GDC 12

Don’t trust the HiPPOs:
A/B Testing Online Games

Steve Collins
CTO / Swrve

GAME DEVELOPERS CONFERENCE®
SAN FRANCISCO, CA
MARCH 5-9, 2012
EXPO DATES: MARCH 7-9
2012

Me, me, me.

1986 1998 2007 2010

The HiPPO

Highest Paid Person's Opinion

http://www.kaushik.net - Occam's Razor Blog

Example #1

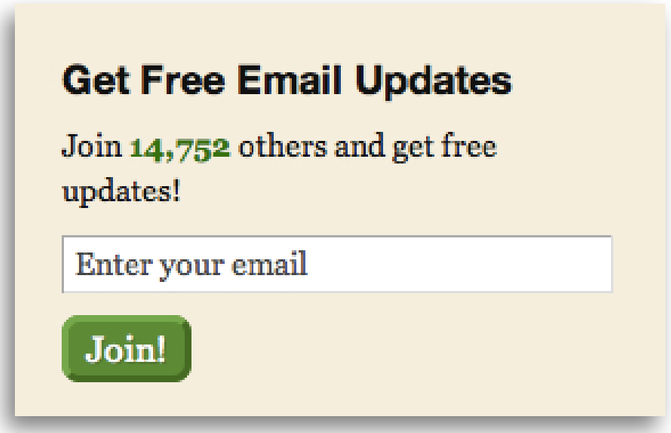
A

+218%

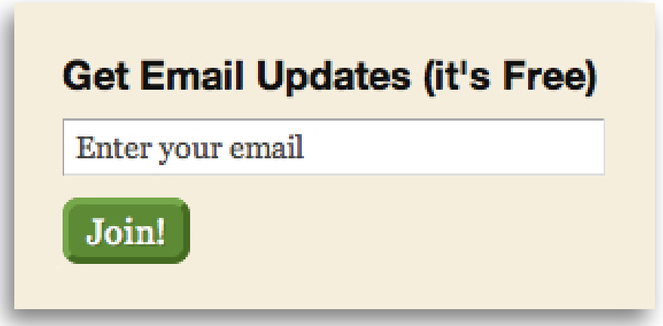
B

www.abtests.com and skritter.com

Example #2



A



B

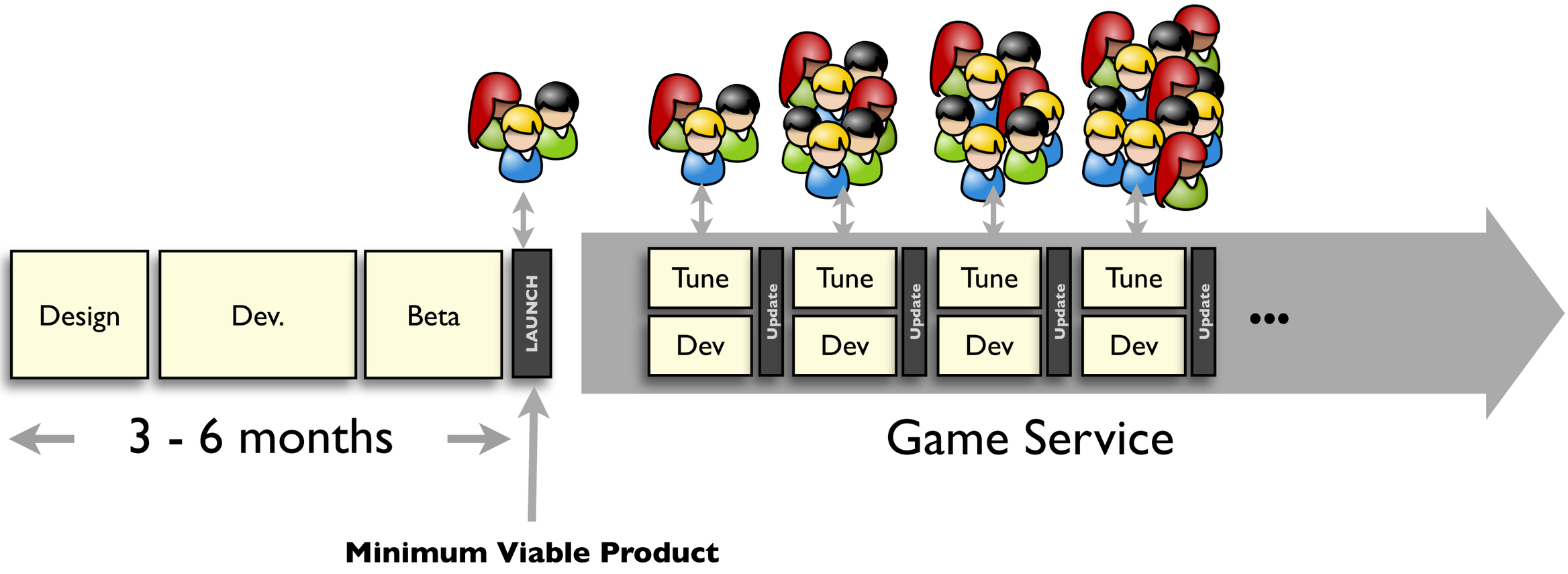
+102%

www.abtests.com and diythemes.com

"One accurate measurement is worth 1,000 expert opinions."

Admiral Grace Murray Hopper

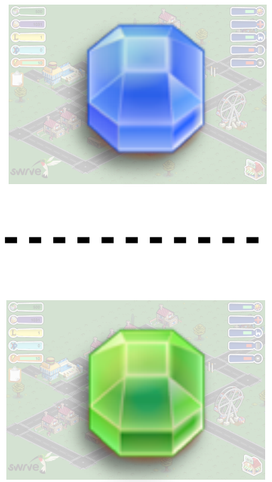
Game Service



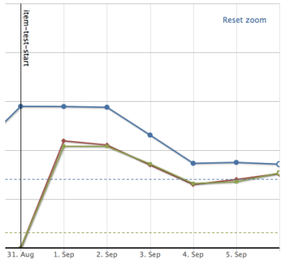
What is testing?

A/B Testing Overview

1. Split population



2. Show variations



3. Measure response

4. Choose winner

5. Deploy winner

What not to do...

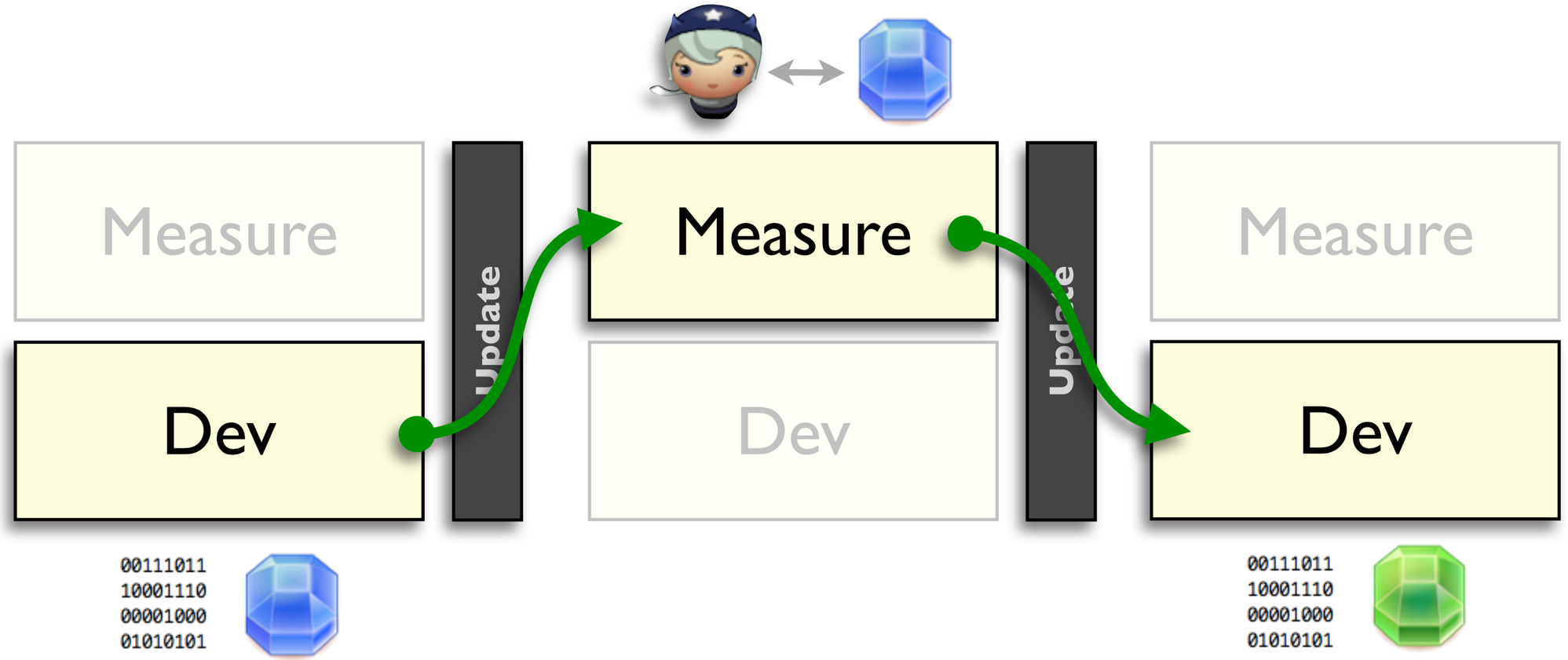
```
// Check membership of gems test?
if(ABTestHash(userID) > 0.50f)
{
    // Show variant of gems test
    icon_resouce = "cdn:/icons/gemos/gem01.png";

    // Check membership of popup103-UI-layout test
    if(!ABTestHash(userID) < 0.33f)
    {
        // in gems test but also in layout test
        layout_resouce = "cdn:/UI/XML/gem-layout-screen02.png";
    } else
    {
        // in gems test, but not in layout test
        layout_resouce = "cdn:/UI/XML/gem-layout-screen02.png";
    }
} else
{
    // Show control for gems test
    icon_resouce = "cdn:/icons/gemos/gem01.png";
}
```

