STP SOFTWARE TEST PROFESSIONALS CONFERENCE SPRING 2012

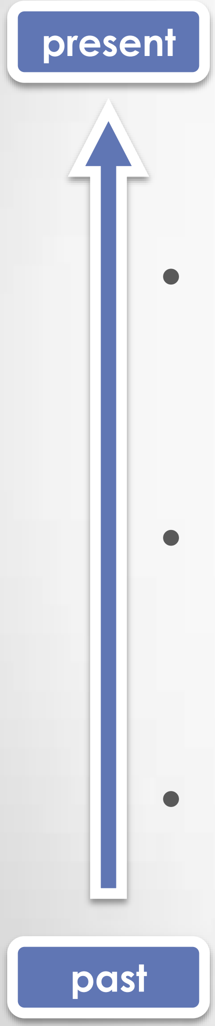
Testing in Production A to Z

TiP Methodologies, Techniques, and Examples

Seth Eliot, Senior Knowledge Engineer,
Test Excellence
Microsoft

Software Test Professionals, Spring – March 28, 2012

About Seth



Microsoft Engineering Excellence

- Best practices for services and cloud

- Bing “Cosmos”
    - Massive, distributed, data processing service
- Microsoft Experimentation Platform
    - Data Driven Decision Making
- Amazon.com Digital Media
    - Video, Music, and Kindle eBook services

TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Killing production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

What is TiP?

Testing in Production

TiP \tip\

- **Noun**: TiP is a set of software testing methodologies that utilizes real users and/or live environments to leverage the diversity of production while mitigating risks to end users.

- **Verb** [trans.]: TiP, TiPed,
  TiPing

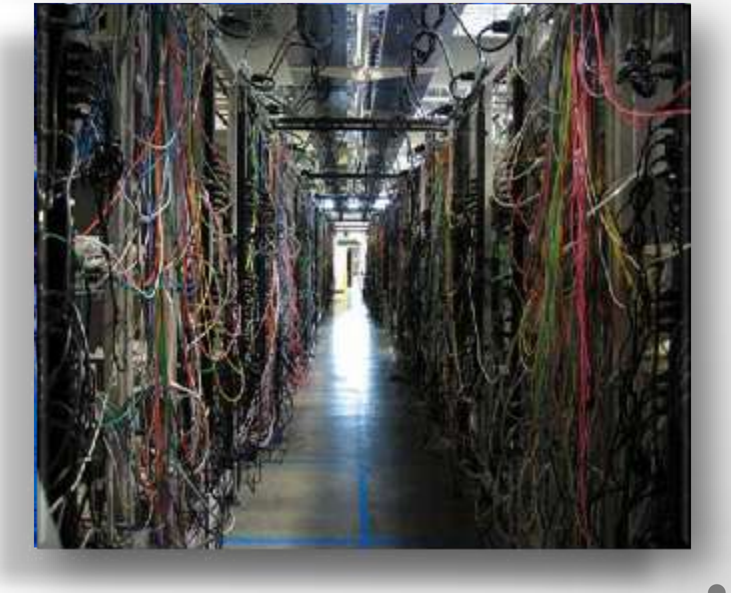
Tester Mindshift

I DON'T ALWAYS TEST MY CODE
BUT WHEN I DO I DO IT IN PRODUCTION
memegenerator.net

- “Stay buggy, my friends...”
- “That's our motto here. Doesn't work to well in practice, actually. “
- “Then blame all issues on QA -_- “
- “you don't even have to hire testers....they're just called ‘clients’.”

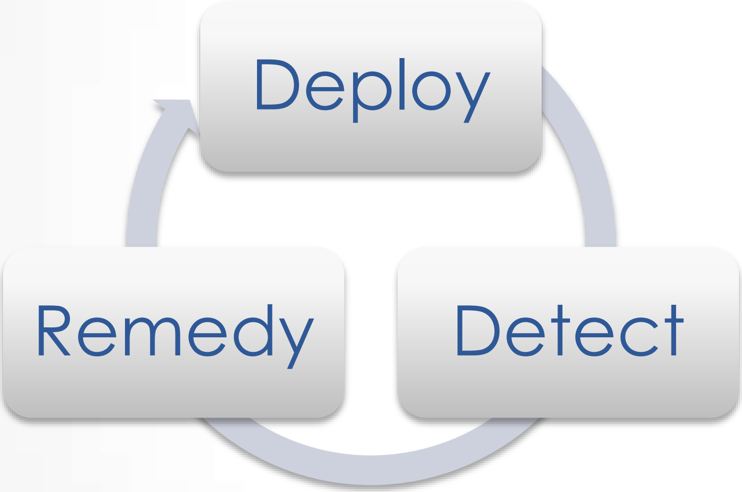
WhY do we TiP?

- Leverage the diversity of real users
- ...and real prod environment...
- ...to find bugs you cannot find pre-production



Why is TiP about Services?

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



- Iterative Virtuous Cycle
-

Google: All engineers have access to the production machines: "…deploy, configure, monitor, debug, and maintain"
[Google Talk, June 2007 @ 21:00]

Facebook: engineers must be present in a specific IRC channel for "roll call" before the release begins or else suffer a public "shaming"
[Facebook ships, 2011]

TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Methodologies

The many ways we can TiP

Ten Methodologies

- Based on observation across Microsoft and Industry
- Your team may categorize differently

Data Mining	Dogfood/beta
User Performance Testing	Synthetic Tests in Production
Environment Validation	User Scenario Execution
Experimentation for Design	Load Testing in Production
Controlled Test Flight	Destructive Testing

1 2 3 4 5 6 7 8 9 10

TiP Methodologies in Three Stages

Input

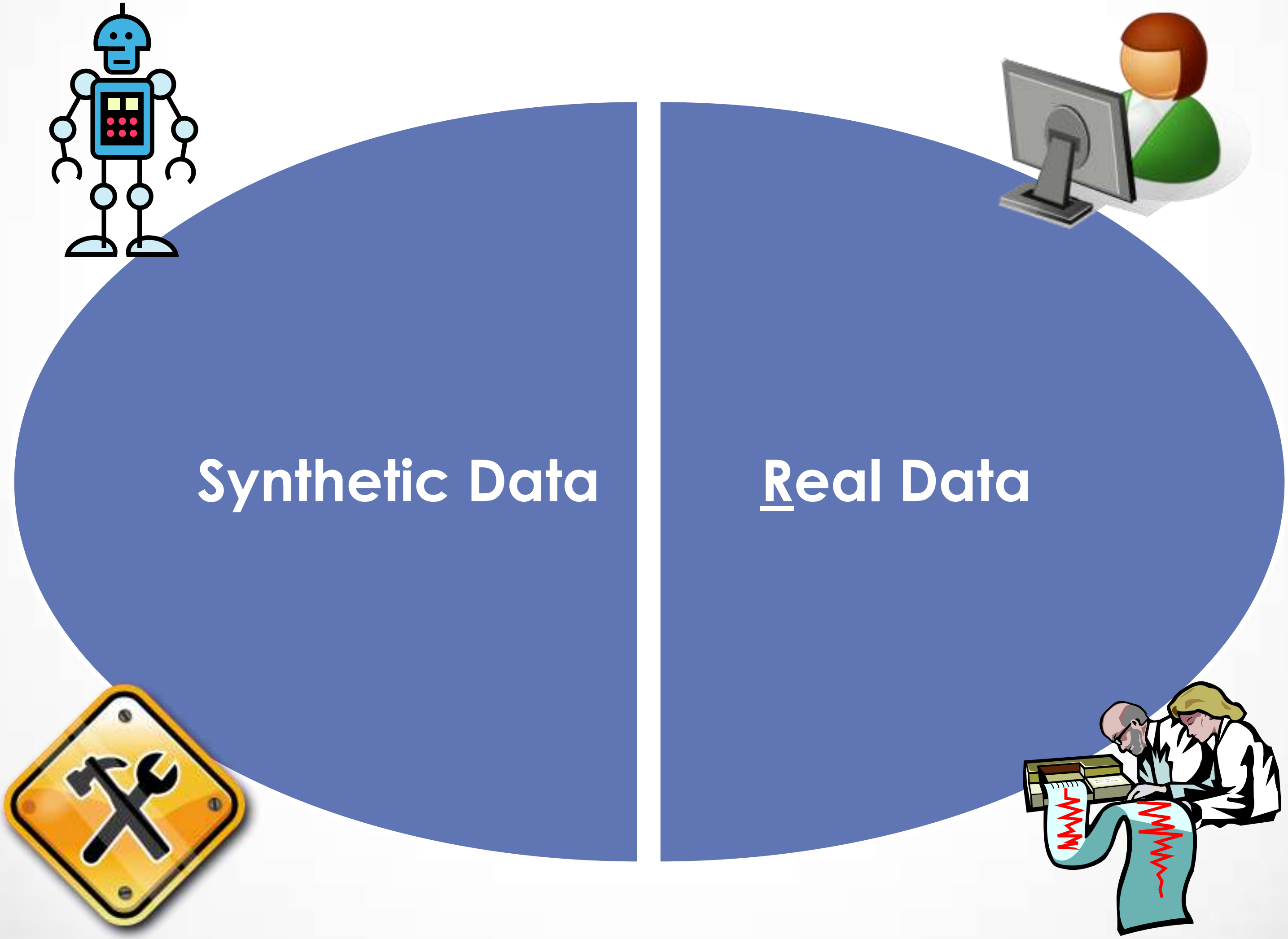
Effect

Observe

- **Input**
  - Where does the data driving your tests come from?
- **Effect**
  - Does the test change or act on production, and how?
- **Observation**
  - How do we measure the test results?

TiP Methodologies – Inputs

- Where does the data driving your tests come from?

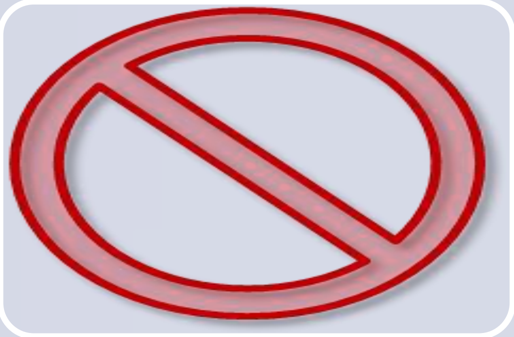


Synthetic Data

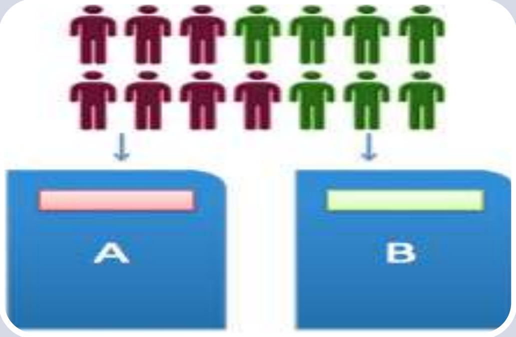
Real Data

TiP Methodologies - Effects

- Does the test change or act on production, and how?



No Change



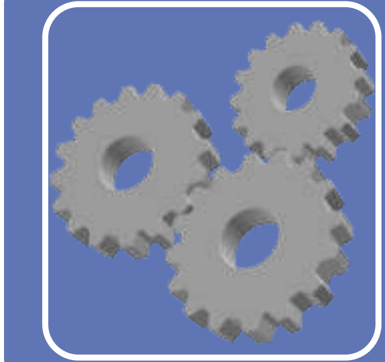
Experiment with Users

Act on or Change Service

TiP Methodologies - Observations

- How do we measure the test results?

User Behavior



System Behavior

Methodology	Inputs are…	Effect is…	We Observe…
Data Mining	Real User Data	None	User Behavior (also System Behavior)
User Performance Testing	Real User Data	None	System Behavior
Environment Validation	Real System Data	None	System Behavior
Experimentation for Design	Real User Data	Experiment with Users	User Behavior
Controlled Test Flight	Real User Data	Experiment with Users	System Behavior
Dogfood/beta	Real User Data	Experiment with Users	System Behavior (also User Behavior)
Synthetic Tests in Production	Synthetic User Data	Acting on System	System Behavior
User Scenario Execution	Synthetic User Data	Acting on System	System Behavior
Load Testing in Production	Synthetic User Data	Stress System	System Behavior
Destructive Testing	Synthetic System Data	Stress System	System Behavior

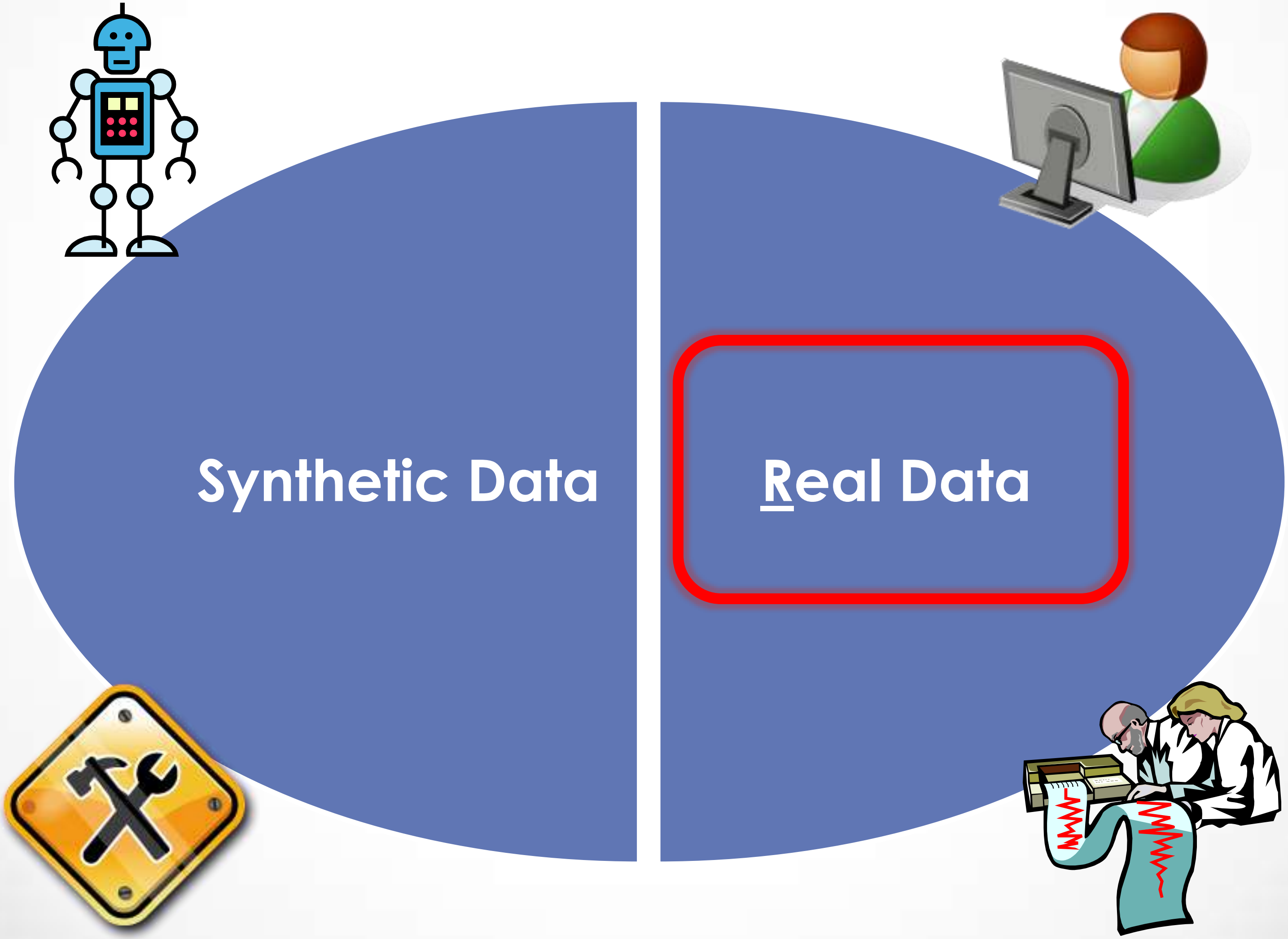
TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Real Data Input

TiP Methodologies – Inputs

- Where does the data driving your tests come from?



