users who purchase recco items of those who visit checkout, or
  - average revenue per user
- Google Website Optimizer
  - Conversion Rate: % of users with PV on Page[A] or Page[B] who had a PV on Page[convert]

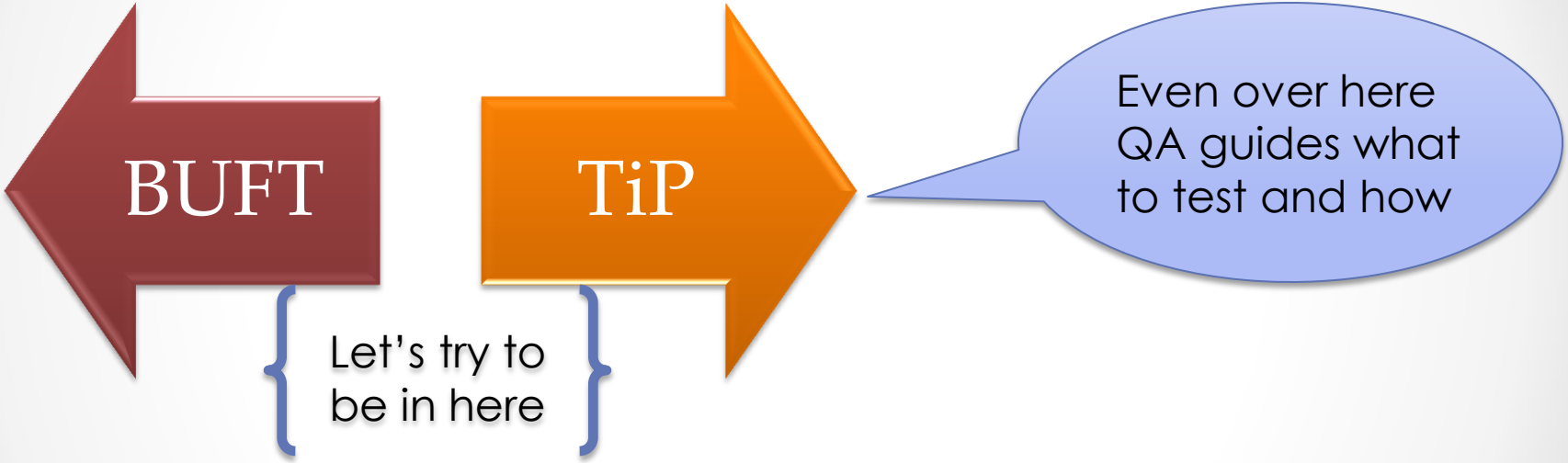
How Does This Apply to SQA?

…

Testing in Production (TiP)

Exposure control + Monitoring & Measurement = TiP

"Fire the Test team and put it in production…"?



Leverage the long tail of production, but be smart and mitigate risk.

Testing in Production (TiP)

TiP can be used with Services (includes Websites)

- Testing
  - Functional and Non-Functional
- Production
  - Data Center where V-curr runs
  - Real world user traffic

What is a Service?

- You control the deployment independent of user action.
- You have direct monitoring access.

Deploy, Detect, Patch

"It is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change." - Charles Darwin

Examples:

- Google: All engineers have access to the production machines: "…deploy, configure, monitor, debug, and maintain" [Google Talk, June 2007 @ 21:00]
- Amazon: Apollo Deployment System, PMET Monitoring System - company wide supported frameworks for all services.

TiP is Not New

amazon.com Prime

Hello, Seth Eliot. We have recommendations for you. (Not Seth?)

Kindle: The #1 Bestseller on Amazon

Seth's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards | Your Account | Help

Shop All Departments | Search Movies & TV | GO | Cart | Your Lists

Movies & TV | Advanced Search | Browse Genres | New Releases | Bestsellers | DVD & Blu-Ray Deals | TV Shows | Blu-Ray | Video On Demand

Prime
Member: Seth Eliot

Seth Eliot: This item is eligible for Amazon Prime. Click here to turn on 1-Click and make Prime even better for you.
(With 1-Click enabled, you can always use the regular shopping cart as well.)

RDD DVD Test ASIN 2

Format: DVD

No customer reviews yet. Be the first.

List Price: ~~$19.99~~
Price: $17.99 & eligible for free shipping with **Amazon Prime**
You Save: $2.00 (10%)

Temporarily out of stock.
Order now and we'll deliver when available. We'll e-mail you with an estimated delivery date as soon as we have more information. Your account will only be charged when we ship the item.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Quantity: 1
Add to Cart
or
Sign in to turn on 1-Click ordering.
Add to Wish List
Add to Shopping List

Express Checkout with PayPhrase
Your PayPhrases

Share

TiP is Not New

SEARCH

BUY PLAY

98074

Enter your City and State/Province or Postal Code

New! Search using event start and end dates.