- ✕ Inflexible
- ✕ Error prone
- ✕ Complex
- ✕ Forces app update

Solution: Data Driven Approaches

http://bit.ly/oWVvX3

Serial Cycles

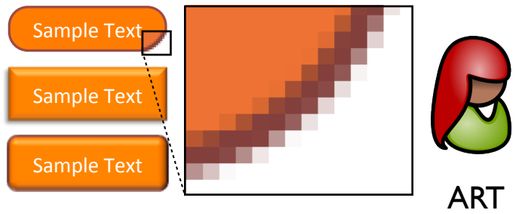


Meta-data

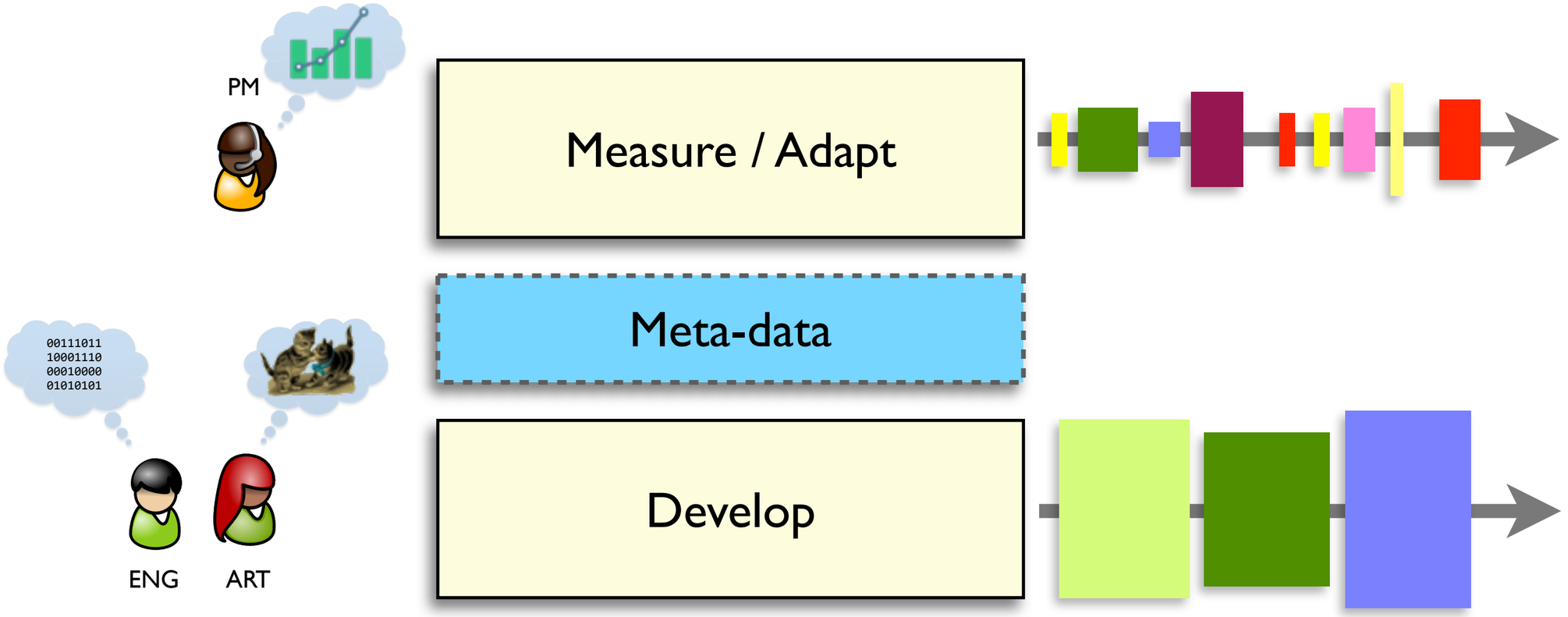
Add to cart =

button-style	3D-bevel
call-to-action	“Add to cart”
colour	Orange

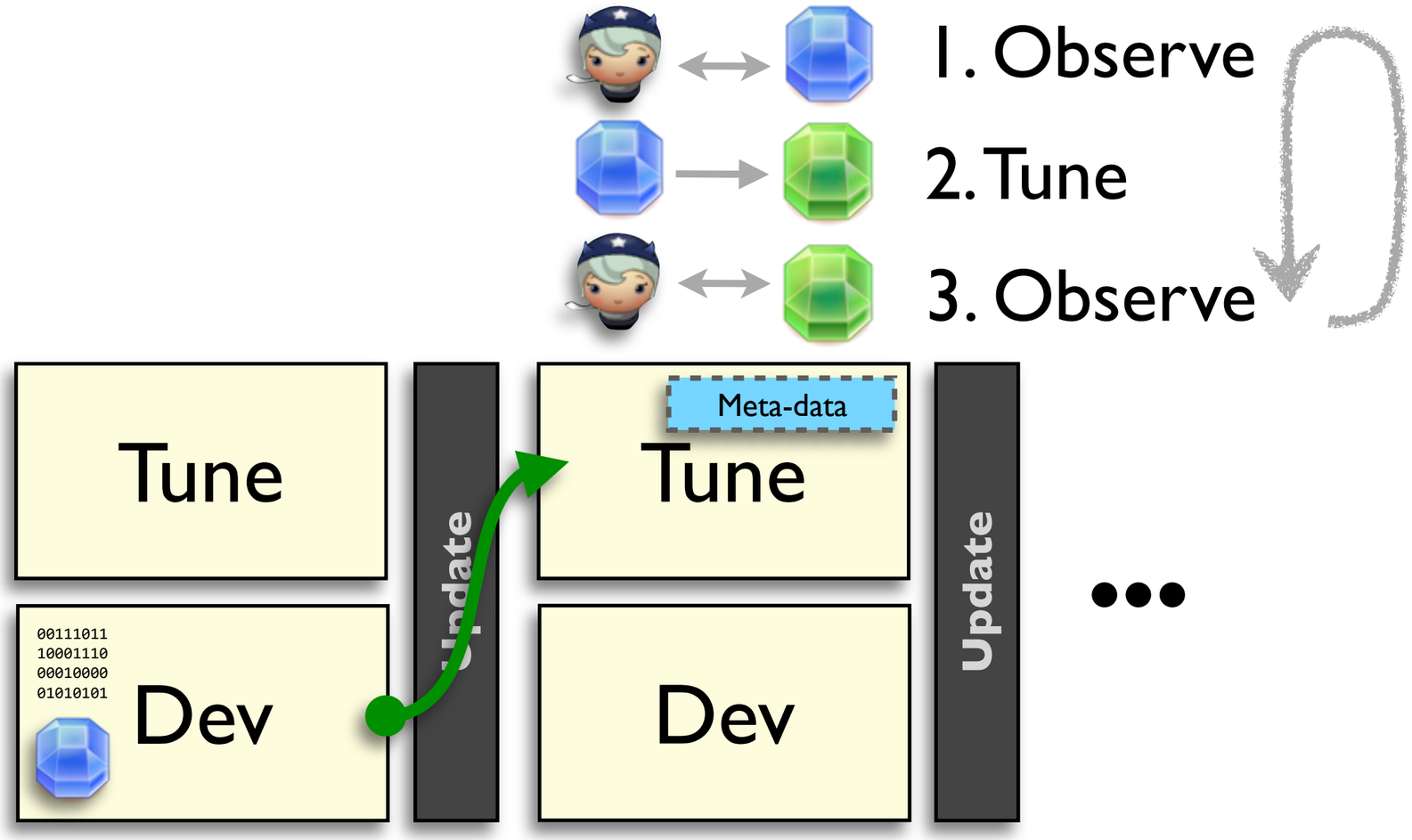

```
this.__defineSetter__('time', function(t) {
  var validMorphs = [];
  var morphDict = this.points.morphTargetDictionary;
  for(var k in morphDict) {
    if(k.indexOf('morphPadding') < 0) {
      validMorphs.push(morphDict[k]);
    }
  }
  validMorphs.sort();
  var l = validMorphs.length-1;
  var scaledt = t*l+1;
  var index = Math.floor(scaledt);
  for (i=0;i<validMorphs.length;i++) {
    this.points.morphTargetInfluences[validMorphs[i]] = 0;
  }
```