Synthetic Data

Real Data

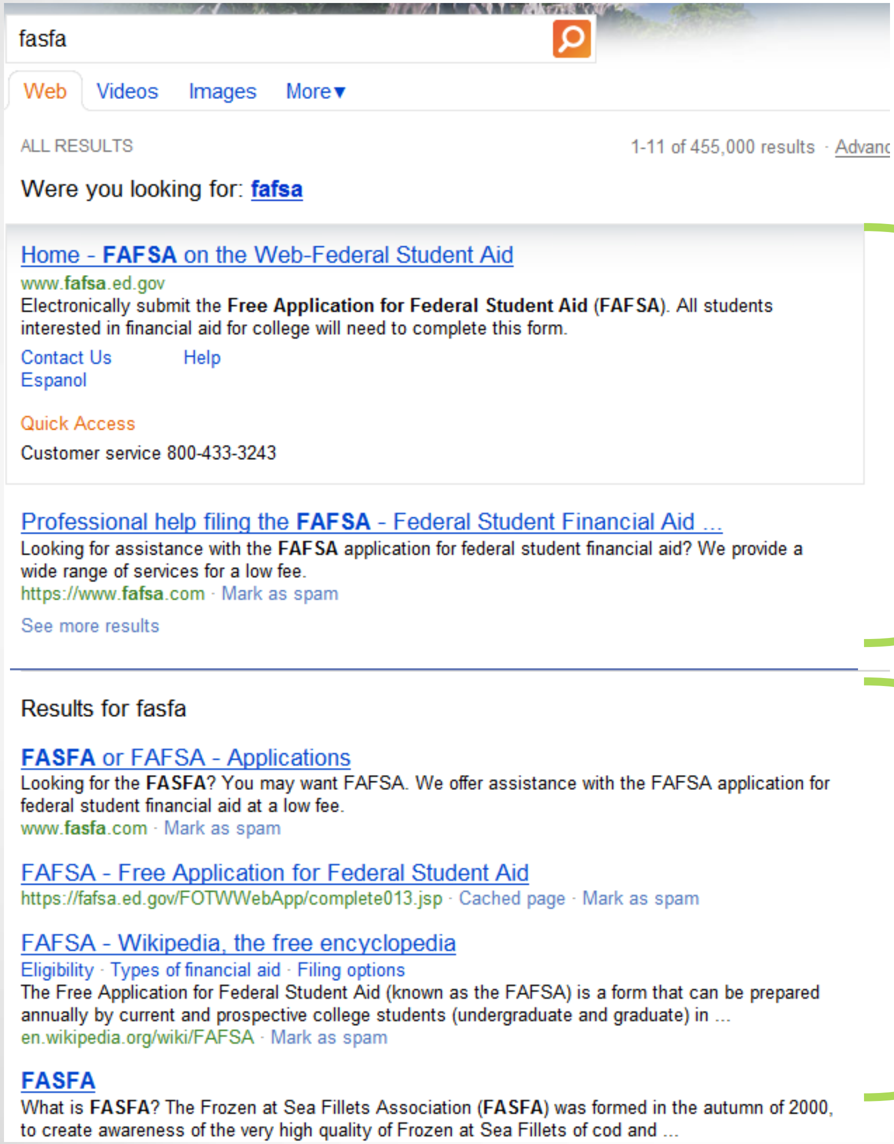
1

Data Mining

- Analysis of Real User Data
  - Large Data Sets
  - Real Usage Patterns
  - Tease out the Defects

- Different Approaches
  - Real-time feedback loop
  - Find patterns for humans to follow-up

Data Mining: Speller Split Screen



fasfa

ALL RESULTS 1-11 of 455,000 results · Advanc

Were you looking for: **fafsa**

Home - **FAFSA** on the Web-Federal Student Aid
www.**fafsa**.ed.gov
Electronically submit the **Free Application for Federal Student Aid** (**FAFSA**). All students interested in financial aid for college will need to complete this form.
Contact Us Help
Espanol

Quick Access
Customer service 800-433-3243

Professional help filing the **FAFSA** - Federal Student Financial Aid ...
Looking for assistance with the **FAFSA** application for federal student financial aid? We provide a wide range of services for a low fee.
https://www.**fafsa**.com · Mark as spam
See more results

Results for fasfa

FASFA or **FAFSA** - Applications
Looking for the **FASFA**? You may want FAFSA. We offer assistance with the FAFSA application for federal student financial aid at a low fee.
www.**fasfa**.com · Mark as spam

FAFSA - Free Application for Federal Student Aid
https://fafsa.ed.gov/FOTWWebApp/complete013.jsp · Cached page · Mark as spam

FAFSA - Wikipedia, the free encyclopedia
Eligibility · Types of financial aid · Filing options
The Free Application for Federal Student Aid (known as the FAFSA) is a form that can be prepared annually by current and prospective college students (undergraduate and graduate) in ...
en.wikipedia.org/wiki/FAFSA · Mark as spam

FASFA
What is **FASFA**? The Frozen at Sea Fillets Association (**FASFA**) was formed in the autumn of 2000, to create awareness of the very high quality of Frozen at Sea Fillets of cod and ...

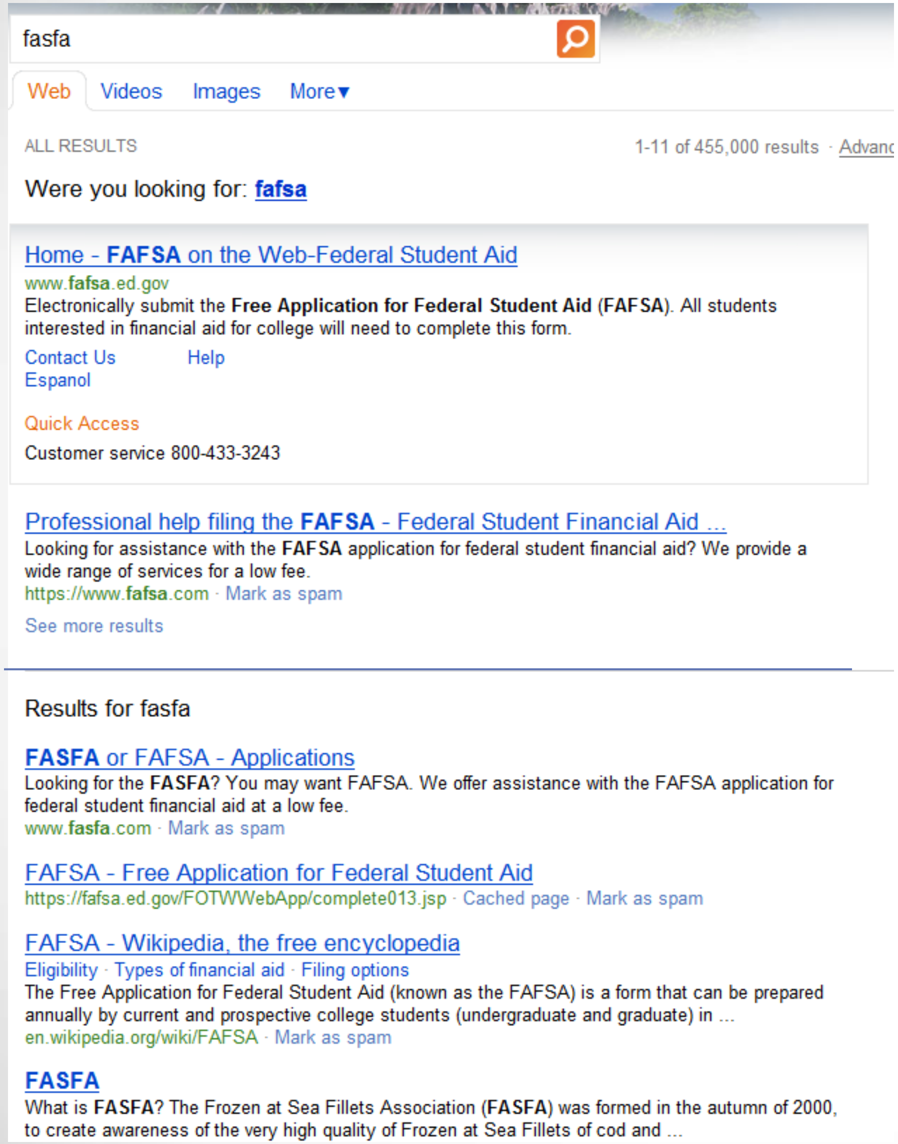
Top Pane

Result 1 and Result 2 from **Corrected** Query

Bottom Pane

Results from **Original** Query

Data Mining: Speller Split Screen



fasfa

Web Videos Images More▾

ALL RESULTS 1-11 of 455,000 results · Advanc

Were you looking for: **fafsa**

Home - **FAFSA** on the Web-Federal Student Aid
www.**fafsa**.ed.gov
Electronically submit the **Free Application for Federal Student Aid** (FAFSA). All students interested in financial aid for college will need to complete this form.
Contact Us Help
Espanol

Quick Access
Customer service 800-433-3243

Professional help filing the **FAFSA** - Federal Student Financial Aid ...
Looking for assistance with the **FAFSA** application for federal student financial aid? We provide a wide range of services for a low fee.
https://www.**fafsa**.com · Mark as spam
See more results

Results for fasfa

FASFA or FAFSA - Applications
Looking for the **FASFA**? You may want FAFSA. We offer assistance with the FAFSA application for federal student financial aid at a low fee.
www.**fasfa**.com · Mark as spam

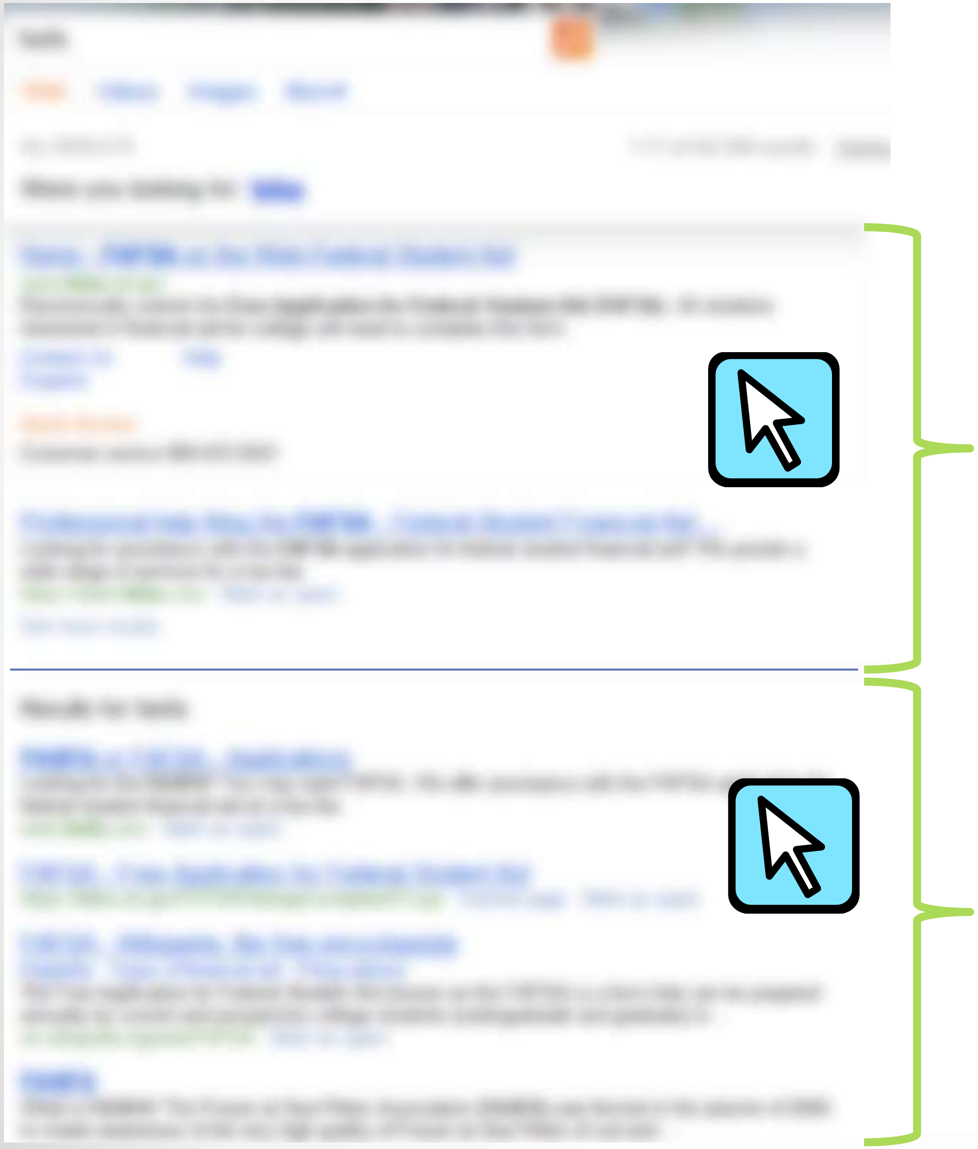
FAFSA - Free Application for Federal Student Aid
https://fafsa.ed.gov/FOTWWebApp/complete013.jsp · Cached page · Mark as spam

FAFSA - Wikipedia, the free encyclopedia
Eligibility · Types of financial aid · Filing options
The Free Application for Federal Student Aid (known as the FAFSA) is a form that can be prepared annually by current and prospective college students (undergraduate and graduate) in ...
en.wikipedia.org/wiki/FAFSA · Mark as spam

FASFA
What is **FASFA**? The Frozen at Sea Fillets Association (**FASFA**) was formed in the autumn of 2000, to create awareness of the very high quality of Frozen at Sea Fillets of cod and ...

- Split screen is a poor compromise
- Want to do better:
  - Just fix the spelling
  - Or leave query as-is

Data Mining: Speller Split Screen



Alter query…

If top pane clicks > 90%

Leave query as-is…

If bottom pane clicks > 90%

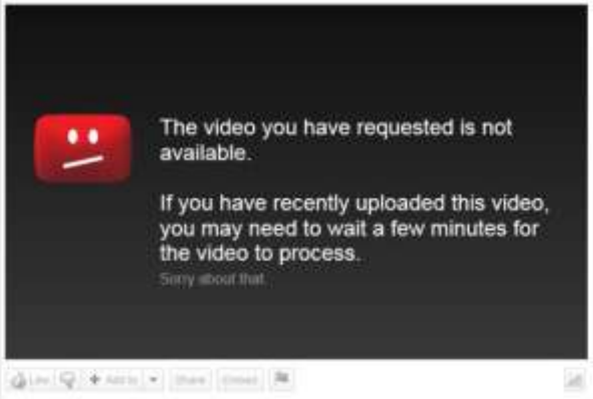
[Unpingco, Feb 2011]

Data Mining: Speller Split Screen

Some Top Results

Query	Correction	Top Clicks	Bottom Clicks	Keep Original	Make Correction
yah	yahoo	99.8%	0.2%		✓
fasfa	fafsa	90.5%	9.5%		✓
utube music	youtube music	90.4%	9.6%		✓
facebookcom	facebook.com	98.4%	1.6%		✓
imbd	imdb	96.3%	3.7%		✓
evony	ebony	0.5%	99.5%	✓	
century link	centrelink	2.0%	98.0%	✓	
yout	youth	3.8%	96.2%	✓	

Detecting Video Takedown Pages



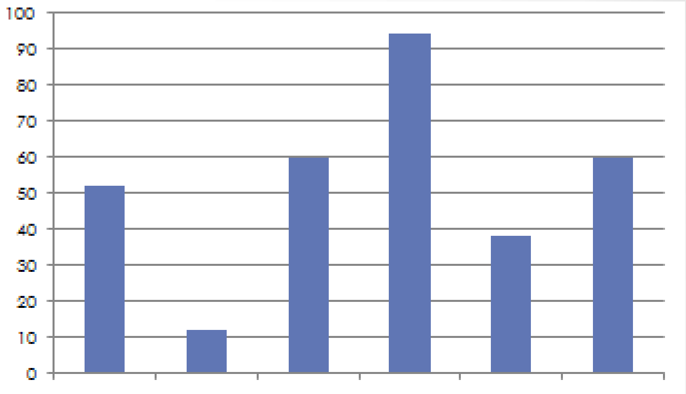
The video you have requested is not available.

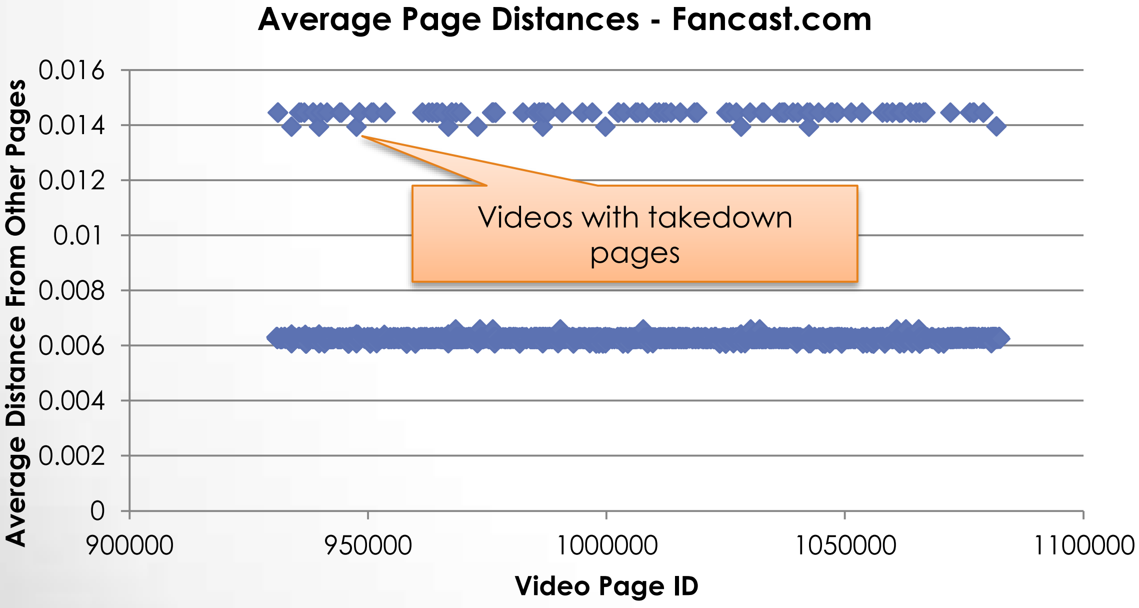
If you have recently uploaded this video, you may need to wait a few minutes for the video to process.