SEARCH OPTIONS

Wizards Play Network

Search for WPN locations offering:

- Magic: The Gathering
  - Pre-Releases, Launch Parties, Game Days
  - Friday Night Magic
  - Pro Tour Qualifiers
  - Grand Prix Trials
  - National Qualifiers
  - Other Magic Events
- Dungeons & Dragons
  - Encounters
  - Worldwide D&D Game Day
  - Other D&D Events
- Other Products

Start New Search GET MAP

Results 1 - 2 of about 2

Renticon
15612 SE 128th St (8.7 miles)
Renton, WA 98059
Get Directions: To Here - From Here

Test store - USA
1600 Lind Ave (11.8 miles)
Renton, WA 98057
Get Directions: To Here - From Here

Event Link

Test store - USA

1600 Lind Ave (11.8 miles)
Renton, WA 98057
United States

Get Directions: To Here - From Here

Events

Name: Dungeons & Dragons WPN: D&D WPN-2010-02-20-Renton
Format: D&D WPN
Phone Number: 1 (555) 555-5555
Coordinator: tom.ko@wizards.com
Date: 2/20/2010

Other Events by this Location

Name: D&D Encounters: Undermountain
Format: D&D WPN
Phone Number: 1 (425) 687-2135
Coordinator: brian.larabee@wizards.com

Add: Immunization

*=Required field

Type of immunization *

- Select -
meningococcal vaccine, NOS
mumps virus vaccine
no vaccine administered
parainfluenza-3 virus vaccine
pertussis vaccine
plague vaccine
pneumococcal conjugate vaccine, polyvalent
pneumococcal polysaccharide vaccine
pneumococcal vaccine, NOS
poliovirus vaccine, inactivated
poliovirus vaccine, live, oral
poliovirus vaccine, NOS
Q fever vaccine
rabies immune globulin
rabies vaccine, for intradermal injection
rabies vaccine, for intramuscular injection
rabies vaccine, NOS
RESERVED - do not use
respiratory syncytial virus immune globulin, intravenous
respiratory syncytial virus monoclonal antibody (palivizumab), intramuscular
rheumatic fever vaccine
Rift Valley fever vaccine
rotavirus vaccine, NOS
rotavirus, live, monovalent vaccine
rotavirus, live, pentavalent vaccine
rotavirus, live, tetravalent vaccine
rubella and mumps virus vaccine

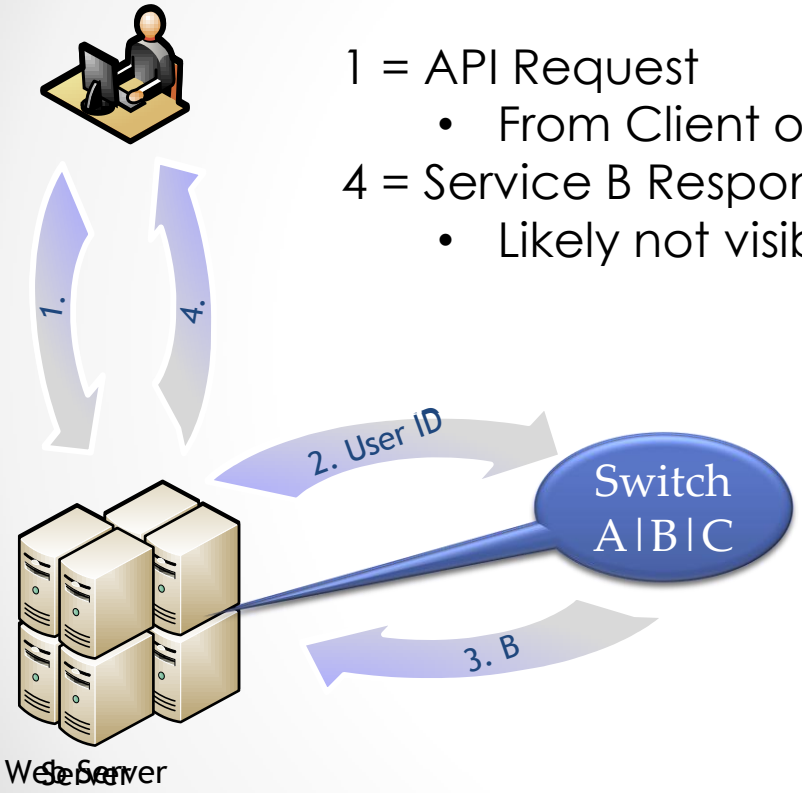
But leveraging it as a legitimate Test Methodology may be new…let's do this right

How Does This Apply to SQA?