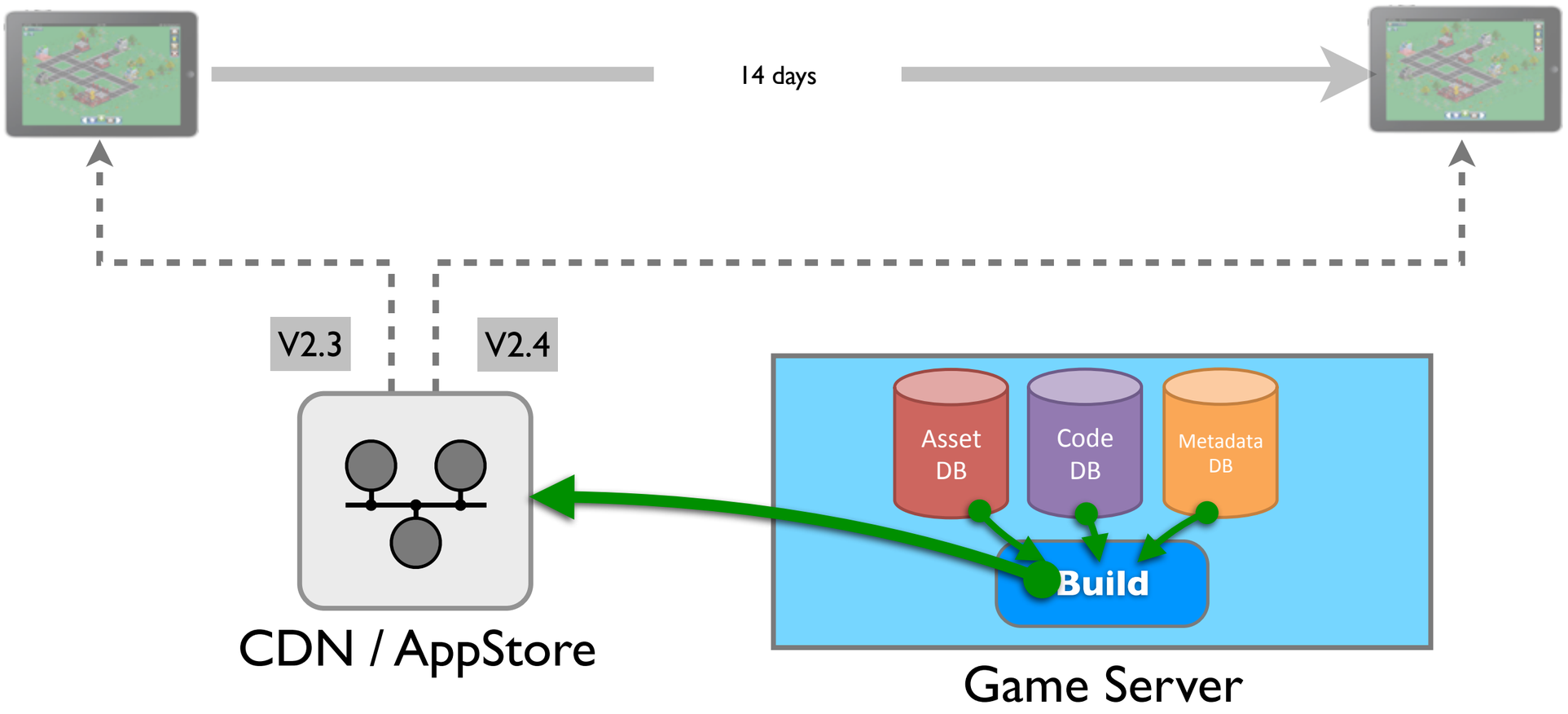
Serial Cycles



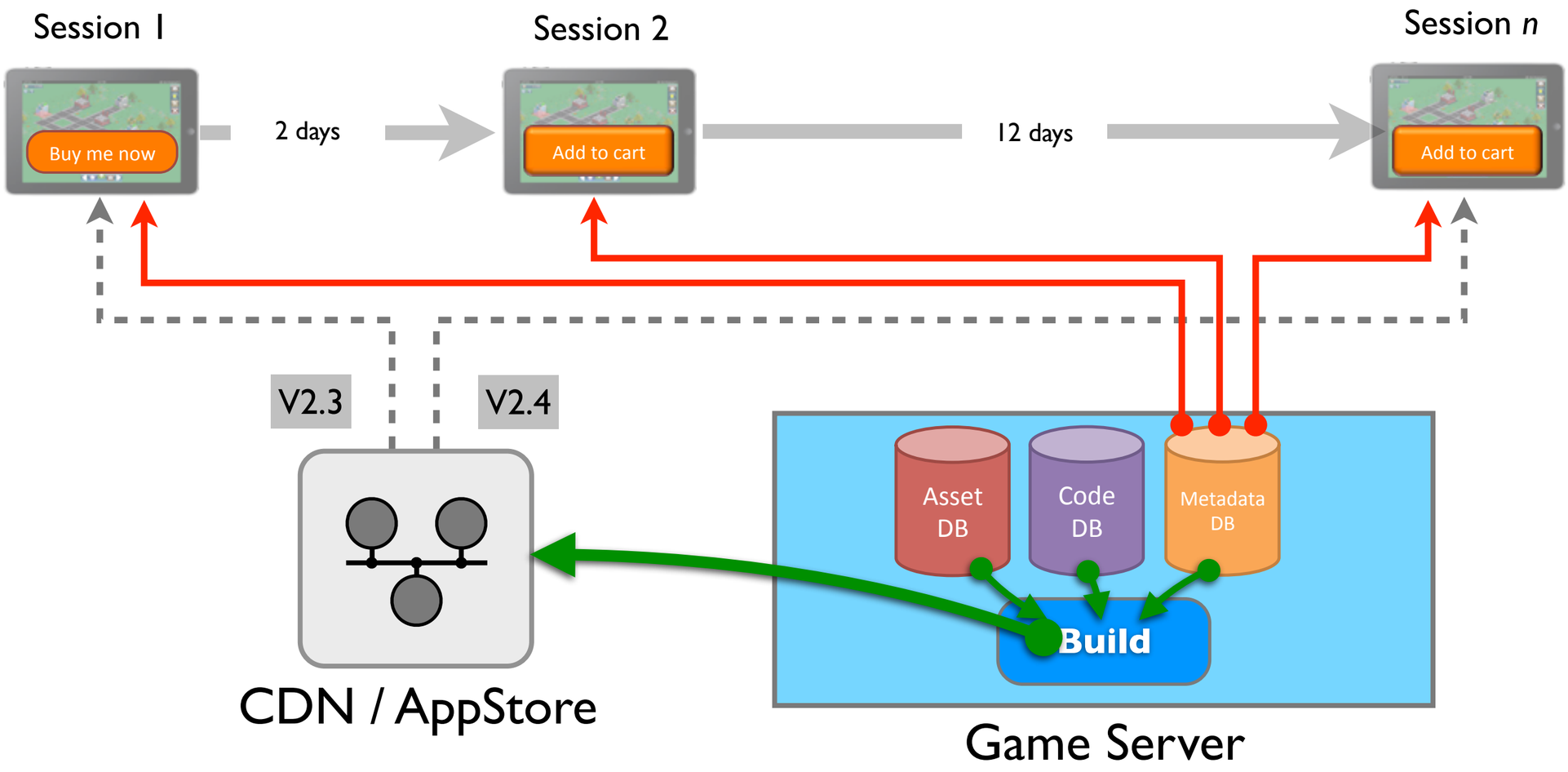
Parallel Cycles



Sessions

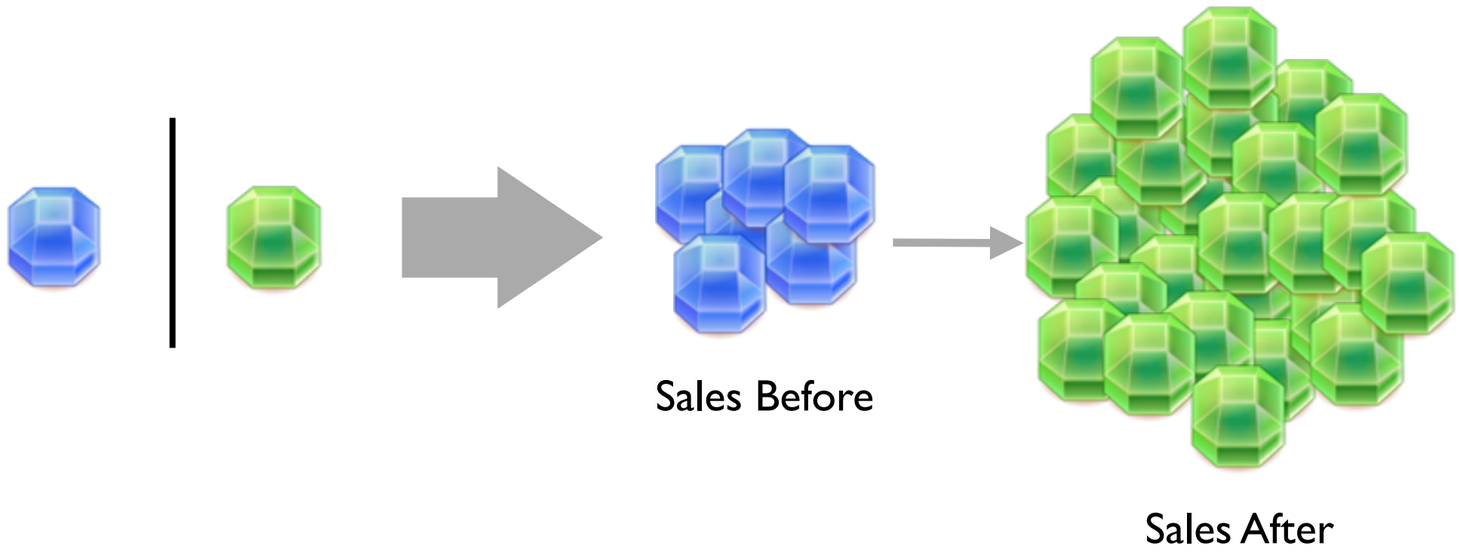

14 days
V2.3
V2.4
CDN / AppStore
Asset DB
Code DB
Metadata DB
Build
Game Server

Sessions

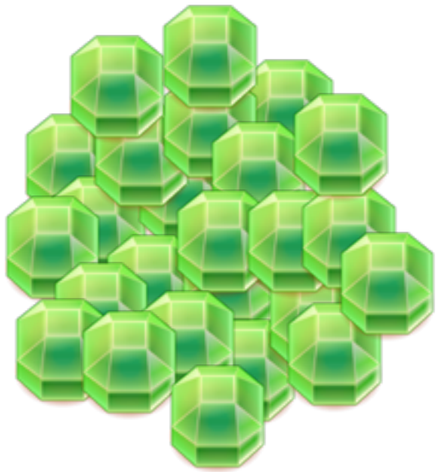

Session 1
Session 2
Session n
Buy me now
2 days
Add to cart
12 days
Add to cart
V2.3
V2.4
CDN / AppStore