Sorry about that.

Automated heuristic: Most videos on a video site will play. Takedown pages look different; find outliers.

Video Histogram "Fingerprint"





- **3.31%** of our video results lead to a video takedown page.
- **5.56%** of our YouTube videos lead to a video takedown page.
- More than **80,000** users experience this per week.
- **We are feeding this back into the product!**

bing

2 User Performance Testing

- Collect specific telemetry about how long **stuff** takes from user point of view

- Real User Data – Real User Experience

- End to End = complete request and response cycle
  - From user to back-end round-trip
  - Include traffic through CDN or data providers
  - Measured from the user point of view

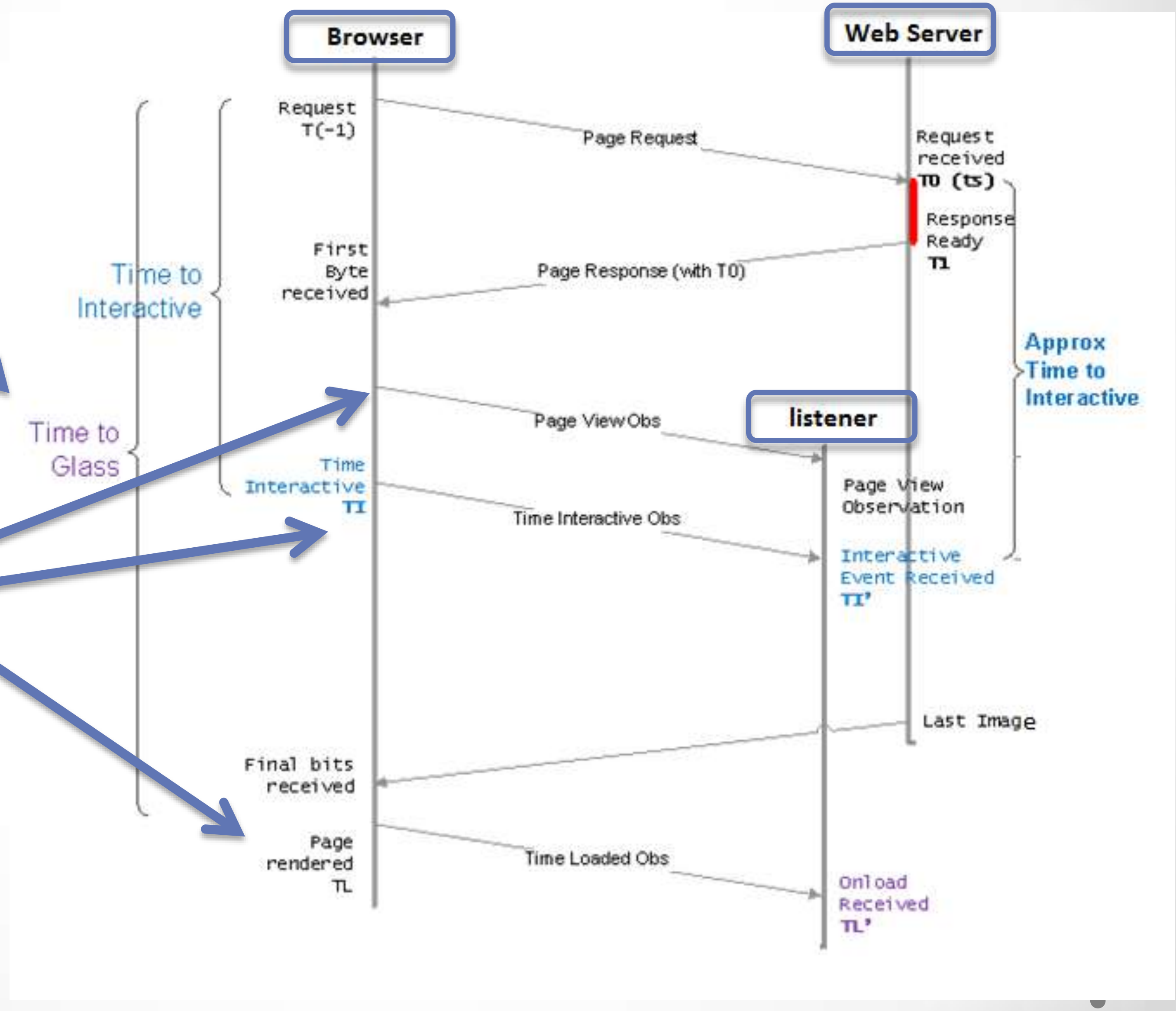
CDN – Content Data Network

- From around the world
- From diversity of browsers, OS, devices

Stuff includes… Page Load Time (PLT)

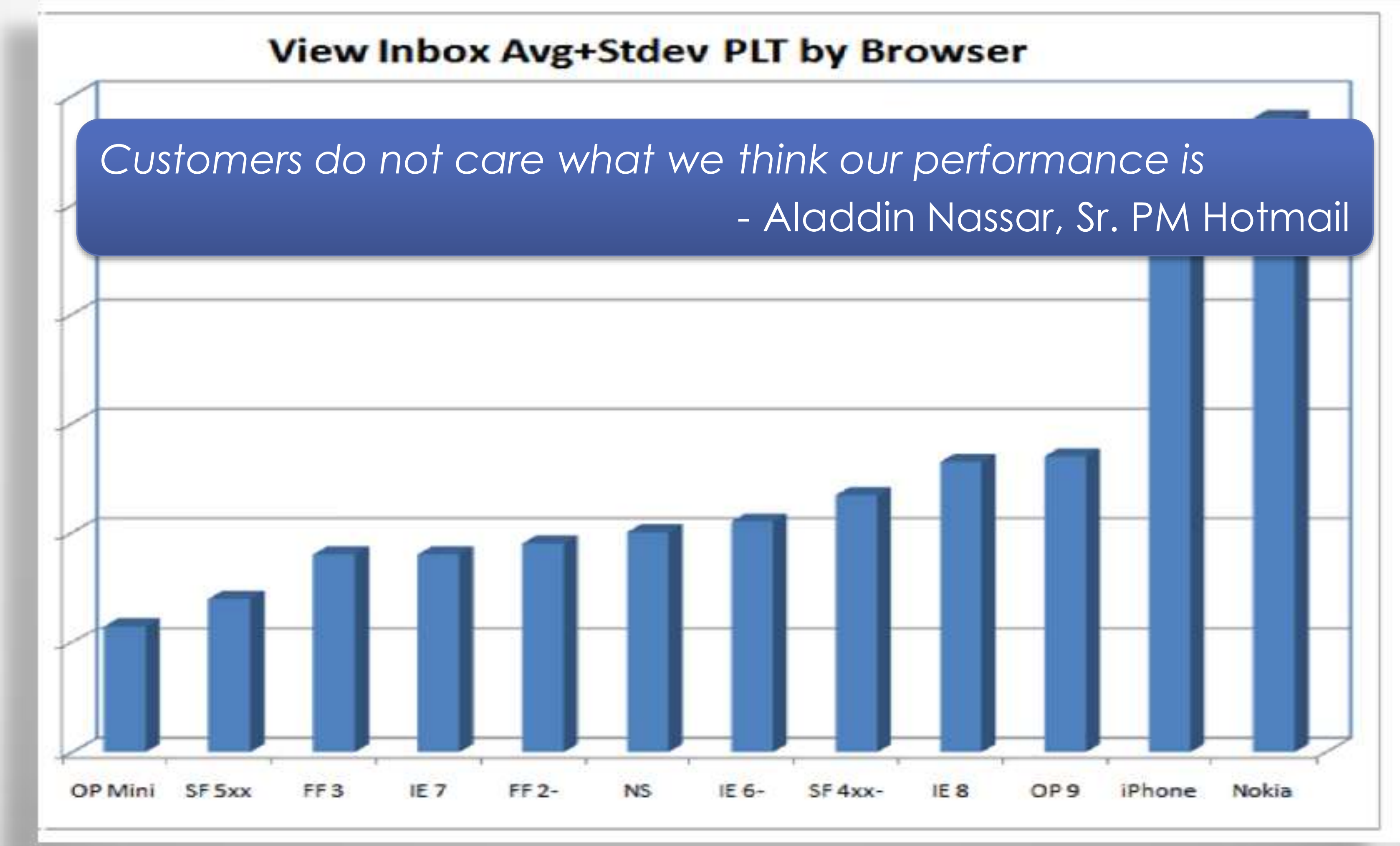
TTG, real PLT for real users

Client Instr. sends Beacons



Hotmail JSI User Performance Testing

- PLT by browser, OS, country, cluster, etc..



Customers do not care what we think our performance is

- Aladdin Nassar, Sr. PM Hotmail

User Performance Testing Examples

- Hotmail
  - Re-architected from the ground up around performance
  - Read messages are 50% faster

- Windows Azure™
  - **Every API:** Tracks how many calls were made; how many succeeded, and how long each call took to process

- Bing™ PerfPing
  - **Measures user perceived performance**
    - Measurement points occur at the client

3

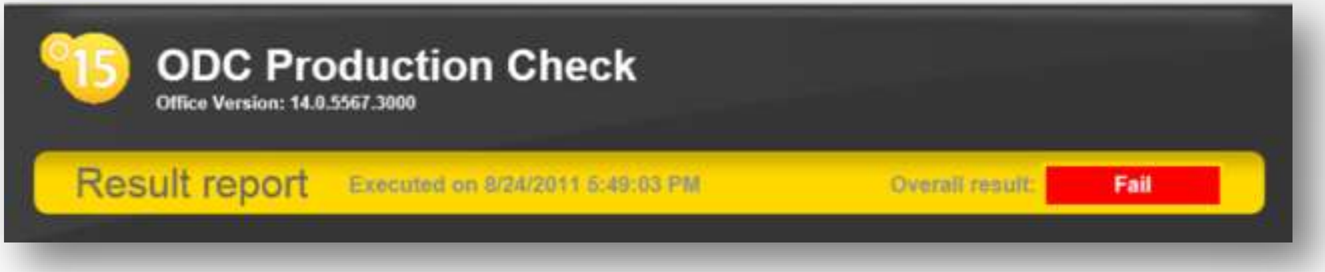
Environment Validation

- What is the most dangerous time for a service?

- System = Service + Environment
- Environment Validation checks…
  - DLL/File version compatibility
  - Connection Health
  - Certificate Installation and Validity
  - Content propagation across servers (freshness)
  - Other…

- Runs at deployment time
  - Or, all the time (always)

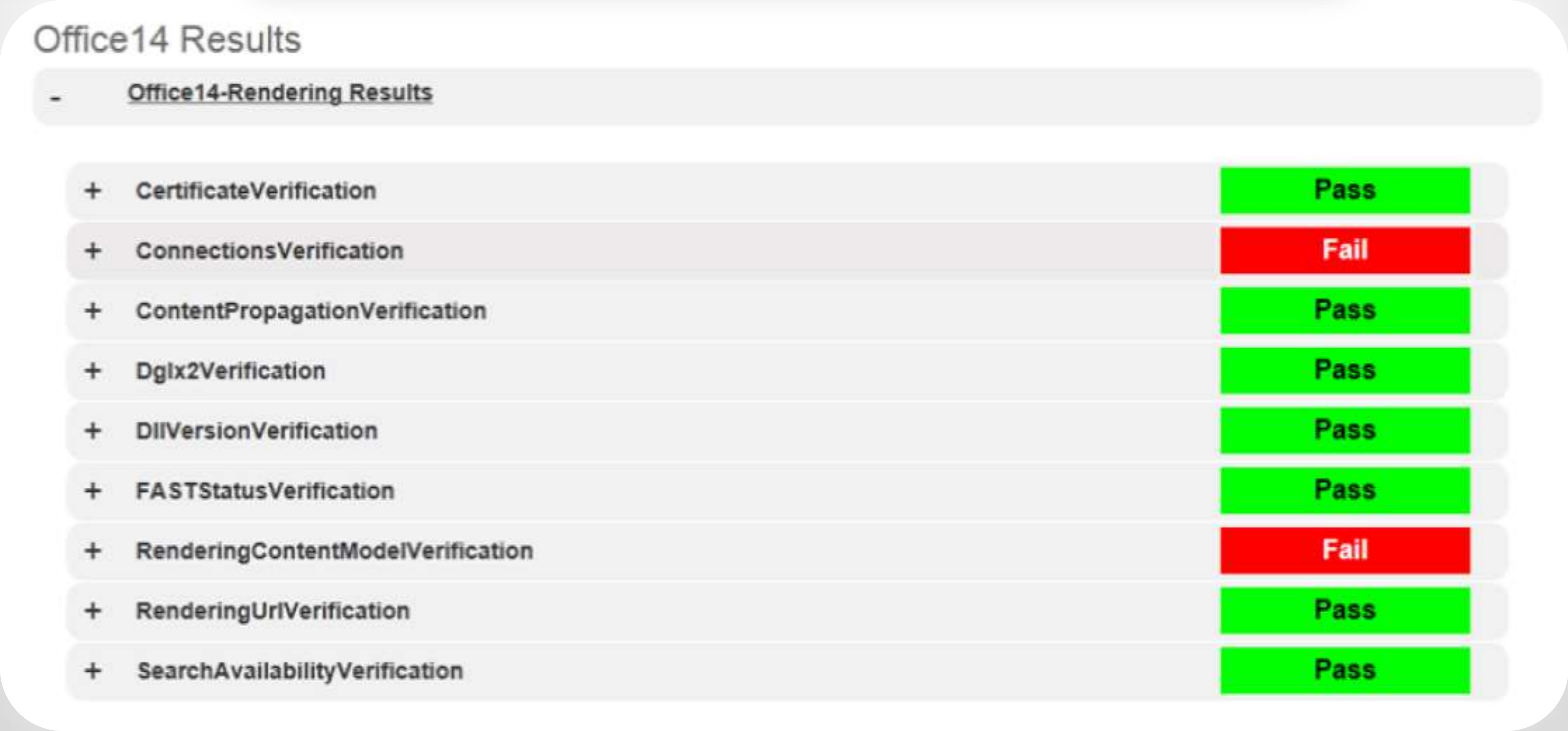
Environment Validation for Office.com



ODC Production Check

Office Version: 14.0.5567.3000

Result report Executed on 8/24/2011 5:49:03 PM Overall result: Fail



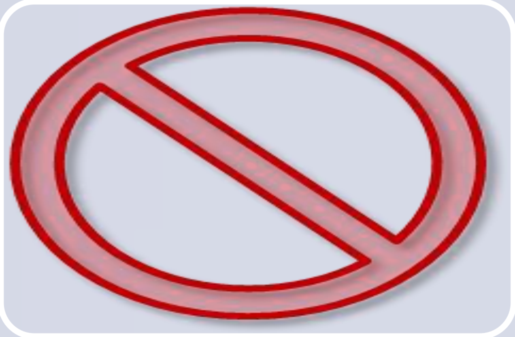
Office14 Results

- Office14-Rendering Results

+	CertificateVerification	Pass
+	ConnectionsVerification	Fail
+	ContentPropagationVerification	Pass
+	Dglx2Verification	Pass
+	DllVersionVerification	Pass
+	FASTStatusVerification	Pass
+	RenderingContentModelVerification	Fail
+	RenderingUrlVerification	Pass
+	SearchAvailabilityVerification	Pass

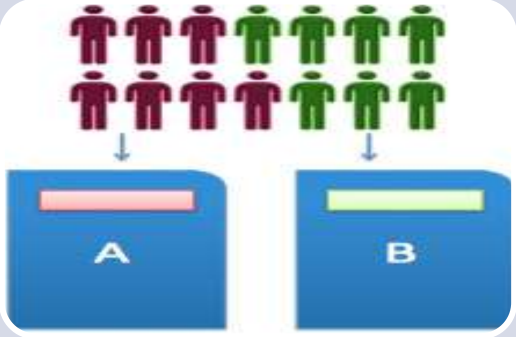
TiP Methodologies - Effects

- Does the test change or act on production, and how?