…

Testing Services (not Websites)

How can we Experiment with Services?



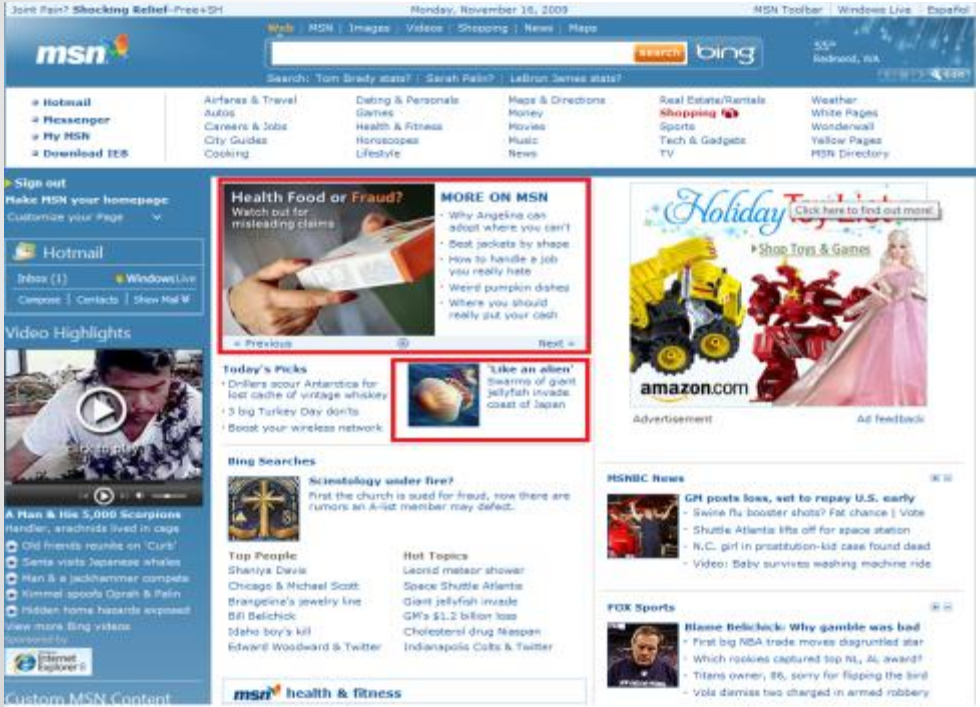
1 = API Request
- From Client or Another Service in the Stack

4 = Service B Response
- Likely not visible to the user

- Microsoft ExP can do this
- So can Amazon WebLab
- Public Platforms Cannot
  - Their "Switch" is Client-Side JavaScript

Example: MSN HOPS

Goal: Increase Clicks on Page per User via Headline Optimization



Which has higher page clicks per user…???

A. Control - Editor Selected
B. **Treatment – HOPS +2.8%**
C. Neither

- and +7% to +28% increase in clicks on modules per user
- but -0.3% to -2.2% cannibalization elsewhere

Example: Amazon ordering pipeline

- Amazon's ordering pipeline (checkout) systems were migrated to a new platform.
- Team had tested and was going to launch.
- Quality advocates asked for a limited user test using Exposure Control.
- Five Launches and Five Experiments until A=B (showed no difference.)
- The cost had it launched initially to the 100% users could have easily been in the millions of dollars of lost orders.

Example: Google Talk

- Use an "Experimentation Framework"
- Limit launch to
  - Explicit People
  - Just Googlers
  - Percent of all users
- Not just features, but it could be a new caching scheme

[Google Talk, June 2007 @ 20:35]

How Does This Apply to SQA?

…

Services TiP with Shadowing

What is Shadowing?

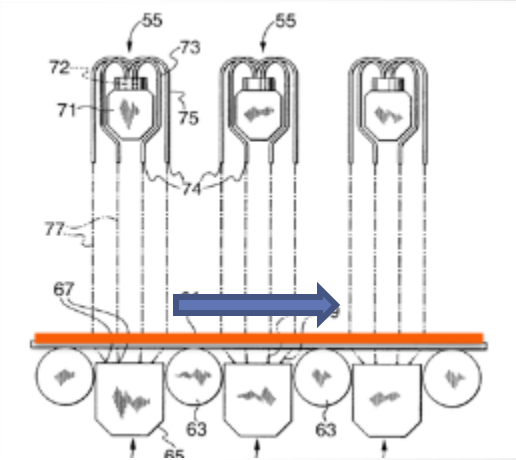
- TiP Technique
- Like ramp-up use real user data in real-time, but mitigate risk by not exposing results to the user
- The ultimate unbiased population assignment
- Controlled experiment
- A+B instead of A/B