Asset DB
Code DB
Metadata DB
Build
Game Server

Implementing Testing

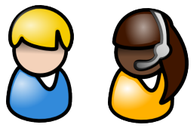
State Hypothesis



Agree OEC (overall evaluation criterion)



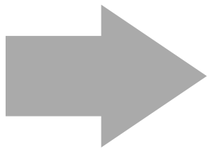
%age of players who purchase



1. Split population

User - Bucket Assignment

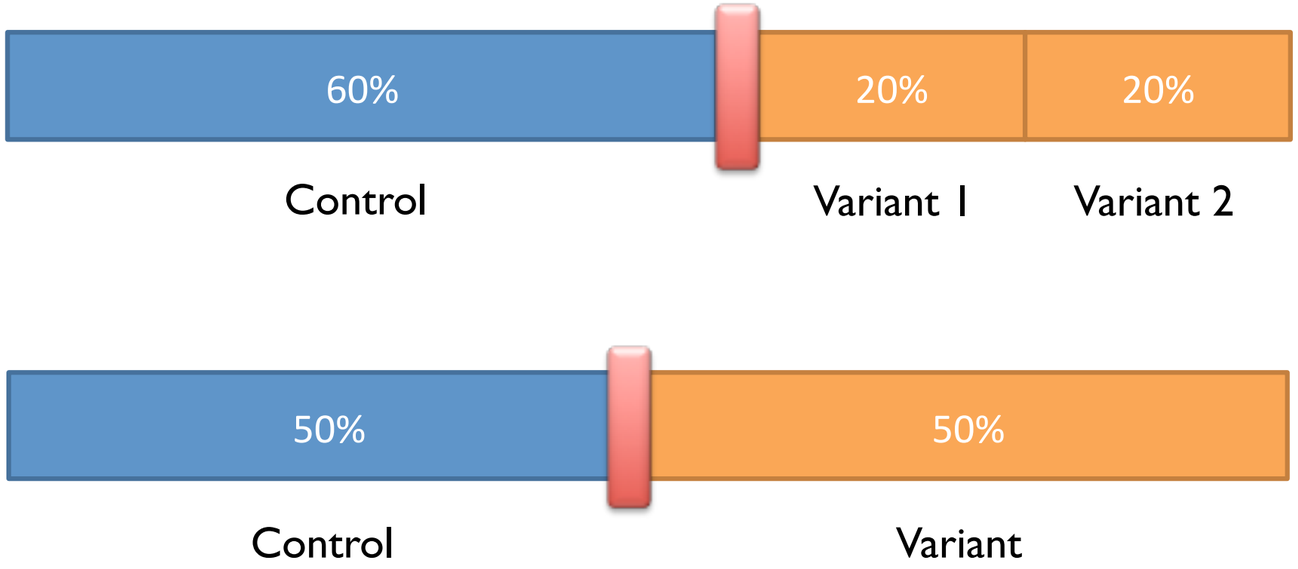
Consistent
Unbiased
Efficient



```
md5_hash(UUID + testID)
```