No Change

- Data Mining
- User Perf. Testing
- Environment Validation

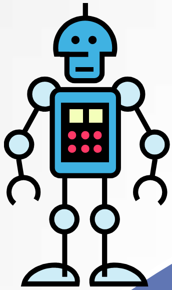


Experiment with Users

Act on or Change Service

TiP Methodologies - Inputs

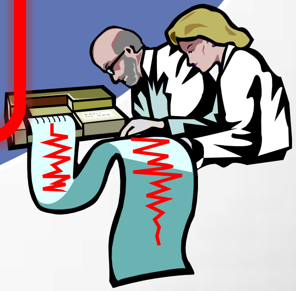
Where does the data driving your tests come from?



Synthetic Data

Real Data

- Data Mining
- User Performance Testing
- Environment Validation



TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Experiment with Users

Experimentation

"To have a great idea, have a lot of them"

-- Thomas Edison

FAIL HARDER
MOVE FAST AND BREAK THINGS
WHAT WOULD YOU DO IF YOU WEREN'T AFRAID?

- Try new things… in production
- Build on Successes
- Cut your losses… before they get expensive

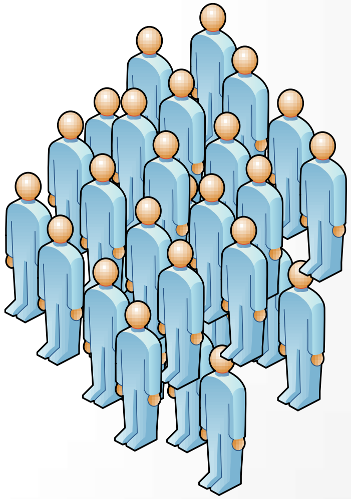
  - A/B Testing - aka Controlled Experimentation
  - Un-Controlled Experimentation

Mitigate Risk with eXposure Control

- Launch a new Service – Everyone sees it
- Exposure Control – only *some* see it

By Browser

By Location



By Percent (scale)

Who’s doing Exposure Control?

Three concentric push phases [Facebook ships, 2011]

- p1 = internal release
- p2 = small external release
- p3 = full external release

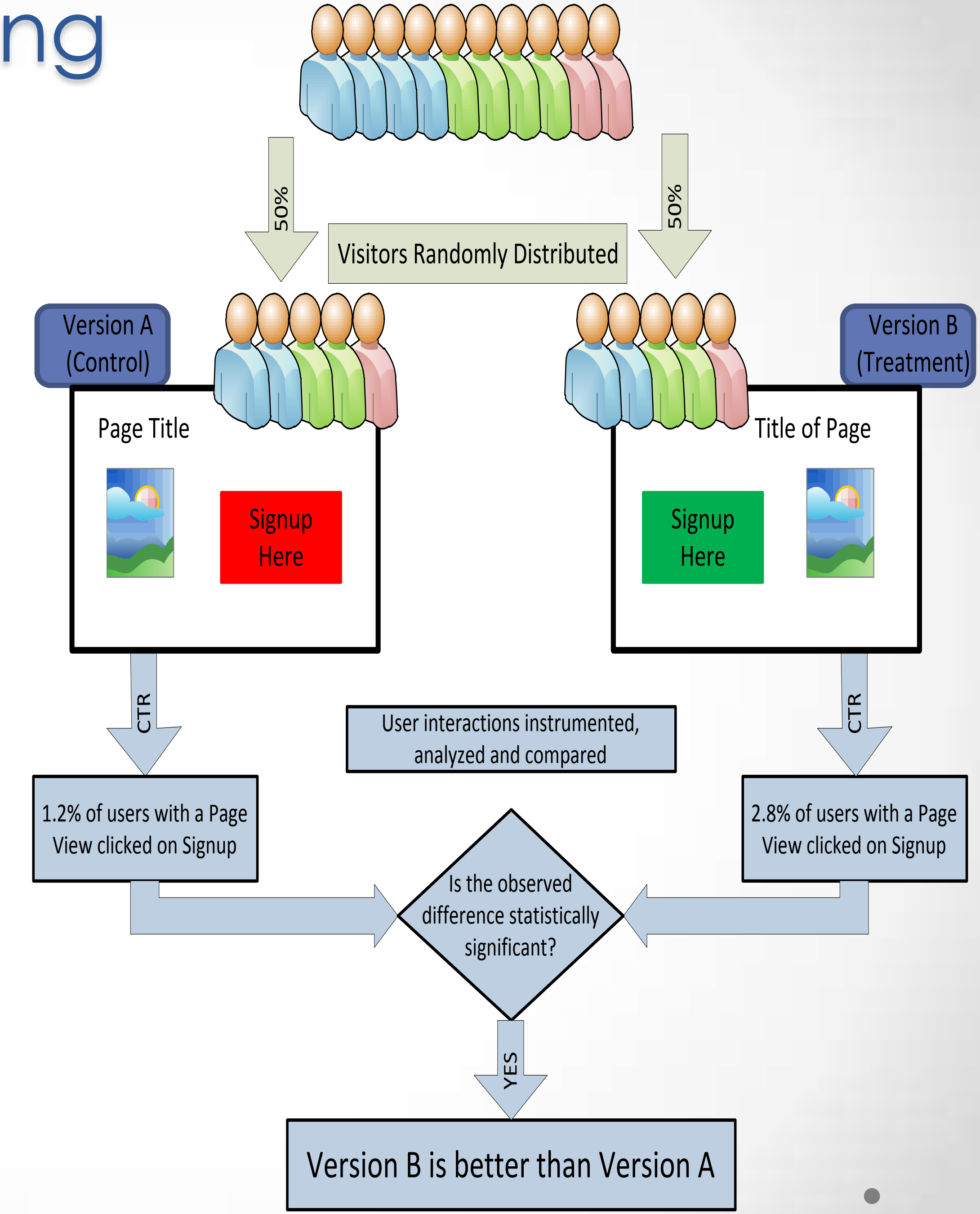
“We do these 1% launches where we float something out and measure that. We can dice and slice in any way you can possibly fathom.”

-Eric Schmidt, former CEO, Google

[Google BusinessWeek, April 2008]

A/B Testing

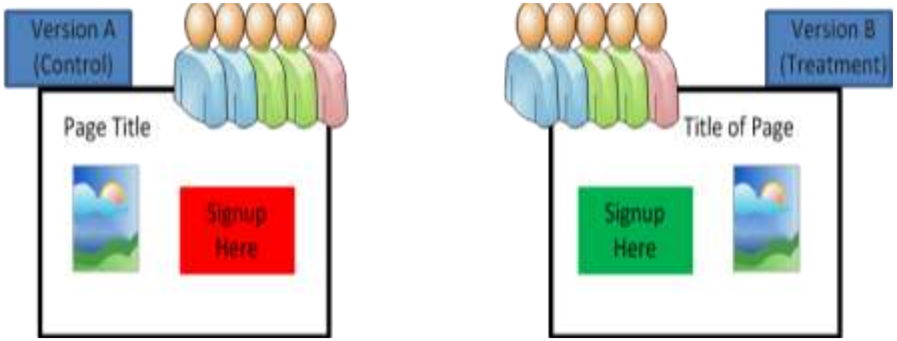
- Aka: Controlled Online Experimentation


50%
50%
Visitors Randomly Distributed
Version A (Control)
Version B (Treatment)
Page Title
Signup Here
Title of Page
Signup Here
CTR
CTR
User interactions instrumented, analyzed and compared
1.2% of users with a Page View clicked on Signup
2.8% of users with a Page View clicked on Signup
Is the observed difference statistically significant?
YES
Version B is better than Version A

AB Testing: Two Different TiP Methodologies

4

- Experimentation for Design
  - Business Metrics – Did we build the right thing?



5

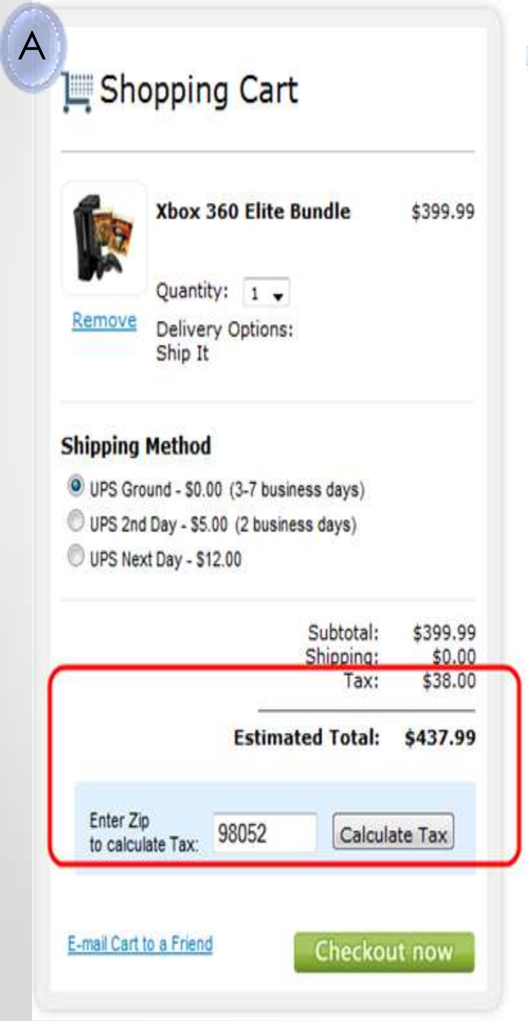
- Controlled Test Flight
  - Assess quality of new code →deployment Go/No-go
  - Code Quality Metrics –

Did we build it right?

Experimentation for Design

Example: Microsoft Store

Goal: Increase Average Revenue per User



A

Shopping Cart

Xbox 360 Elite Bundle $399.99

Quantity: 1

Remove

Delivery Options:
Ship It

Shipping Method

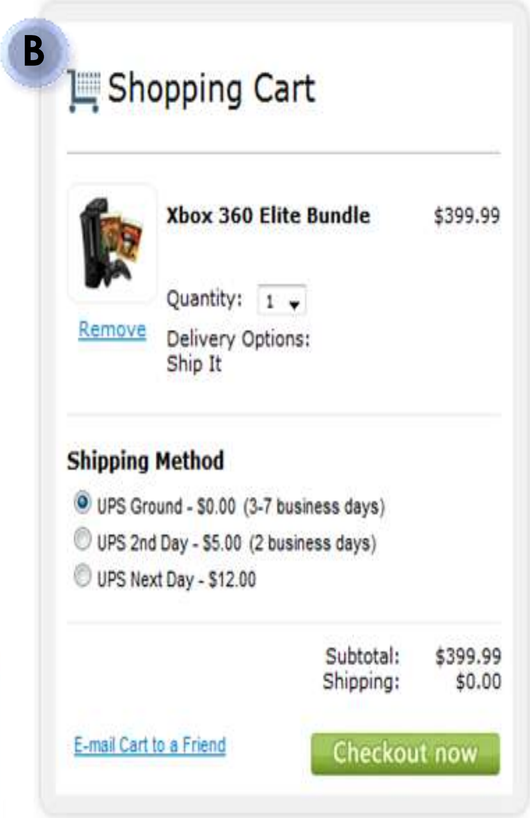
● UPS Ground - $0.00 (3-7 business days)
○ UPS 2nd Day - $5.00 (2 business days)
○ UPS Next Day - $12.00

Subtotal:	$399.99
Shipping:	$0.00
Tax:	$38.00
Estimated Total:	**$437.99**

Enter Zip to calculate Tax: 98052 Calculate Tax

E-mail Cart to a Friend

Checkout now



B

Shopping Cart

Xbox 360 Elite Bundle $399.99

Quantity: 1

Remove

Delivery Options:
Ship It

Shipping Method

● UPS Ground - $0.00 (3-7 business days)
○ UPS 2nd Day - $5.00 (2 business days)
○ UPS Next Day - $12.00

Subtotal:	$399.99
Shipping:	$0.00

E-mail Cart to a Friend

Checkout now

Which increased revenue…?

A. Control

B. Treatment – up 3.3%

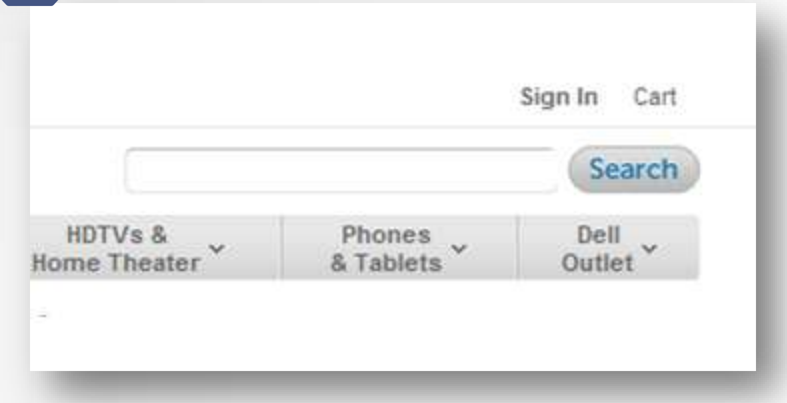
C. Neither

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

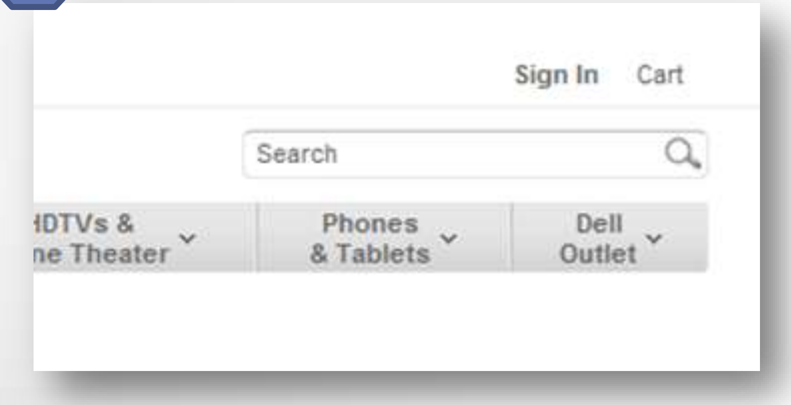
Experimentation for Design

Example: Dell

A



B

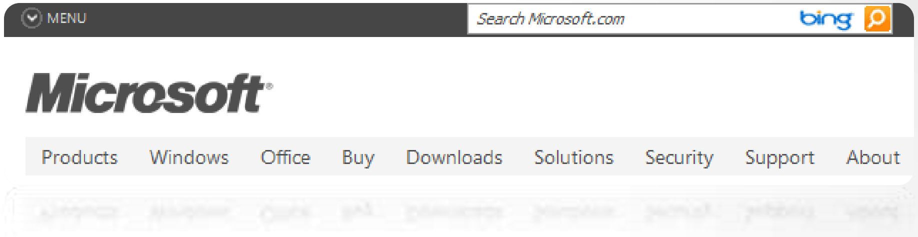


Which Increased Revenue Per Visitor?

A. **Control – up 6.3%**
B. Treatment
C. Neither

[Dell, 2012]

Example: Controlled Test Flight – Microsoft.com



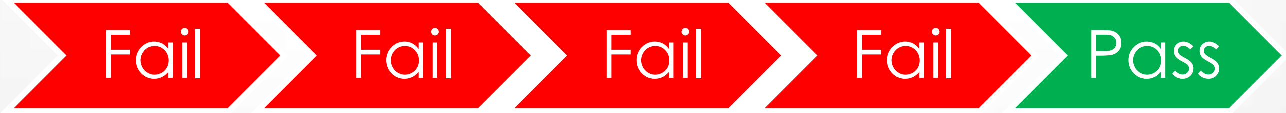
- Microsoft.com
  - Put a single server with v-next in production
  - Monitor the server and applications hosted on the server
  - Capture traffic, volume, performance, and availability
  - Pull it back out and "crawl" the logs

- No Functional difference observable by user
- Not always truly random and un-biased

[Microsoft.com, TechNet]

Example Controlled Test Flight : 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.