Example: ExP RO Shadowing

- RO = RecordObservation, a REST Service for client-side observations.
- Migrate to a new platform.
- Send all observations to BOTH systems via dual beacons.
- Saw Differences – Fixed Bugs.
- Controlled Experiment: both in same Data Center
  - if not, then network introduces bias

Example: USS Cooling System Shadowing

- Based on steel alloy, input speed and temperature, determine number of laminar flows needed to hit target temperature.
- System A: A Human Operator
- System B: An Adaptive Automation
- B has no control, just learn until matches operator.



Example: Google Talk Shadowing

- Google Talk Server provides Presence Status
  - Billions of packets per day

- Orkut integration
  - Started fetching presence without showing anything in UI for weeks before launch
  - Ramp-up slowly from 1% of Orkut PVs
- GMail chat integration:
  - Users logged in/out: used this data to trigger presence status changes w/o showing anything on the UI

[Google Talk, June 2007 @ 9:00]

How Does This Apply to SQA?

…

The Power of Complex Measurements

TTG at Microsoft

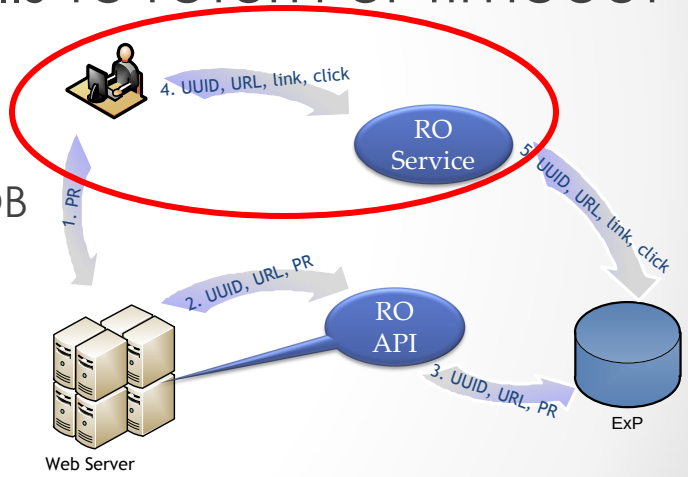
- Use of Experimentation Platform for Complex Measurements
- TTG = Time To Glass
  - "PLT" with a real population over all browsers and bandwidths
  - Includes Browser Render Time
- Calculate TTG from Observations
  - Onload - PageRequest = TTG
- Can analyze results by Browser, Region, etc
  - But Correlation does not imply Causation

Better than monitoring tools like Gomez/Keynote

Form Tracking at Microsoft

- Submit a form (or click a link) and send a beacon to a tracking system and ExP.
- Wait a fixed time or wait for calls to return or timeout (OOB)
- Experiment
  - Variants: Different Wait Times, Fixed vs. OOB
  - Metric: % Data Lost per submit
- Longer time should mean

  Less Data Loss
- Yes, but…..



Resources

...

More Information

- seth.eliot@microsoft.com
- Seth's Blog: http://blogs.msdn.com/seliot/
- ExP Website: http://exp-platform.com

References

Quinn, et al,1999
Quinn GE, Shin CH, Maguire MG, Stone RA (May 1999). "Myopia and ambient lighting at night". *Nature* **399** (6732): 113–4. doi:10.1038/20094. PMID 10335839.

Zadnik, et al, 2000
Zadnik K, Jones LA, Irvin BC, *et al.* (March 2000). "Myopia and ambient night-time lighting". *Nature* **404** (6774): 143–4. doi:10.1038/35004661. PMID 10724157.

Goodbye, Google, Mar 2009
http://stopdesign.com/archive/2009/03/20/goodbye-google.html)

Greg Linden, Apr 2006
Greg Linden's Blog: http://glinden.blogspot.com/2006/04/early-amazon-shopping-cart.html

Timothy Fitz, Feb 2009
IMVU, Continuous Deployment at IMVU: Doing the impossible fifty times a day,
http://timothyfitz.wordpress.com/2009/02/10/continuous-deployment-at-imvu-doing-the-impossible-fifty-times-a-day/

Google Talk, June 2007
Google: Seattle Conference on Scalability: Lessons In Building Scalable Systems, Reza Behforooz
http://video.google.com/videoplay?docid=6202268628085731280

END

BW4. Testing with Real Users
Seth Eliot

Thank you