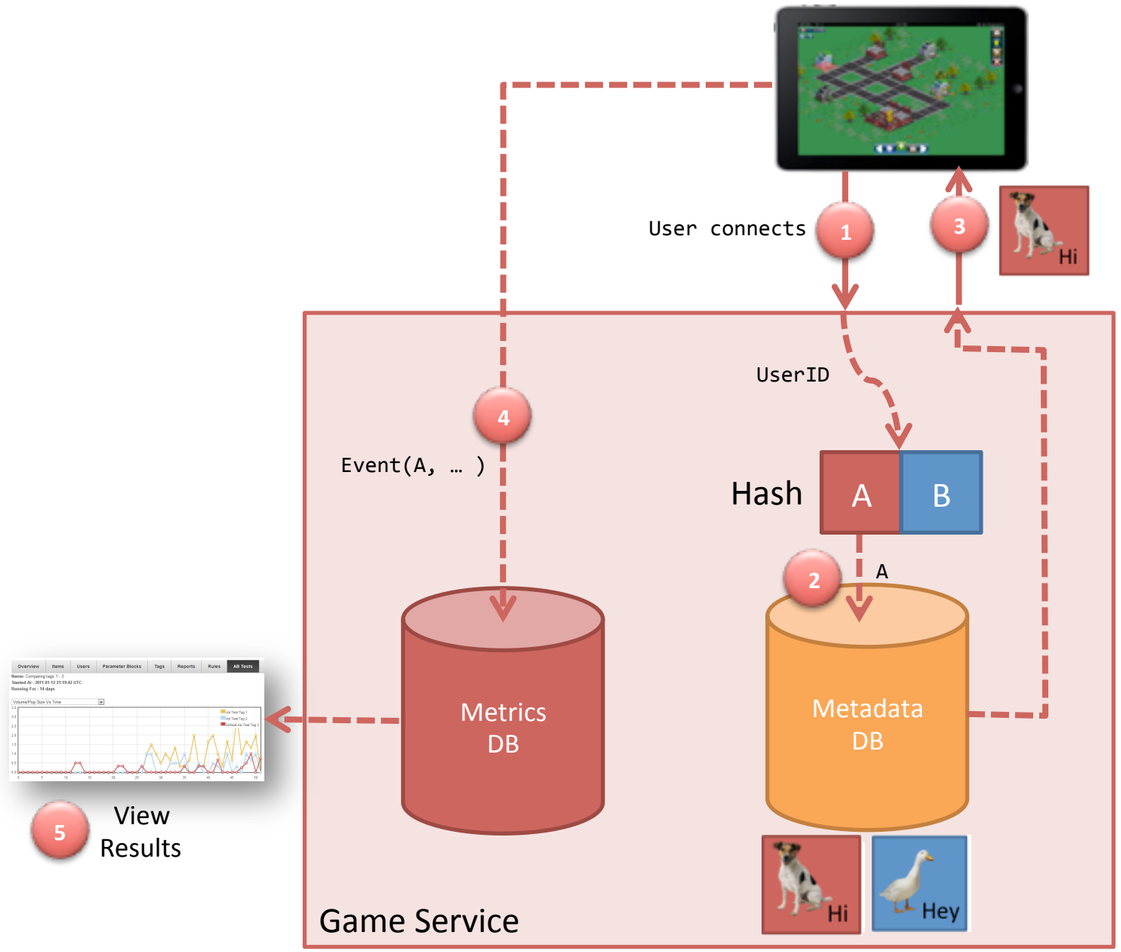
Ensure users are bucketed independently for each test

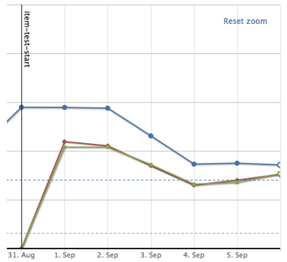
Split population (into independent groups)



2. Show variations

Testing Architecture





3. Measure response

An incredibly brief summary of...

Statistics of testing

Conversions

Conversion Event {
- Item Purchase
- Buy-in
- Completed Tutorial
- Fired in-game event
- Added a friend
- Watched movie
- ...

#players = n

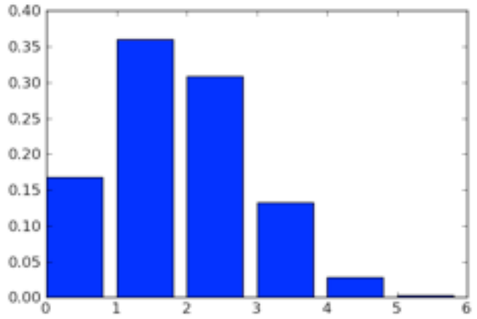
#conversions = r

conversion rate = $r/n = p$

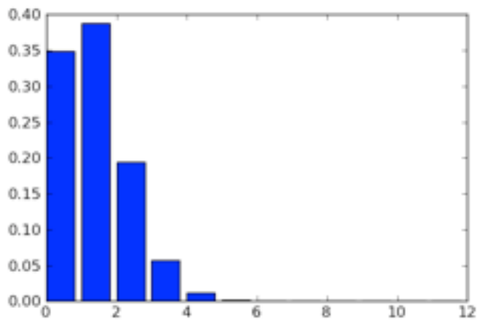
Binomial Distribution

$$P[r,n] = \frac{n!}{(n-r)!r!}\, p^r (1-p)^{n-r}$$

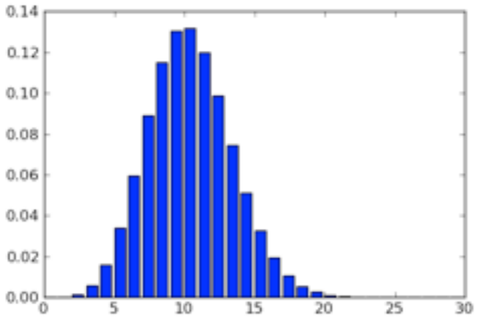
Probability of r successes given n trials and probability of success p per trial



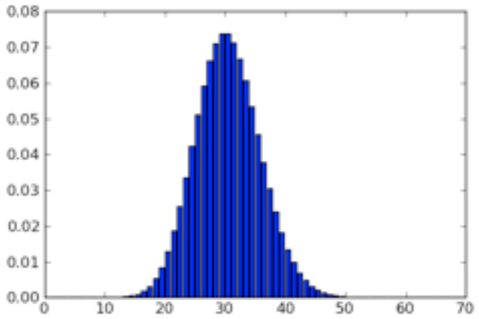
n = 5, p = 0.3



n = 100, p = 0.1



n = 100, p = 0.1



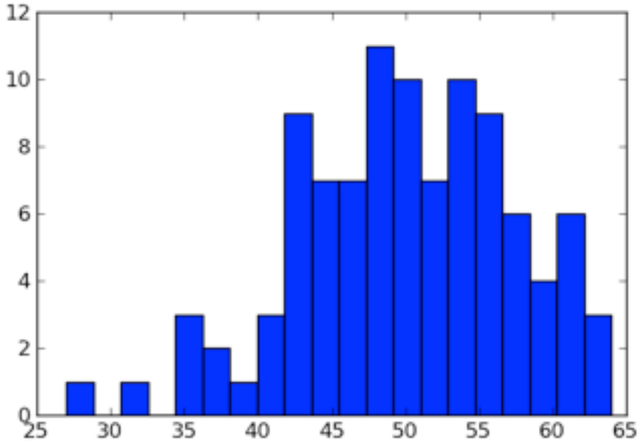
n = 1000, p = 0.05

Binomial Distribution

Run 100 "experiments", $n = 1000$, $p = 0.05$
and compute the average "conversion rate"

$$\mu = np \quad \sigma = \sqrt{np(1-p)}$$

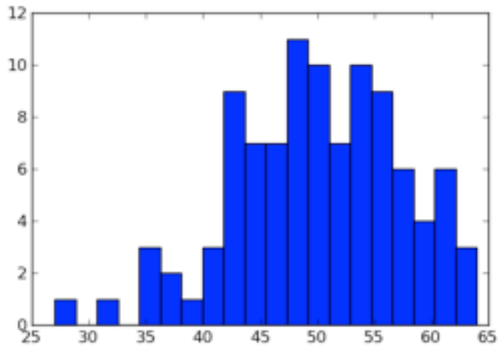
Expected value is 50



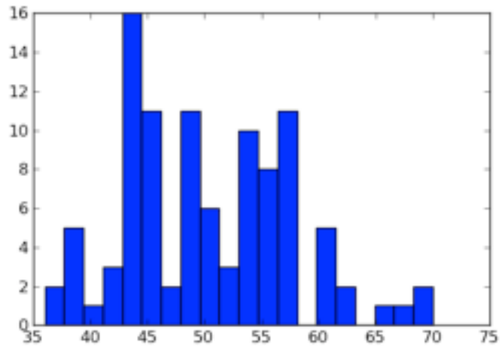
$\overline{x} = 49.96$

Binomial Distribution

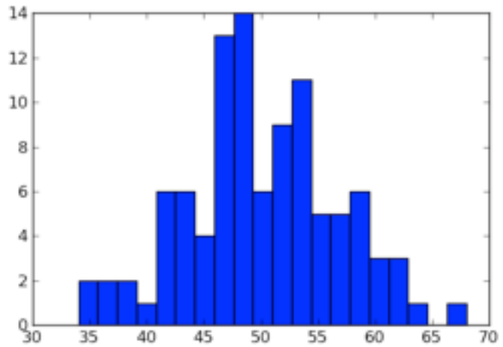
Repeat this process...