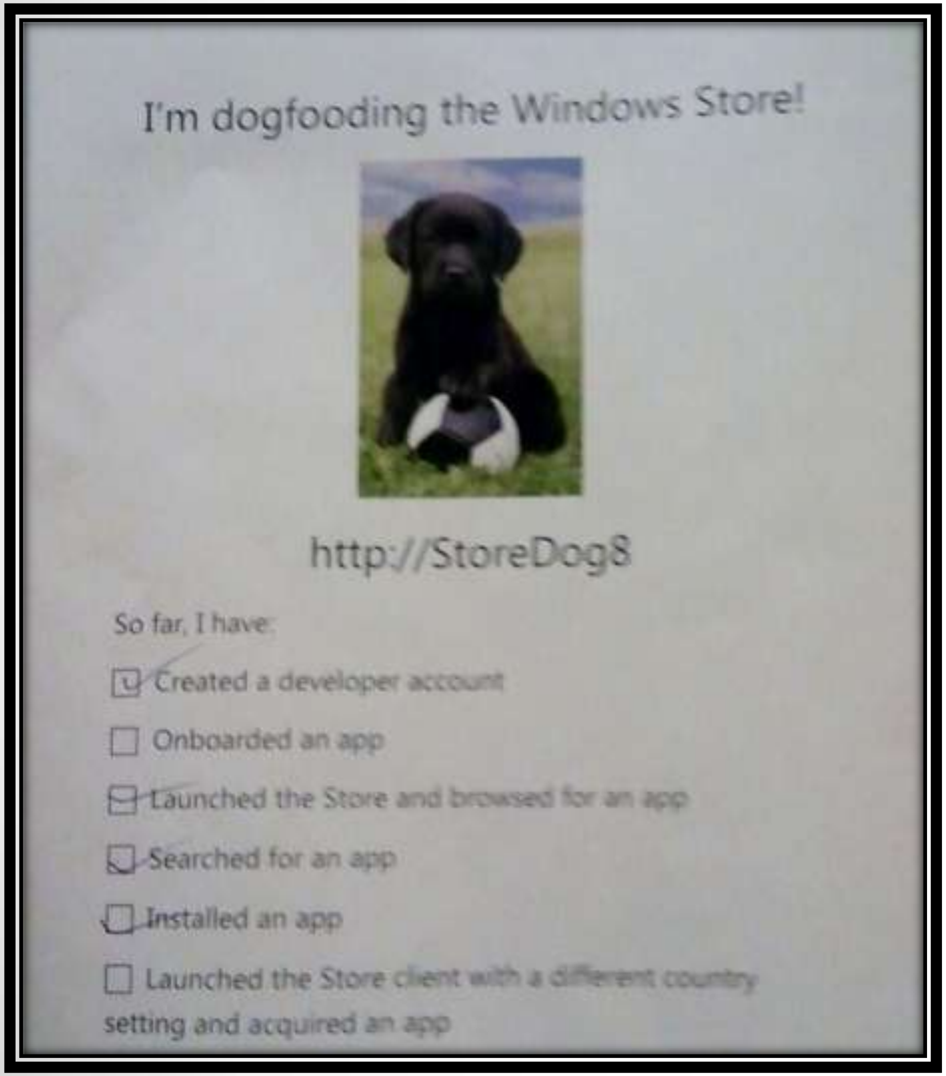
Example: Controlled Test Flight – Netflix

- Deployed to the Cloud (AWS)
- Developers use web based portal to deploy new code alongside old code
  - Put one "canary" instance into traffic
- Go / No-Go
  - If Go, then old instances removed automatically

[Cockcroft, March 2012]

6

Un-Controlled Experimentation



I'm dogfooding the Windows Store!

http://StoreDog8

So far, I have:

☑ Created a developer account

□ Onboarded an app

☑ Launched the Store and browsed for an app

☑ Searched for an app

☑ Installed an app

□ Launched the Store client with a different country setting and acquired an app

Dogfood and Beta

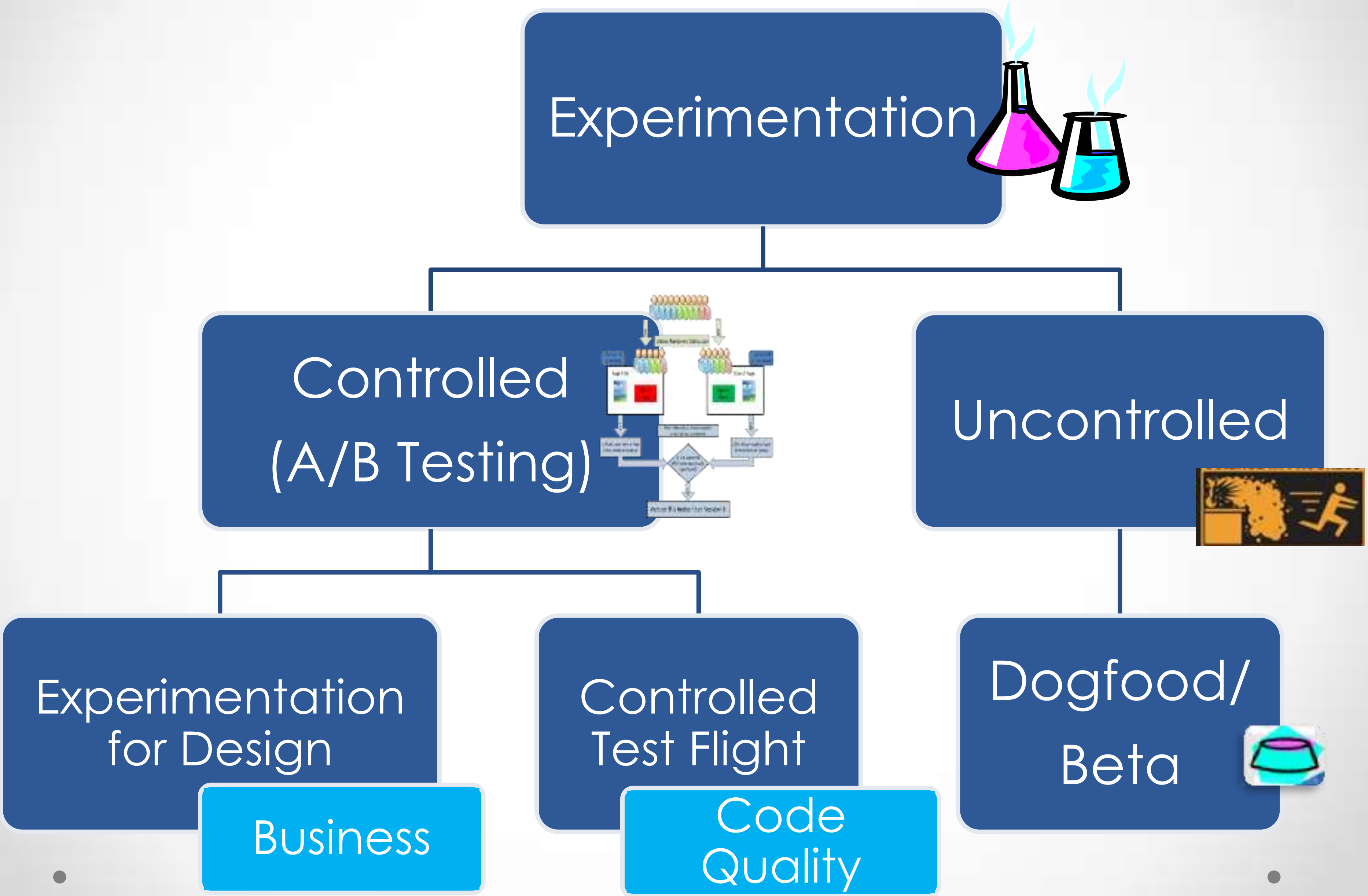
Different from AB Testing:

- Users opt-in; users know!
- Active Feedback (also telemetry)

Dogfood vs. Beta

- DF: Providers use own product
- Beta: limited general audience usage

The Experimentation Family



Experimentation

Controlled (A/B Testing)

Uncontrolled

Experimentation for Design

Business

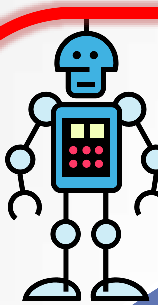
Controlled Test Flight

Code Quality

Dogfood/ Beta

TiP Methodologies - Inputs

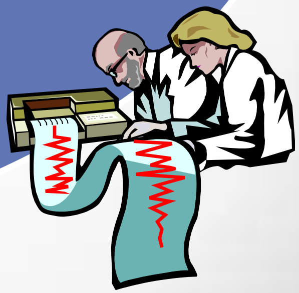
Where does the data driving your tests come from?



Synthetic Data

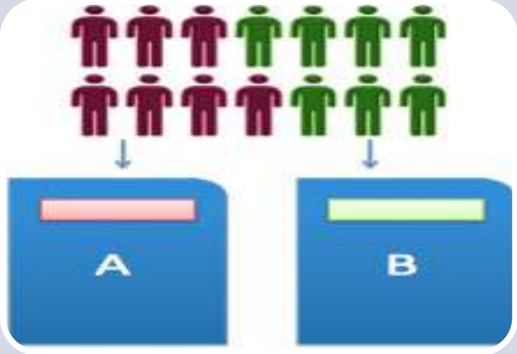
Real Data

- Data Mining
- User Performance Testing
- Environment Validation
- Experimentation for Design
- Controlled Test Flight
- Dogfood/beta



TiP Methodologies - Effects

- Does the test change or act on production, and how?



No Change

Data Mining

User Perf. Testing

Environment Validation

Experiment with Users

Experim. for Design

Controlled Test Flight

Dogfood/Beta

Act on or Change Service

TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Synthetic Data Input
&
Effect on Service

TiP Test Execution

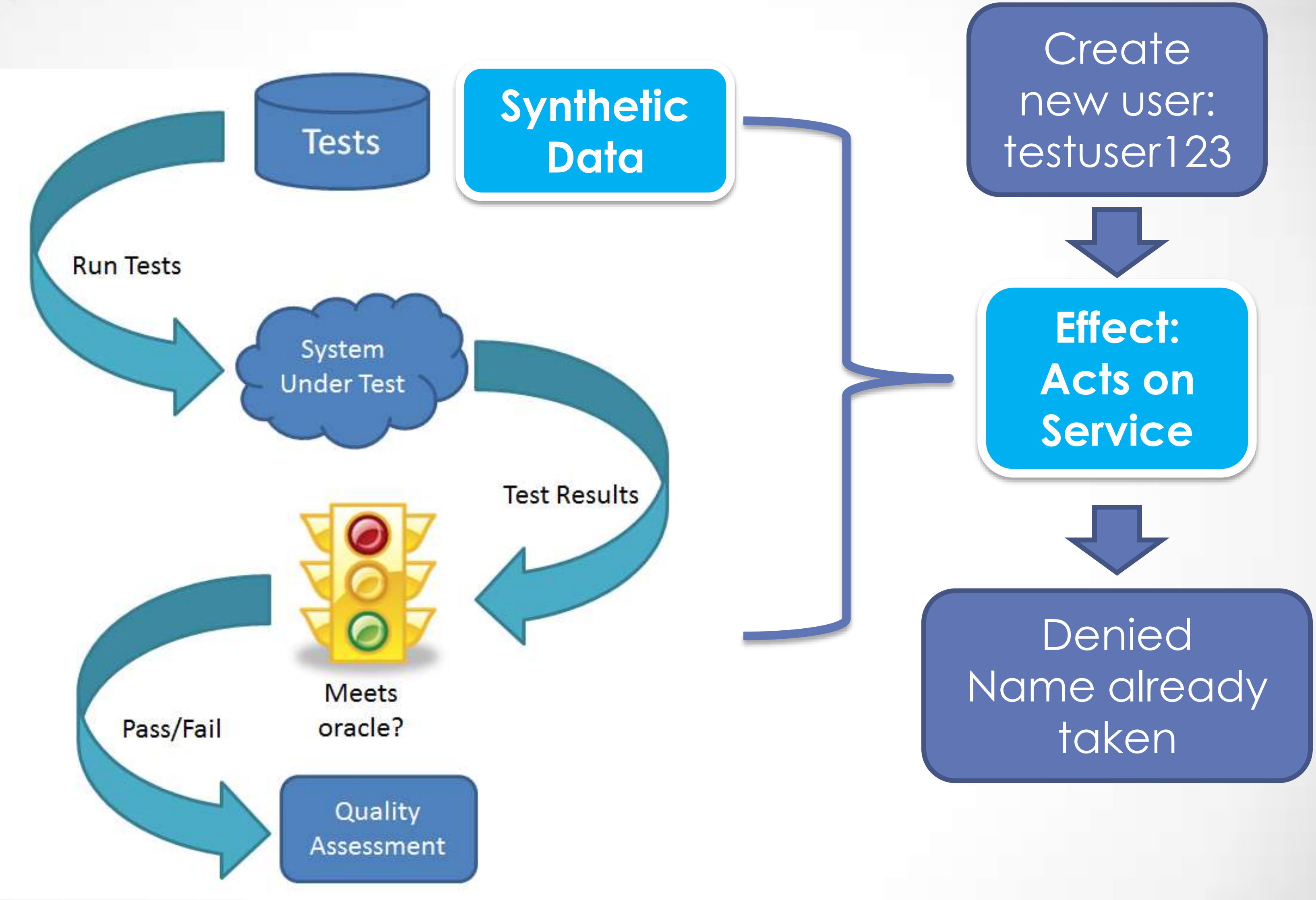
7

- Synthetic Tests in Production
  - Automation
  - Against internal APIs

8

- User Scenario Execution
  - From User Entry Point- Synthetic Transactions
  - E2E Scenario in Production
  - Automation or….
  - Manual

Synthetic Data / Effect on Service



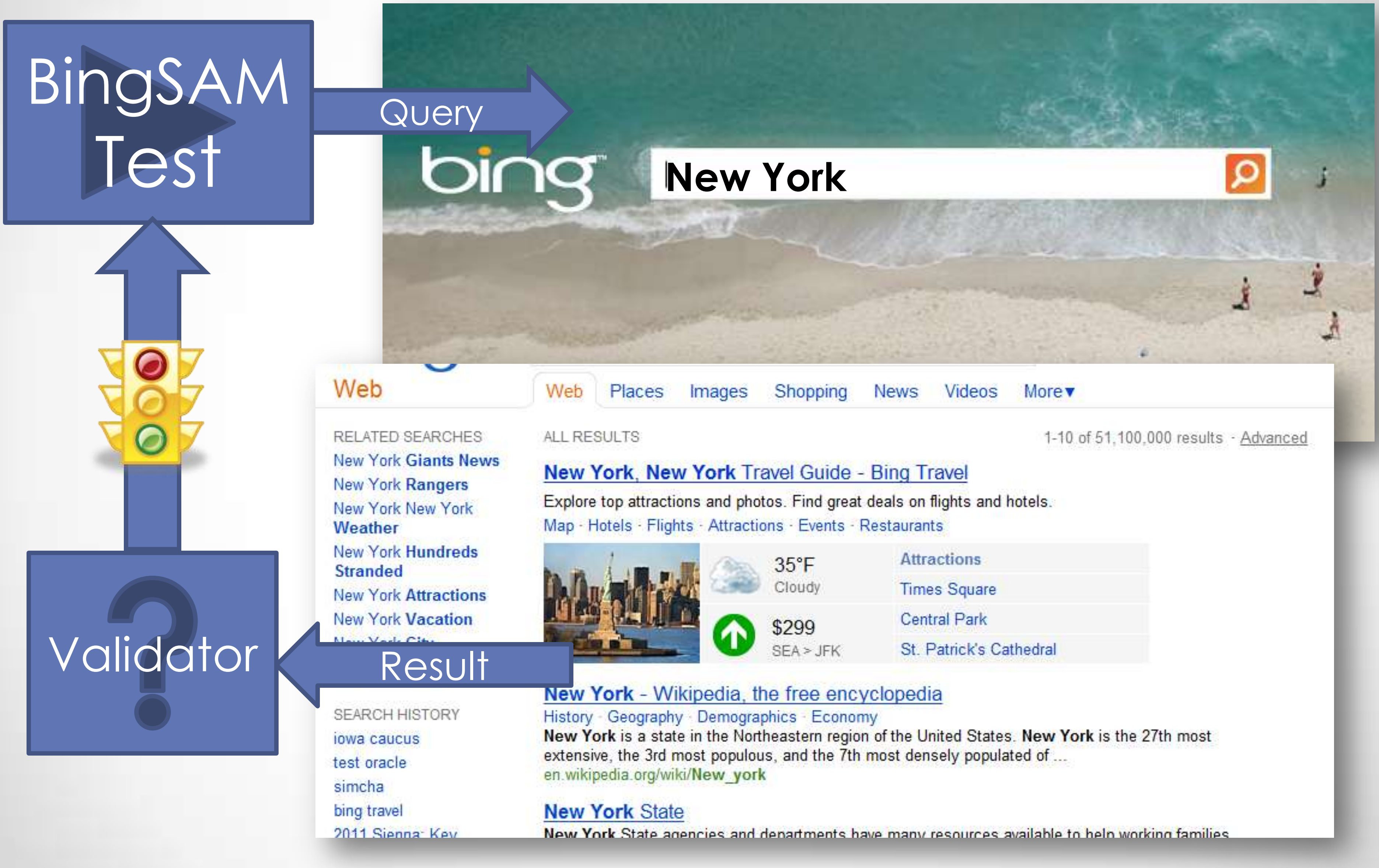
Synthetic Data

Create new user: testuser123

Effect: Acts on Service

Denied Name already taken

User Scenario Execution: BingSAM



bing

New York

Web

Web · Places · Images · Shopping · News · Videos · More▾

RELATED SEARCHES

- New York **Giants News**
- New York **Rangers**
- New York New York **Weather**
- New York **Hundreds Stranded**
- New York **Attractions**
- New York **Vacation**
- New York City

SEARCH HISTORY

- iowa caucus
- test oracle
- simcha
- bing travel
- 2011 Sienna: Key

ALL RESULTS

1-10 of 51,100,000 results · Advanced

New York, New York Travel Guide - Bing Travel

Explore top attractions and photos. Find great deals on flights and hotels.