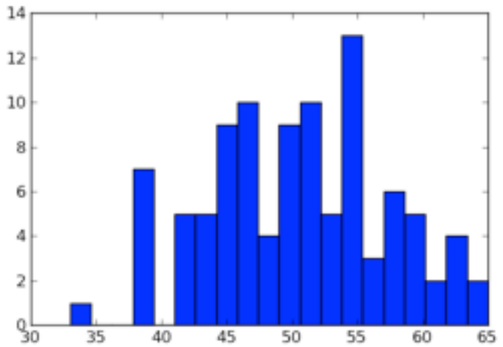
$\overline{x} = 49.96$



$\overline{x} = 50.30$



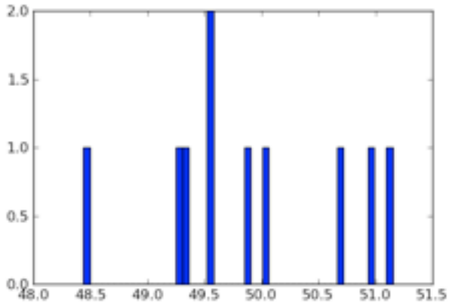
$\overline{x} = 49.98$



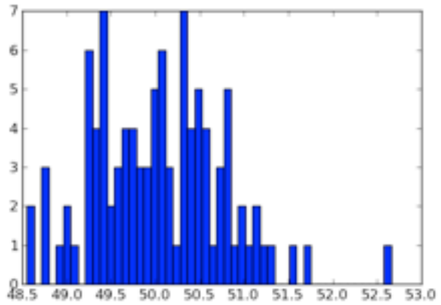
$\overline{x} = 50.36$

Binomial Distribution

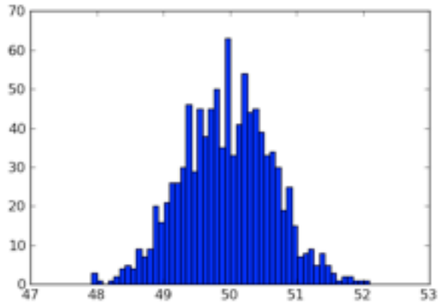
Keep repeating this process and plot the averages...



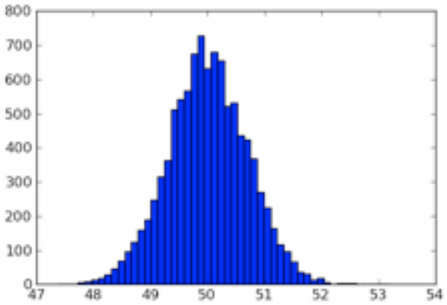
10 runs



100 runs



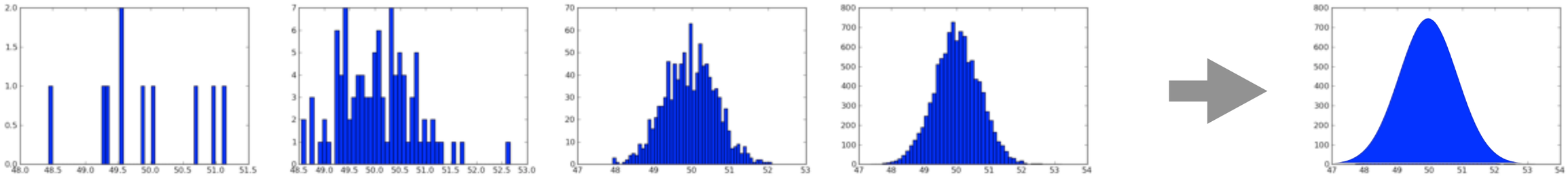
1000 runs



10000 runs

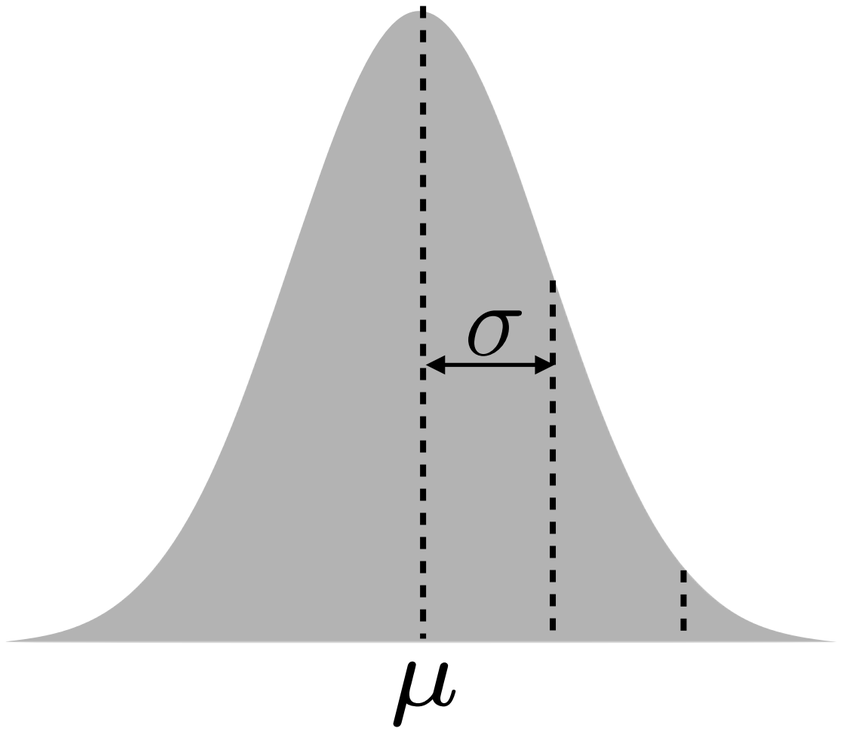
These are *sampling distributions of the mean*