Map · Hotels · Flights · Attractions · Events · Restaurants

		Attractions
	35°F Cloudy	Times Square
		Central Park
	$299 SEA > JFK	St. Patrick's Cathedral

New York - Wikipedia, the free encyclopedia

History · Geography · Demographics · Economy

New York is a state in the Northeastern region of the United States. **New York** is the 27th most extensive, the 3rd most populous, and the 7th most densely populated of ...

en.wikipedia.org/wiki/**New**_**york**

New York State

New York State agencies and departments have many resources available to help working families

Google Staged Approach

Risk Mitigation Index (RMI): the risk of the failing functionality of a product.

- Up-Front Test
  - Internal testers, lower the risk from 100% to 91%.
- User Scenario Execution
  - Crowd-sourcing such as UTest to get to 60%
- Dogfood
  - Released to dog-fooders, get risk down to 48%.
- Beta
  - Beta version is released

[Google GTAC 2010]

Who’s Right?

“These are some number of bugs that simply cannot be found until the house is lived in and software is no different. It needs to be in the hands of real users doing real work with real data in real environments”

James Whittaker, Former Engineering Director, Google

[Google, JW 2009]

[It’s a mistake to assume] all users are early adopters with excellent technical ability

Jon Bach, Director of Live Site Quality, eBay

[STPCon, 2012]

Test Data Handling

- Synthetic Tests + Real Data = Potential Trouble
  - Avoid it
  - Tag it
  - Clean it up

Example: Facebook Test Users

- Cannot interact with real users
- Can only friend other Test Users
- Create 100s
- Programmatic Control

Write Once, Test Anywhere

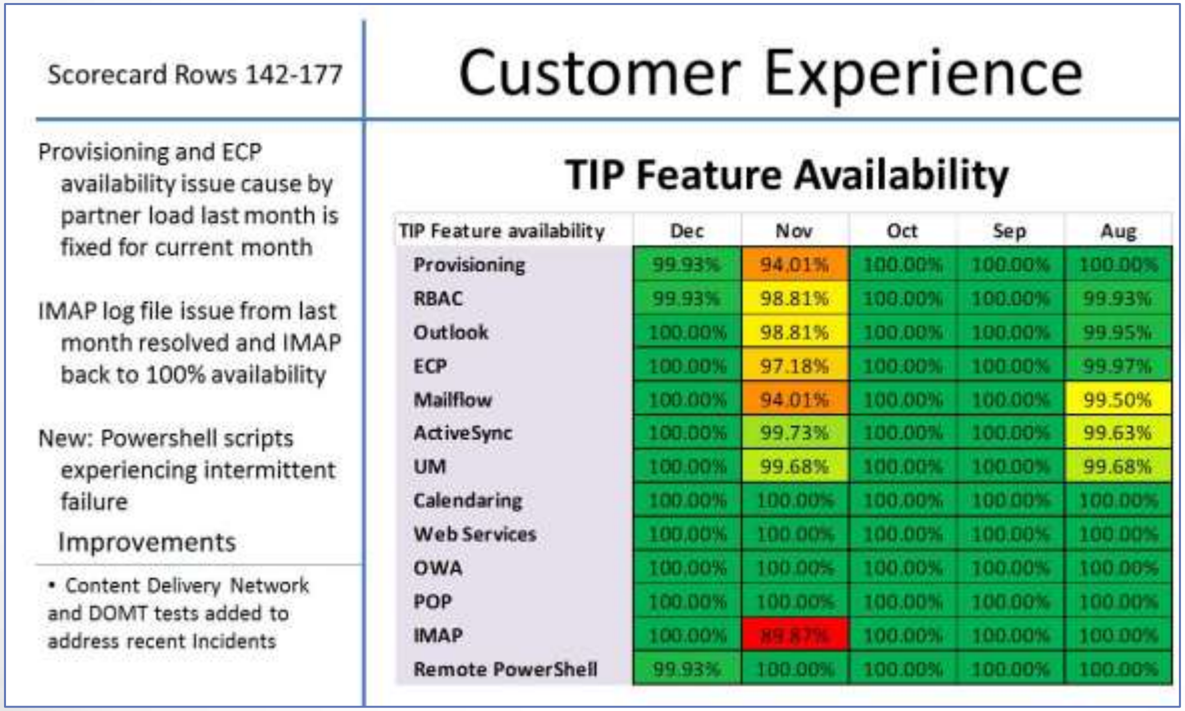
Microsoft Exchange Server 2010

- Microsoft Exchange
  - 70,000 automated test cases in lab
  - Re-engineered to run from the cloud
- TiP Execution Framework
  - Test Harness runs tests from Azure Cloud Platform
- Exchange measured performance
  - Latency: baseline, and measured over time
  - How….?

[Deschamps, Johnston, Jan 2012]

Active Monitoring

- Microsoft Exchange
  - Instead of pass/fail signal look at thousands of continuous runs.
    - Did we meet the "five nines" (99.999%) **availability**?
    - Did we complete the task in less than 2 seconds 99.9% of the time? - **performance**



Scorecard Rows 142-177

Customer Experience

Provisioning and ECP availability issue cause by partner load last month is fixed for current month

IMAP log file issue from last month resolved and IMAP back to 100% availability

New: Powershell scripts experiencing intermittent failure

Improvements

- Content Delivery Network and DOMT tests added to address recent Incidents

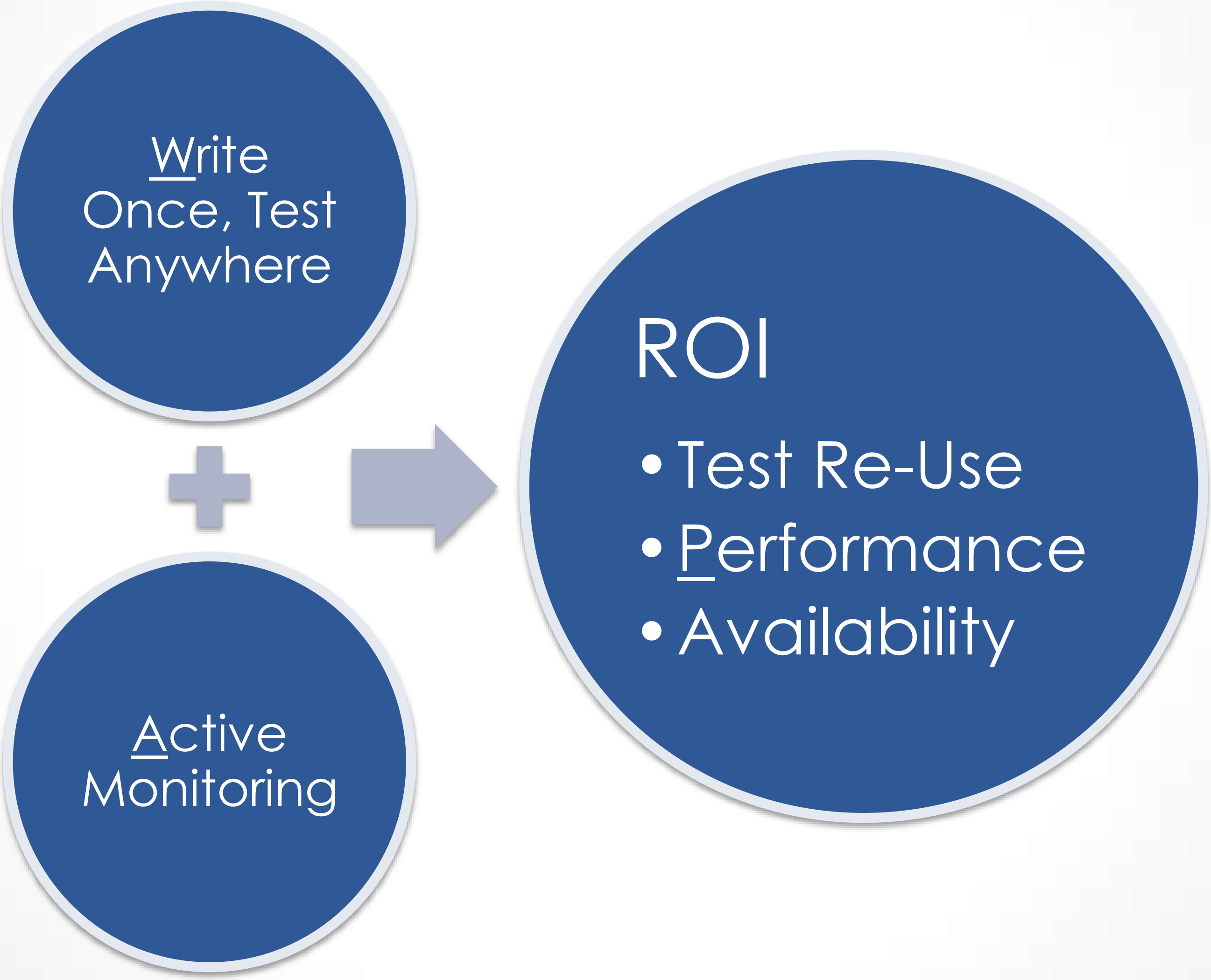
TIP Feature Availability

TIP Feature availability	Dec	Nov	Oct	Sep	Aug
Provisioning	99.93%	94.01%	100.00%	100.00%	100.00%
RBAC	99.93%	98.81%	100.00%	100.00%	99.93%
Outlook	100.00%	98.81%	100.00%	100.00%	99.95%
ECP	100.00%	97.18%	100.00%	100.00%	99.97%
Mailflow	100.00%	94.01%	100.00%	100.00%	99.50%
ActiveSync	100.00%	99.73%	100.00%	100.00%	99.63%
UM	100.00%	99.68%	100.00%	100.00%	99.68%
Calendaring	100.00%	100.00%	100.00%	100.00%	100.00%
Web Services	100.00%	100.00%	100.00%	100.00%	100.00%
OWA	100.00%	100.00%	100.00%	100.00%	100.00%
POP	100.00%	100.00%	100.00%	100.00%	100.00%
IMAP	100.00%	89.87%	100.00%	100.00%	100.00%
Remote PowerShell	99.93%	100.00%	100.00%	100.00%	100.00%

Microsoft Exchange Server 2010

[Deschamps, Johnston, Jan 2012]

More Testing, Less Cost



Write Once, Test Anywhere

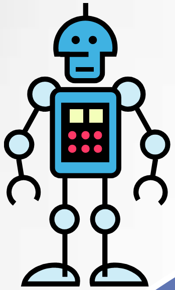
Active Monitoring

ROI

- Test Re-Use
- Performance
- Availability

TiP Methodologies - Inputs

- Where does the data driving your tests come from?



Synthetic Data

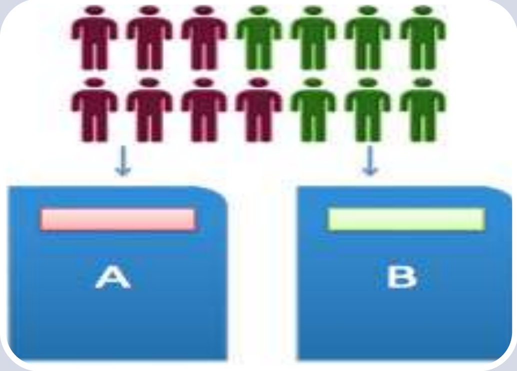
- Synthetic Tests in Production
- User Scenario Execution

Real Data

- Data Mining
- User Perf Testing
- Environment Validation
- Experimentation for Design
- Controlled Test Flight
- Dogfood/beta

TiP Methodologies - Effects

- Does the test change or act on production, and how?



No Change

Data Mining

User Perf. Testing

Environment Validation

Experiment with Users

Experim. for Design

Controlled Test Flight

Dogfood/Beta

Act on or Change Service

Synthetic TiP

User Scenario Execution

TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Validate the Data Center

And the Service

9 Load Testing in Production

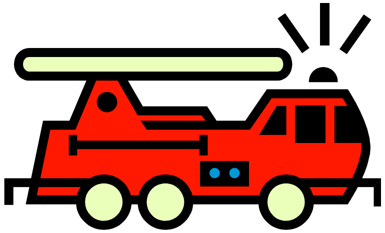
- Injects load on top of real user traffic
- Monitors for **performance**
- Employs alert and back-off protections
- Load should **not** go through CDN or data providers

bing™

- Identified major datacenter power issue
- From
  - 30 Engineers on a con-call
- To
  - 1.5 engineers and a framework

Operations

- "Ops" runs the data center
- Ops needs to be in the loop on TiP
  - Else they may react as if a real problem were occurring

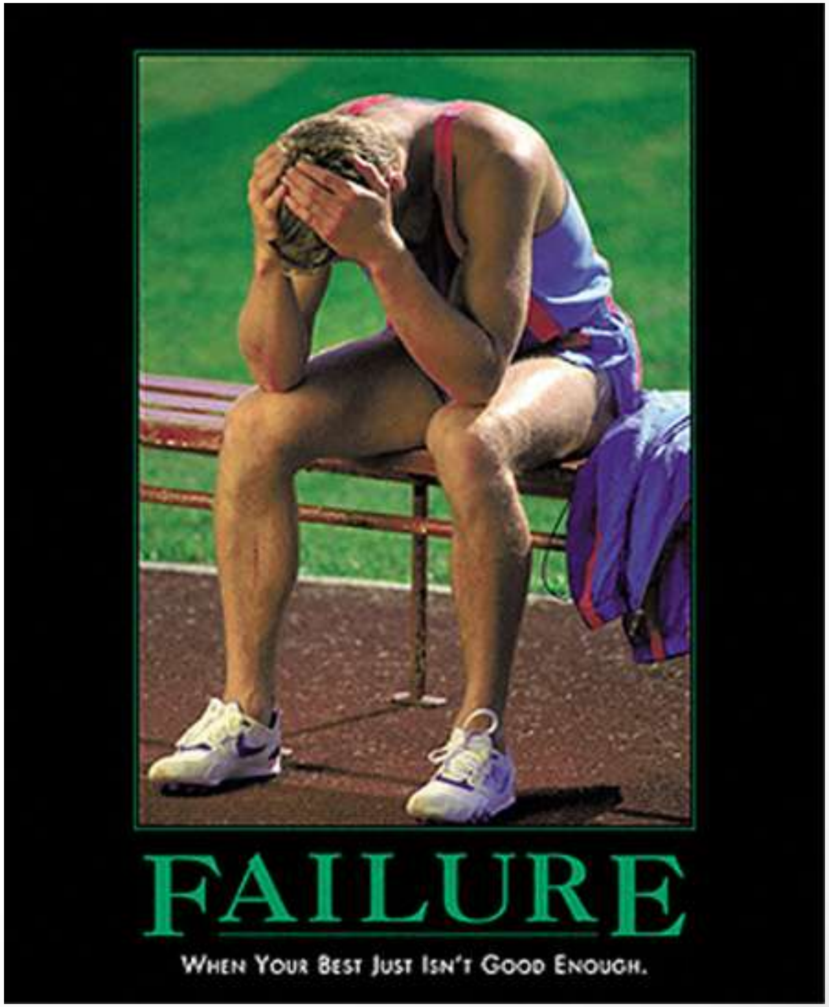


- Ops traditionally does monitoring
  - TiP is synergistic - TestOps
  - …but need to define roles and responsibilities

10 Destructive Testing in Production

- Google first year of a new data center [Google DC, 2008]
  - 20 rack failures, 1000 server failures and thousands of hard drive failures

- High Availability means you must Embrace Failure
  - How do you test this?