Central Limit Theorem



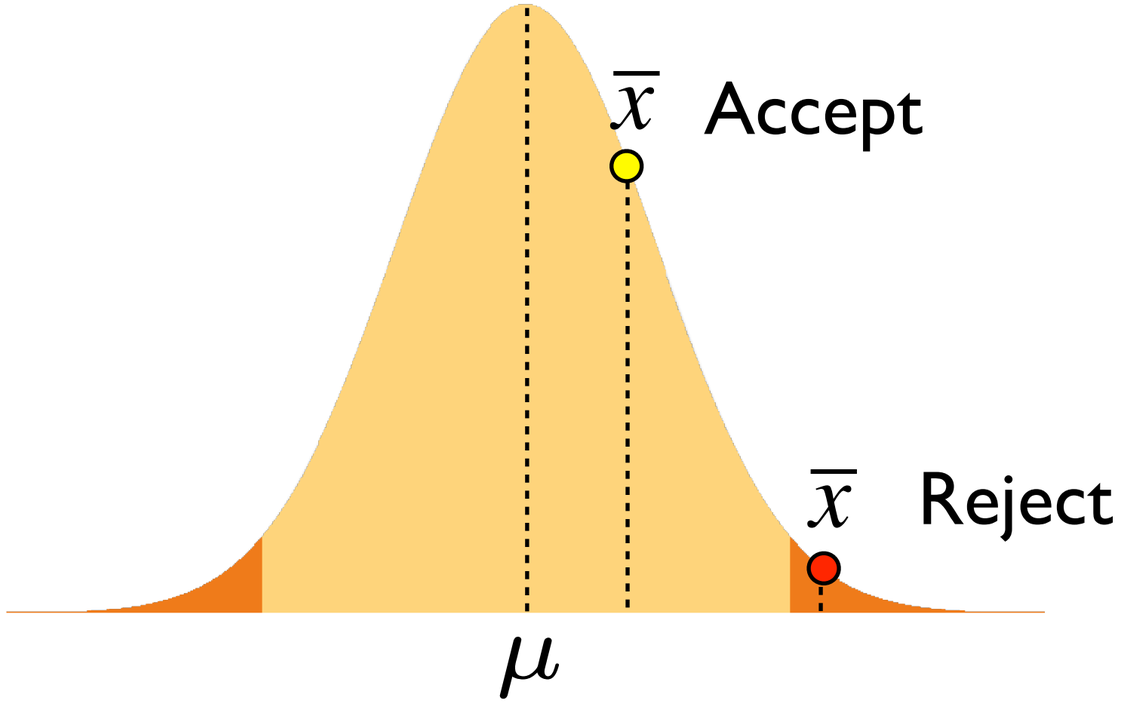
As *n* increases, the sampling distribution of the mean becomes "normal", independent of the underlying distribution

The Normal Distribution



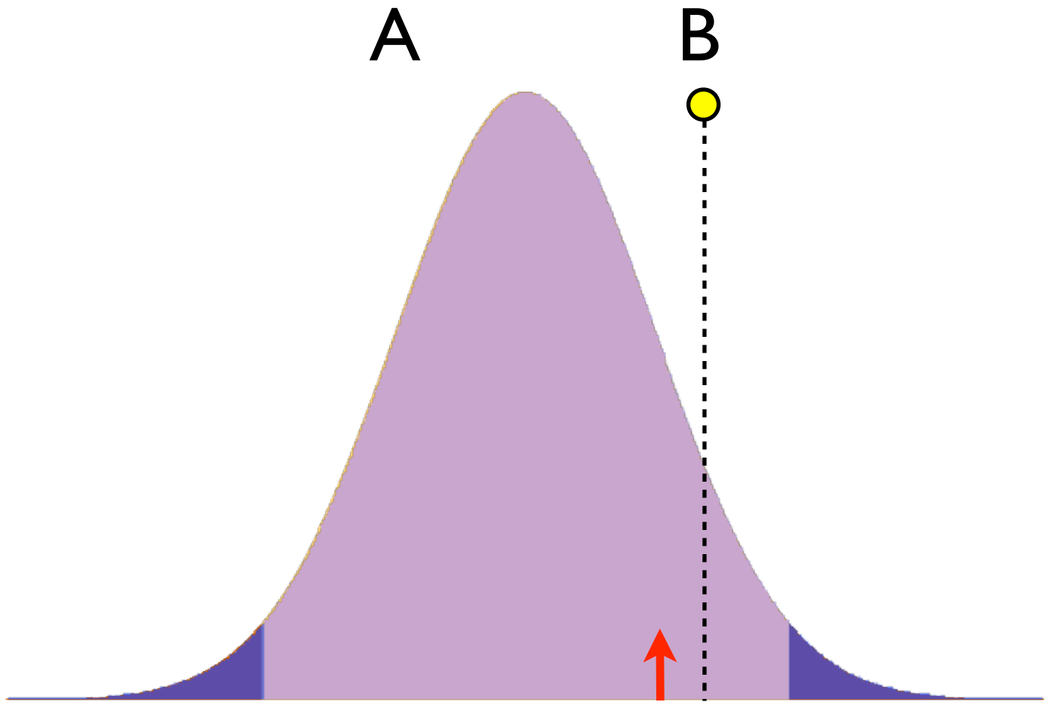
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2)}{2\sigma^2}}$$