Netflix Tests its "Rambo Architecture"

- …system has to be able to succeed, no matter what, even all on its own
- Test with Fault Injection

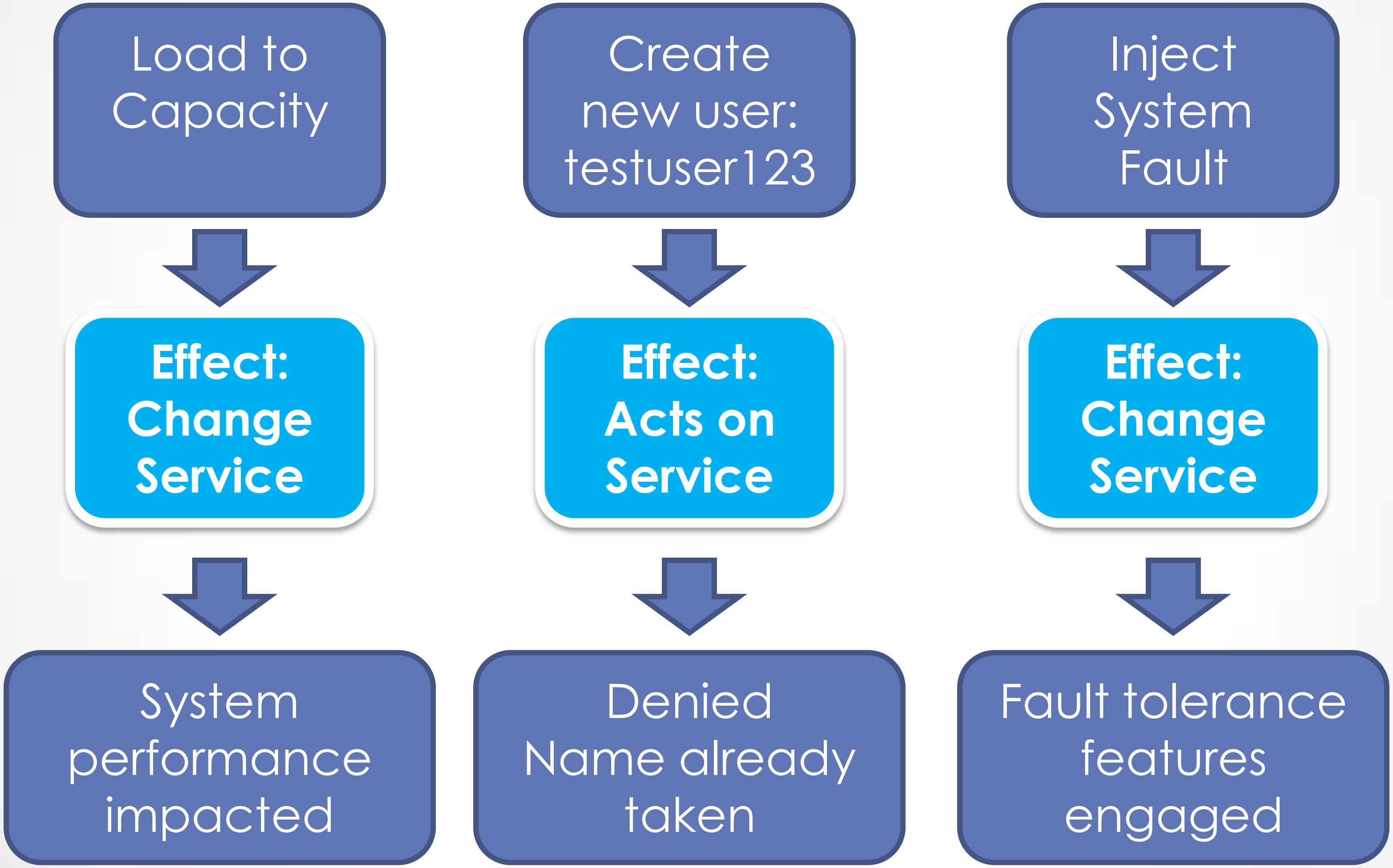
[Netflix Army, July 2011]

- Netflix *Simian Army*
  - **Chaos monkey** randomly kills production instance in AWS
  - **Chaos Gorilla** simulates an outage of an entire Amazon AZ
  - Janitor Monkey, Security Monkey, Latency Monkey…..

NETFLIX

amazon
web services™
NETFLIX
Streaming Instan

Effect: Change Service



Load to Capacity

Effect: Change Service

System performance impacted

Create new user: testuser123

Effect: Acts on Service

Denied Name already taken

Inject System Fault

Effect: Change Service

Fault tolerance features engaged

TiP Methodologies - Effects

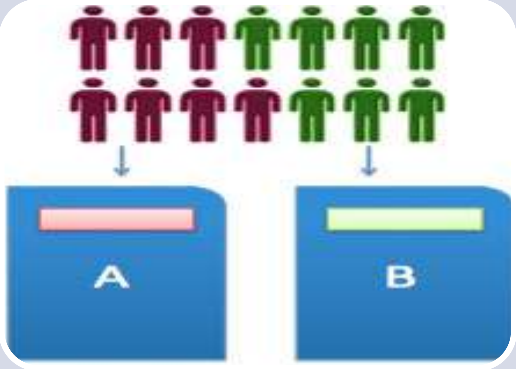
- Does the test change or act on production, and how?

No Change

Data Mining

User Perf. Testing

Environment Validation



Experiment with Users

Experim. for Design

Controlled Test Flight

Dogfood/Beta

Act on or Change Service

Synthetic TiP

User Scenario Execution

Load TiP

Destructive Testing

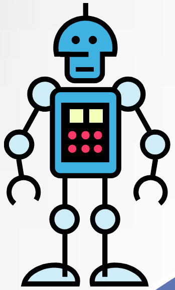
TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Methodology Wrap-up

TiP Methodologies - Inputs

- Where does the data driving your tests come from?



Synthetic Data

- Synthetic Tests in Production
- User Scenario Execution
- Load Testing in Production
- Destructive Testing

Real Data

- Data Mining
- User Perf Testing
- Environment Validation
- Experimentation for Design
- Controlled Test Flight
- Dogfood/beta

TiP Methodologies - Effects

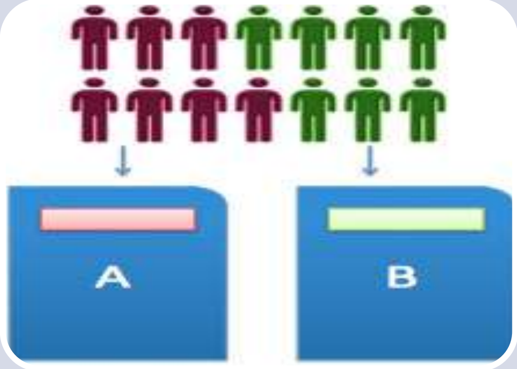
- Does the test change or act on production, and how?

No Change

Data Mining

User Perf. Testing

Environment Validation



Experiment with Users

Experim. for Design

Controlled Test Flight

Dogfood/Beta

Act on or Change Service

Synthetic TiP

User Scenario Execution

Load TiP

Destructive Testing

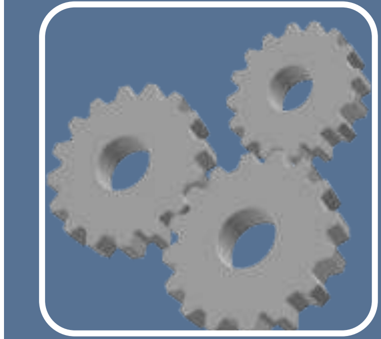
TiP Methodologies - Observations

- How do we measure the test results?

User Behavior

- Data Mining
- Experimentation for Design
- (Dogfood/Beta)

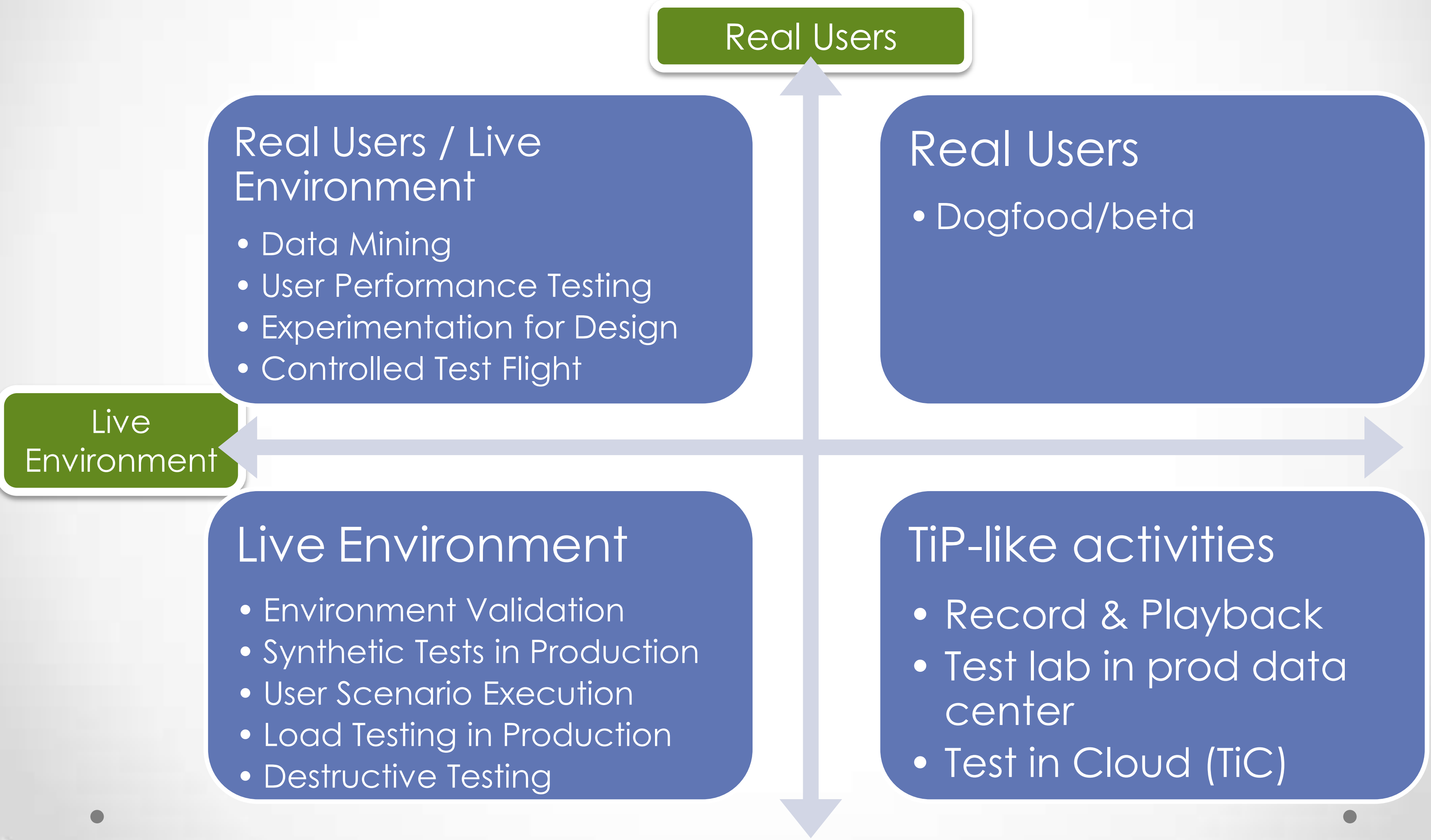
System Behavior



- User Performance Testing
- Controlled Test Flight
- Dogfood/beta
- Load Testing in Production
- Destructive Testing
- Environment Validation
- Synthetic Tests in Production
- User Scenario Execution

Real Users / Live Environments

…utilizes real users **and/or** live environments. ….



Real Users

Live Environment

Real Users / Live Environment

- Data Mining
- User Performance Testing
- Experimentation for Design
- Controlled Test Flight

Real Users

- Dogfood/beta

Live Environment

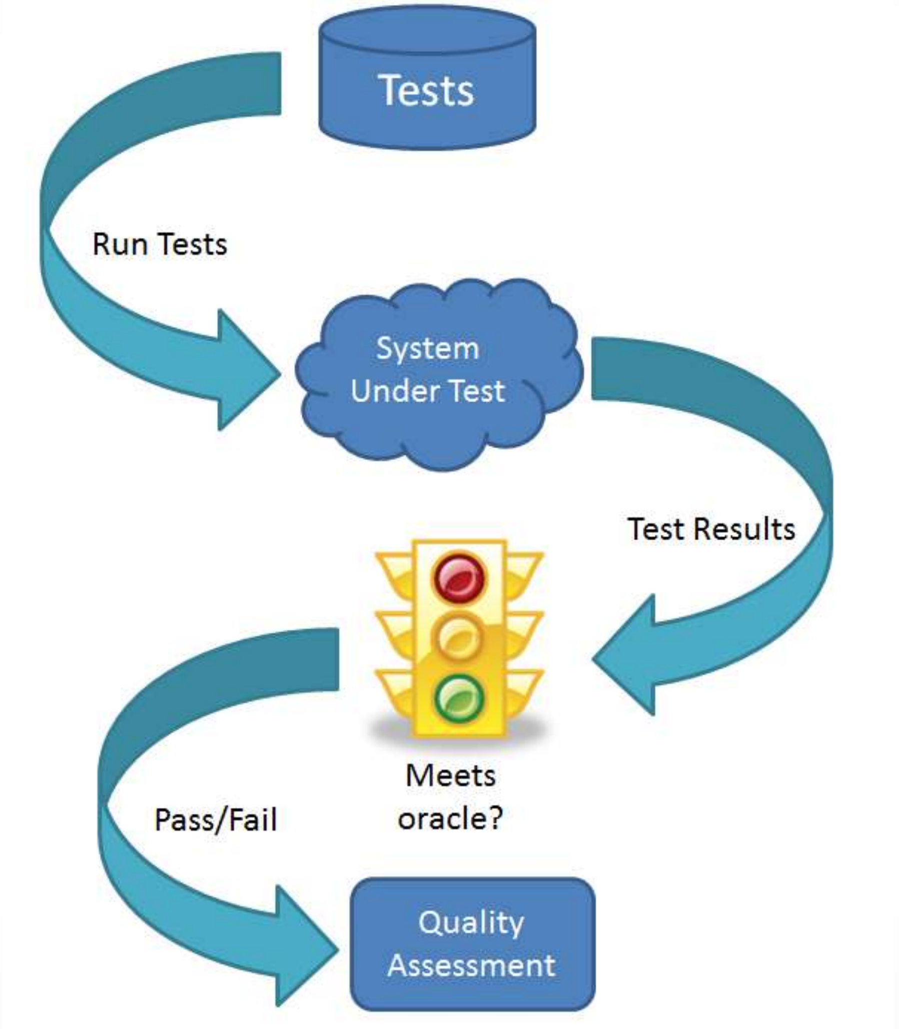
- Environment Validation
- Synthetic Tests in Production
- User Scenario Execution
- Load Testing in Production
- Destructive Testing

TiP-like activities

- Record & Playback
- Test lab in prod data center
- Test in Cloud (TiC)

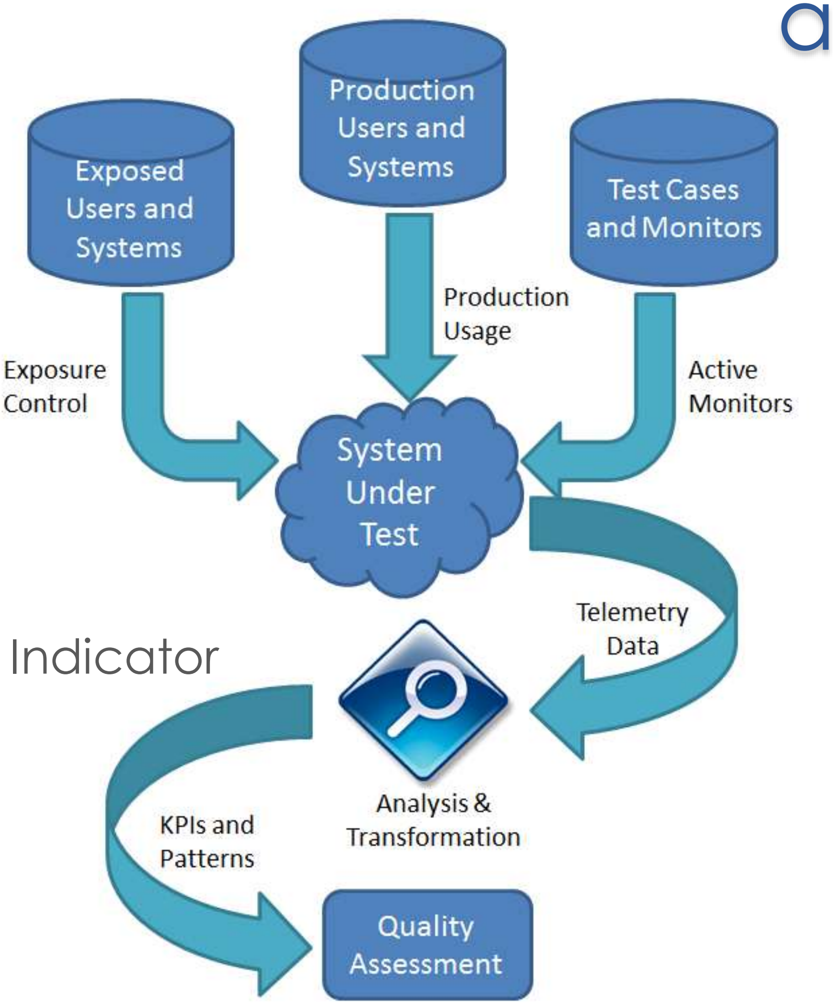
Changing the Quality Signal

Traditional Quality Signal



Big Data Quality Signal

aka TestOps



KPI: Key Performance Indicator

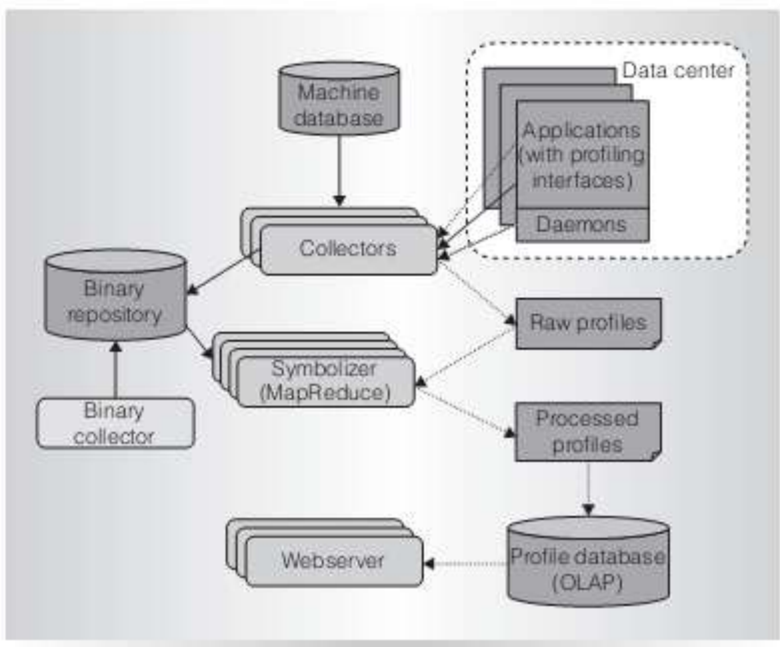
- Request latency
- RPS
- Availability / MTTR

How Big?

- **Google**: more than 3 billion searches, 2 billion video replays and absorbs 24 hours of video per minute.
- **Microsoft Bing** has grown from 10% of U.S. Searches in September 2010 to over 30% as of April 2011
    - Cosmos ingests 1-2 PB/day
- **Facebook**: 800 million active users sharing 30 billion pieces of content per month.
- **Amazon**: more than 120 million user accounts, 2 million sellers and 262 billion objects stored in its S3 cloud storage system.
- **Twitter** reached its one-billionth tweet after just 3 years, 2 months and 1 day.

Google-Wide Profiling (GWP)

- Continuous profiling infrastructure for data centers - draw **performance** insights
- Collects stack traces, hardware events, kernel events etc.,
- From several thousand applications running on thousands of servers
- Compressed profile database grows by several GB every day.



- What are the hottest processes, routines, or code regions?
- How does performance differ across software versions?
- Which locks are most contended?

[Google-Wide Profiling, 2010]

TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

What NOT to Wear!

What Not to Do

…

TiP is a common and costly technical malpractice

“Simply, you should have a separate test system that is identical to your production system”

- If you do not have the resources for this, instead maintain:
  - a scaled-down environment with load generators that duplicate the expected load.
  - a model or simulation of the environment.

“One mistake in particular was that the bank had created two application servers on a single installation of WebSphere Application Server base”

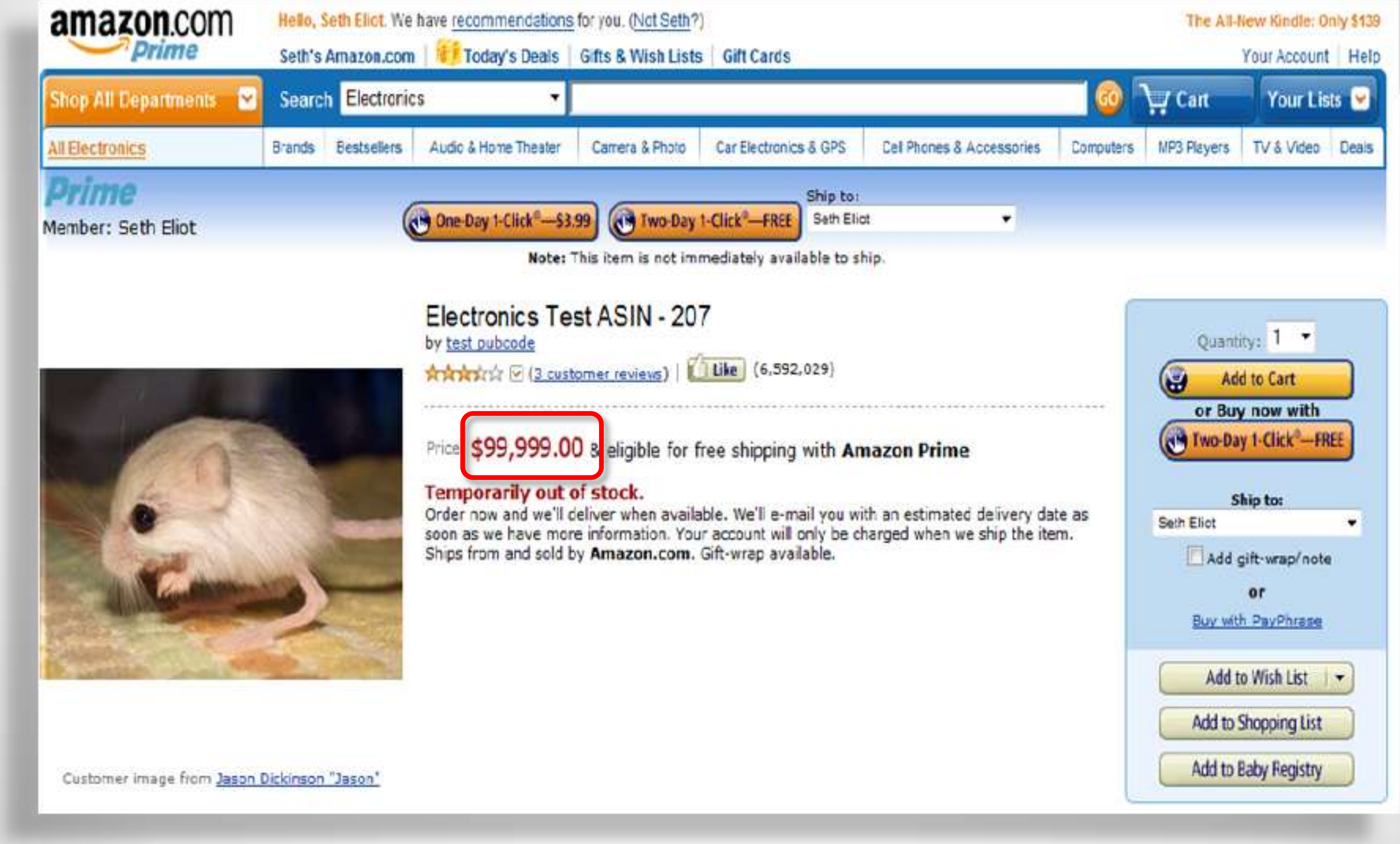
- Because both [prod and test]… ran on the same base WebSphere Application Server
  - their logs are shared
  - any upgrade to the SDK would disrupt both application servers

[IBM, 2011]

What Not to Do, IBM

- Disrupt user experience introduced by testing: for example: outages
- Failure to understand the production environment and the effect of TiP on it.
- Co-mingling/corrupting production data

Wireless Mouse?



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

Member: Seth Eliot

Ship to: Seth Eliot

Note: This item is not immediately available to ship.

Electronics Test ASIN - 207

by test pubcode

☆☆☆☆☆ (3 customer reviews) | Like (6,592,029)

Price: $99,999.00 & eligible for free shipping with **Amazon Prime**

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

or Buy now with

Ship to: Seth Eliot

□ Add gift-wrap/note

or

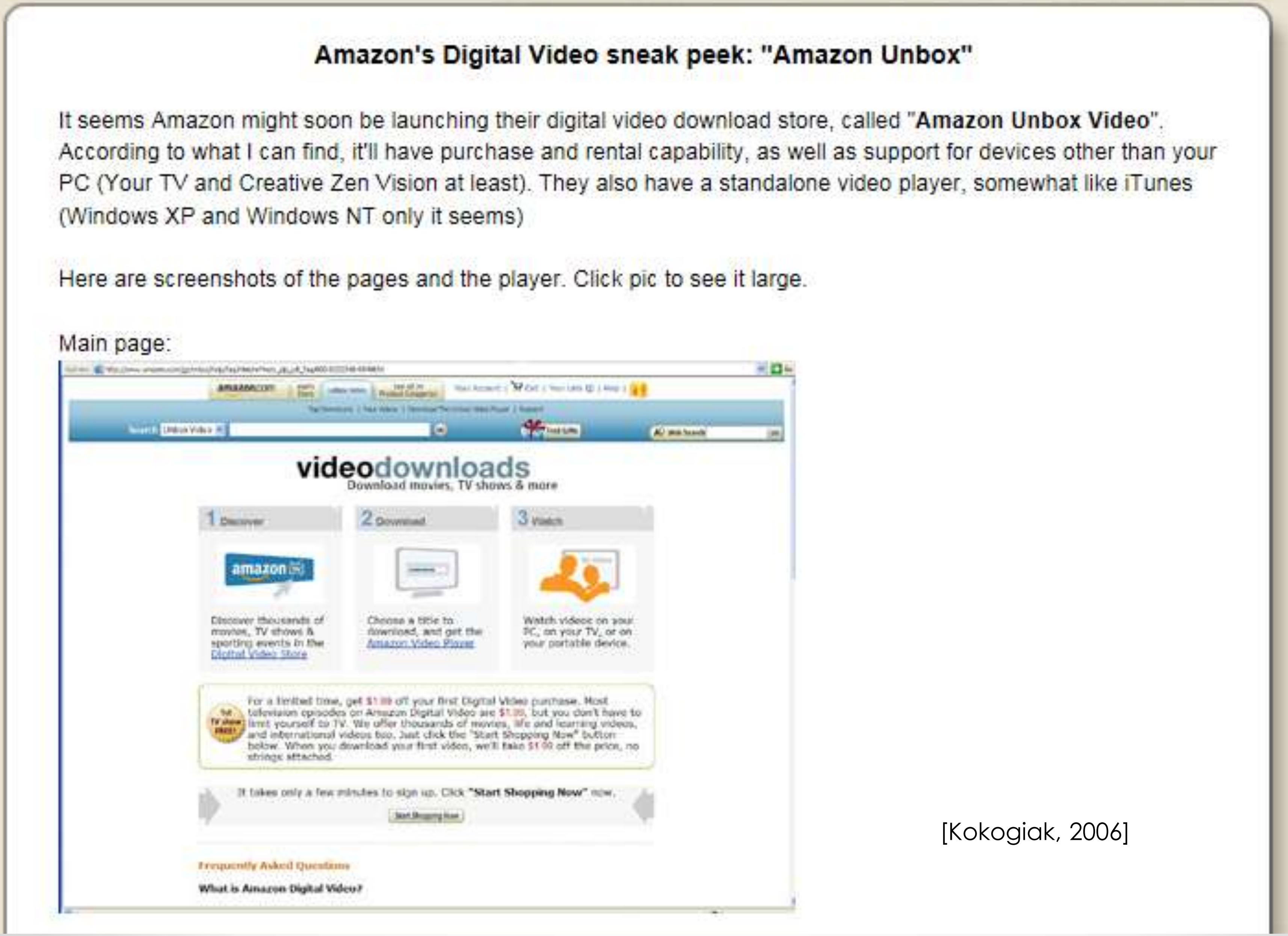
Buy with PayPhrase

Customer image from Jason Dickinson "Jason"

What Not to Do, Amazon

- Exposure of test artifacts to end users: e.g. exposed test items, test emails sent
- Misuse of PII customer data
- Leaking sensitive new features prior to official launch

Amazon's Digital Video sneak peek



Amazon's Digital Video sneak peek: "Amazon Unbox"

It seems Amazon might soon be launching their digital video download store, called "**Amazon Unbox Video**". According to what I can find, it'll have purchase and rental capability, as well as support for devices other than your PC (Your TV and Creative Zen Vision at least). They also have a standalone video player, somewhat like iTunes (Windows XP and Windows NT only it seems)

Here are screenshots of the pages and the player. Click pic to see it large.

Main page:

Discover thousands of movies, TV shows & sporting events in the Digital Video Store

Choose a title to download, and get the Amazon Video Player

Watch videos on your PC, on your TV, or on your portable device.

For a limited time, get $1.99 off your first Digital Video purchase. Most television episodes on Amazon Digital Video are $1.99, but you don't have to limit yourself to TV. We offer thousands of movies, life and learning videos, and international videos too. Just click the "Start Shopping Now" button below. When you download your first video, we'll take $1.99 off the price, no strings attached.

It takes only a few minutes to sign up. Click **"Start Shopping Now"** now.

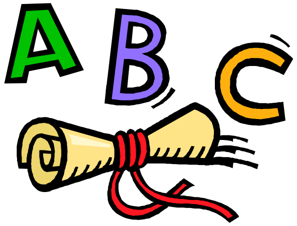
Frequently Asked Questions

What is Amazon Digital Video?

[Kokogiak, 2006]

TiP A-Z

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy



Summary

• • •

The latest version of this slide deck can be found at:
http://bit.ly/seth_stp_2012

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Summary

A	Active Monitoring	N	Not to Do, What
B	Big Data	O	Operations
C	Client Instrumentation	P	Performance Testing
D	Data Mining	Q	Quality Signal
E	Experimentation	R	Real Data Input
F	Fault Injection	S	Service, Effect on; Synthetic Data Input
G	Go / No-go for deployment	T	Three Stages
H	High Availability	U	Users, Experiment with
I	Iterative Virtuous Cycle	V	Validation in Data Center
J	JSI: J-script Instrumentation	W	Write Once, Test Anywhere
K	Kill production instances	X	eXposure Control
L	Load Testing in Production	Y	Y TiP?
M	Methodologies	Z	Zymurgy

Time for a Beer…

Zymurgy

zy·mur·gy (z mûr j) n. The branch of chemistry that deals with fermentation processes, as in brewing.

References

[Google Talk, June 2007]	Google: Seattle Conference on Scalability: Lessons In Building Scalable Systems, Reza Behforooz http://video.google.com/videoplay?docid=6202268628085731280
[Unpingco, Feb 2011]	Edward Unpingco; Bug Miner; Internal Microsoft Presentation, Bing Quality Day
[Barranco, Dec 2011]	René Barranco; Heuristics-Based Testing; Internal Microsoft Presentation
[Dell, 2012]	http://whichtestwon.com/dell%e2%80%99s-site-wide-search-box-test
[Microsoft.com, TechNet]	http://technet.microsoft.com/en-us/library/cc627315.aspx
[Cockcroft, March 2012]	http://perfcap.blogspot.com/2012/03/ops-devops-and-noops-at-netflix.html
[Deschamps, Johnston, Jan 2012]	Experiences of Test Automation; Dorothy Graham; Jan 2012; ISBN 0321754069; Chapter: "Moving to the Cloud: The Evolution of TiP, Continuous Regression Testing in Production"; Ken Johnston, Felix Deschamps
[Google DC, 2008]	http://content.dell.com/us/en/gen/d/large-business/google-data-center.aspx?dgc=SM&cid=57468&lid=1491495 http://perspectives.mvdirona.com/2008/06/11/JeffDeanOnGoogleInfrastructure.aspx

References, *continued*

[Netflix Army, July 2011]	The Netflix Simian Army; July 2011 http://techblog.netflix.com/2011/07/netflix-simian-army.html
[Google-Wide Profiling, 2010]	Ren, Gang, et al. Google-wide Profiling: A Continuous Profiling Infrastructure for Data Centers. [Online] July 30, 2010. research.google.com/pubs/archive/36575.pdf
[Facebook ships, 2011]	http://framethink.blogspot.com/2011/01/how-facebook-ships-code.html
[Google BusinessWeek, April 2008]	How Google Fuels Its Idea Factory, BusinessWeek, April 29, 2008; http://www.businessweek.com/magazine/content/08_19/b4083054277984.htm
[IBM 2011]	http://www.ibm.com/developerworks/websphere/techjournal/1102_supauth/1102_supauth.html
[Kokogiak, 2006]	http://www.kokogiak.com/gedankengang/2006/08/amazons-digital-video-sneak-peek.html
[Google GTAC 2010]	Whittaker, James. GTAC 2010: Turning Quality on its Head. [Online] October 29, 2010. http://www.youtube.com/watch?v=cqwXUTjcabs&feature=BF&list=PL1242F05D3EA83AB1&index=16.
[Google, JW 2009]	http://googletesting.blogspot.com/2009/07/plague-of-homelessness.html
[STPCon, 2012]	STPCon Spring 2012 - Testing Wanted: Dead or Alive – March 26, 2012

Thank You

Session 503

A to Z Testing in Production

Seth Eliot

Thank you for attending this session.
Please fill out an evaluation form.