The Null Hypothesis



$$H_0 : \quad \overline{x} = \mu$$

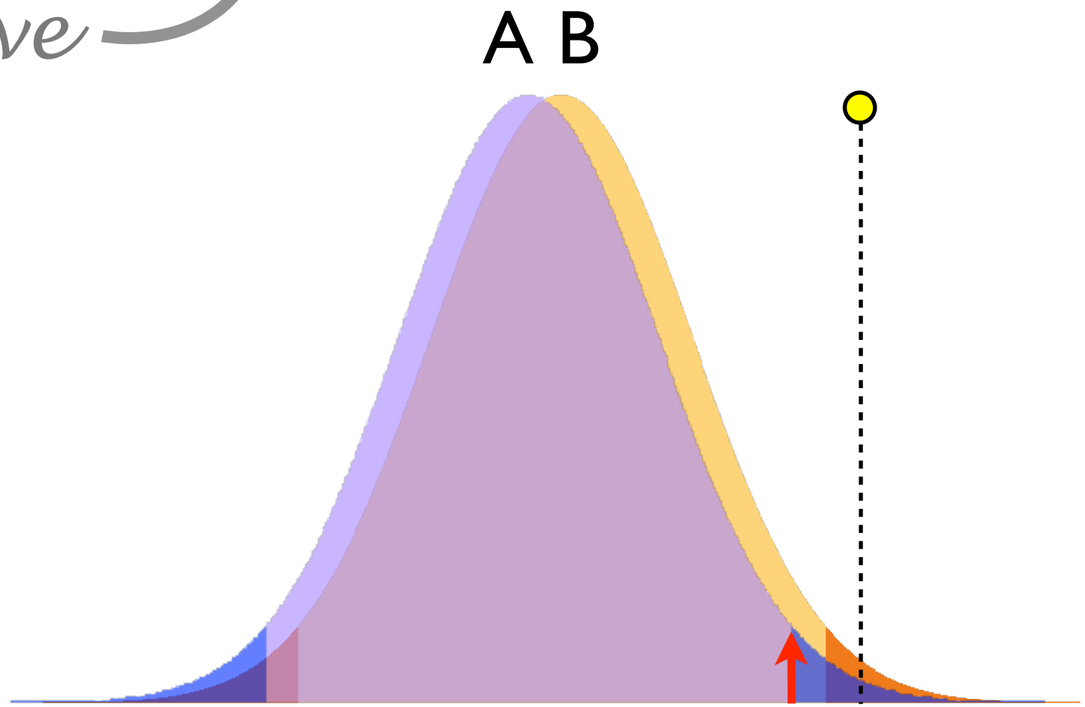
Comparing distributions



$$H_0: \quad \mu_A = \mu_B$$

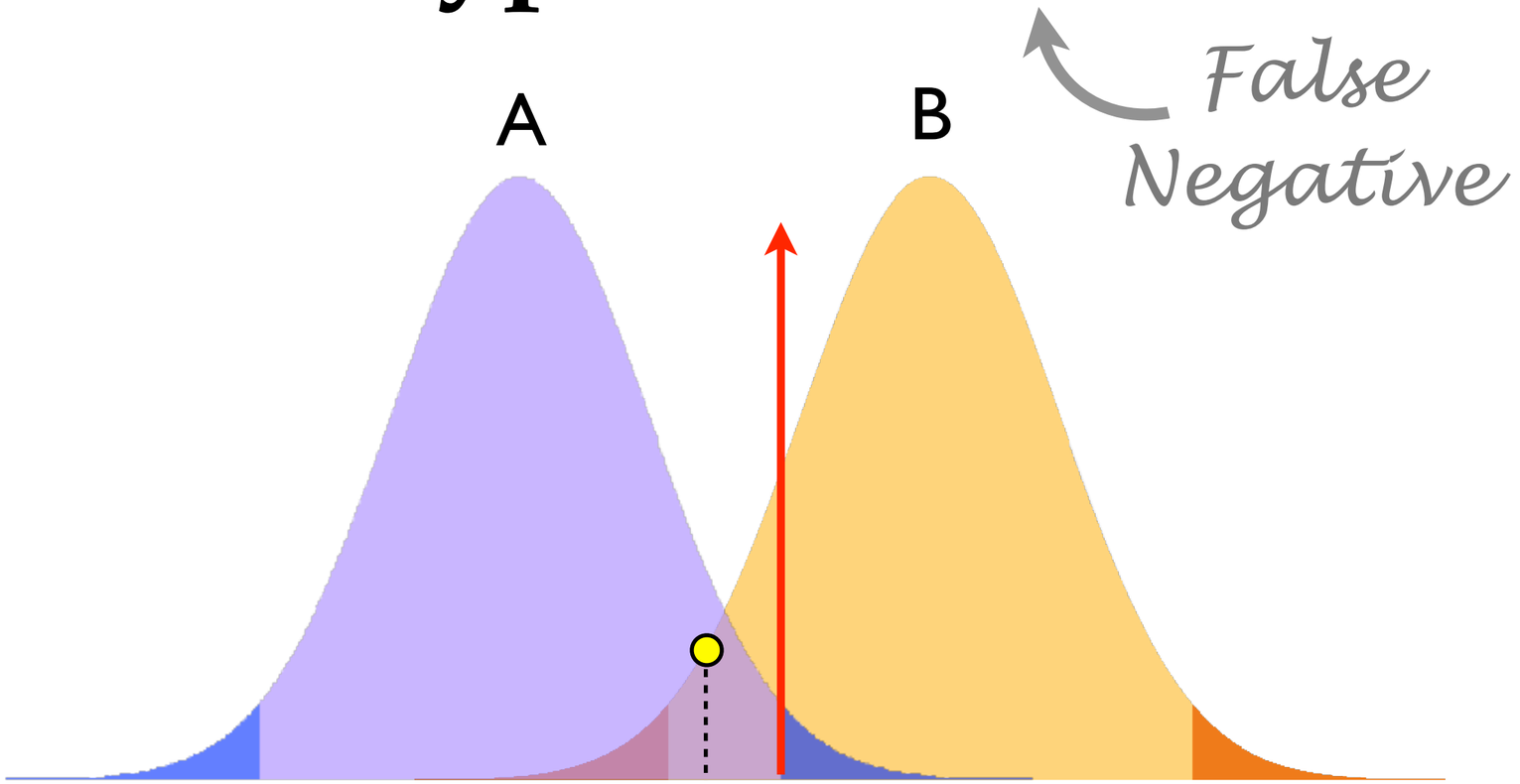
Type-I Error

False Positive



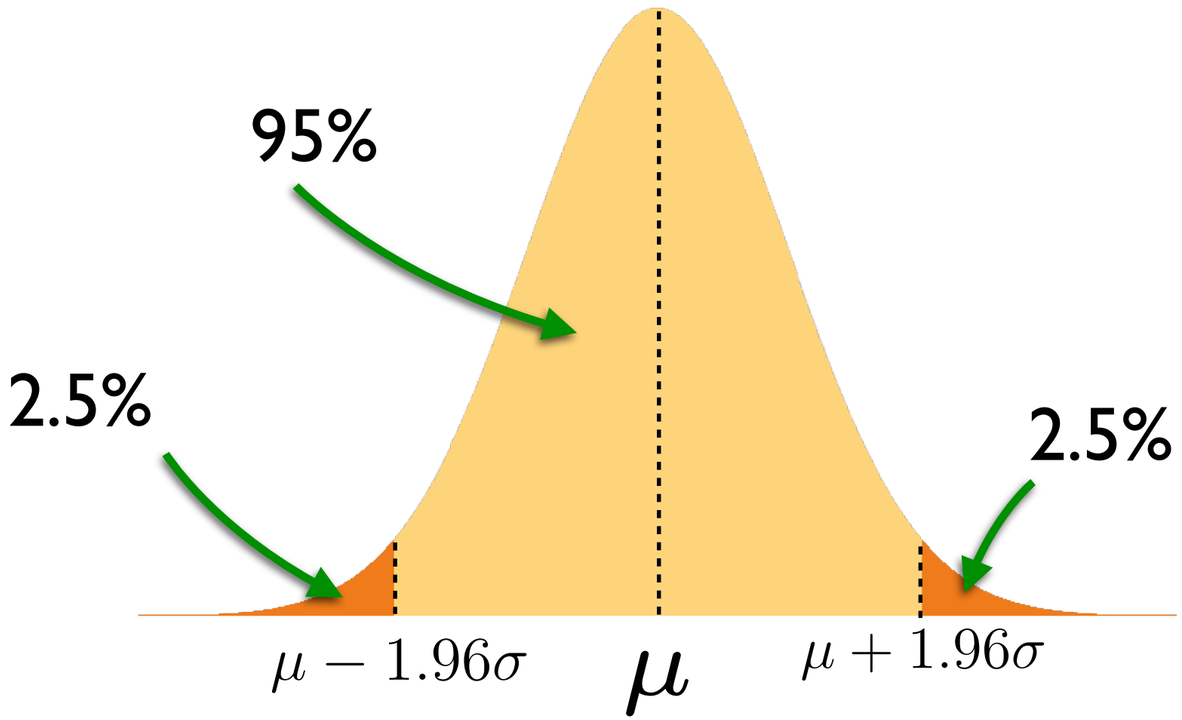
$$\mu_A = \mu_B$$

Type-II Error



$$\mu_A \neq \mu_B$$

p-Value



The probability of observing
as extreme a result
assuming the null hypothesis
is true

p-value = area outside these critical points = 0.05 here

Power

Number of participants per group

Standard deviation

$$n = \frac{(z_\alpha + z_\beta)^2 \, \sigma^2}{\delta^2}$$

Required change

$$n \approx \frac{16\sigma^2}{\delta^2}$$

$$n \approx \frac{16\sigma^2}{\delta^2}$$

Control conversion rate = 5%
Desired increase = **50%** (i.e. to 7.5%)
Standard deviation = 0.1

n = 256 (25 mins for 50k DAU game)*

(* assume 40% retention rate, even daily session distribution and 2 buckets in test)

$$n \approx \frac{16\sigma^2}{\delta^2}$$

Control conversion rate = 5%
Desired increase = **5%** (i.e. to 5.25%)
Standard deviation = 0.1

n = 25,600 (20 hrs for 50k DAU game)

$$n \approx \frac{16\sigma^2}{\delta^2}$$

Control conversion rate = 5%
Desired increase = 5% (i.e. to 5.25%)
Standard deviation = **0.5**

n = 640,000 (21 days for 50k DAU game)

Challenges when Testing

Primacy

- News users behave differently to old users
- Familiarity with existing UI / resources / items etc.

SOLUTION: Restrict tests to new users

Causality

- There may be many reasons for a change in test statistic
- Seasonality, events, trends, errors, etc.

SOLUTION: use tight evaluation criteria

(e.g. sales of item tested NOT overall revenue)

Testing QA

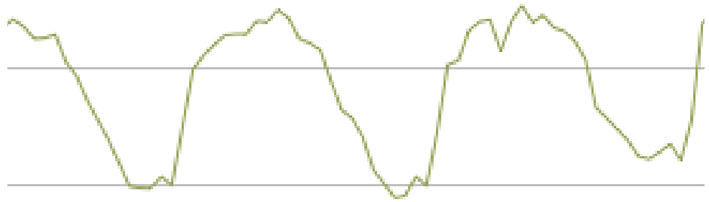
- Tests can (will) introduce errors
- Particularly with many variants

SOLUTION(s)

- ramp-up, roll-back capability
- force user bucket capability

Temporal Effects

- Daily, weekly, yearly
- False signals
- Ramp up bias



SOLUTION

- Run tests for sufficiently long to normalize for effects

Version Control

- Multiple app-versions in flight
- Resources may have changing schema
- Can't force upgrade always

SOLUTION(s)

- Limit to one app-version; careful version control with schema

Testing in Online Games

●●●

Death to HiPPOs

Steve@swrve.com
@stevec64

- Homework:
  - http://exp-platform.com - Ron Kovahi et al.
  - http://statisticsforexperimenters.net/ - George Box et al.
  - http://www.kaushik.net - Occam's Razor Blog
  - http://www.abtests.com/

"Planet Cute" art by Daniel Cook (Lostgarden.com)
Gun art content by Ocean's Dream