A/B Testing for Game Design Iteration: A Bayesian Approach

Steve Collins
CTO / Swrve

steve@swrve.com
@stevec64

Biography

...

havok

kore

swrve

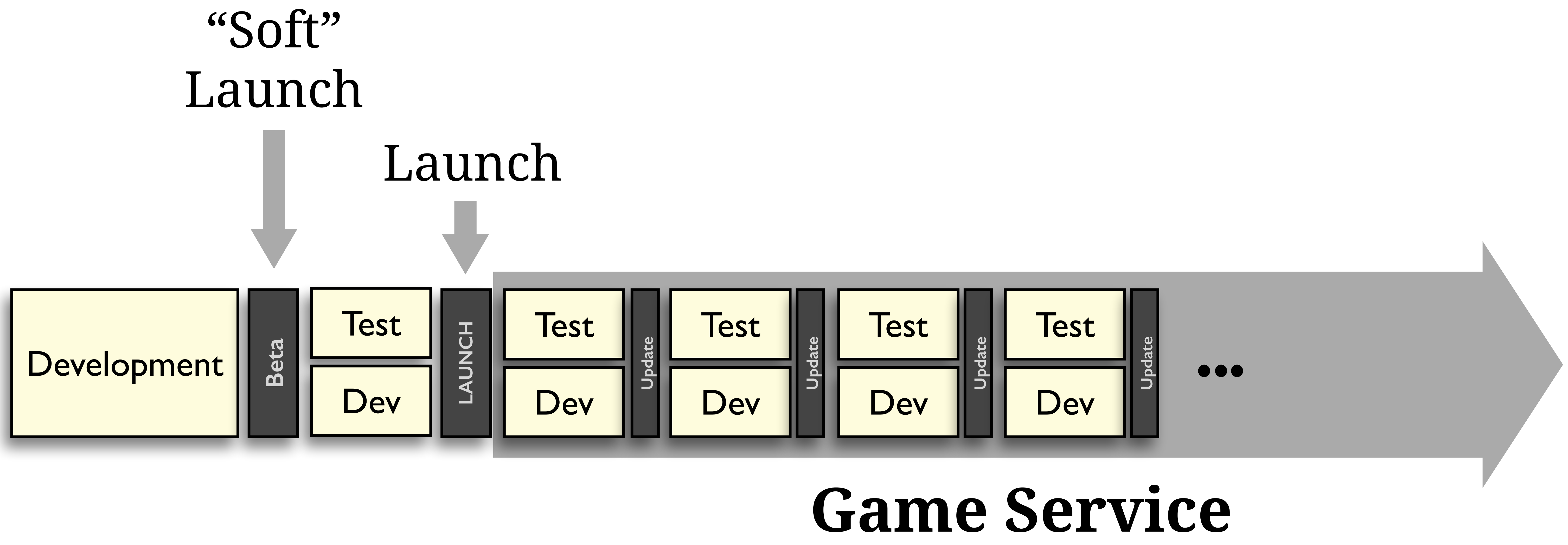
Targeting & Analytics

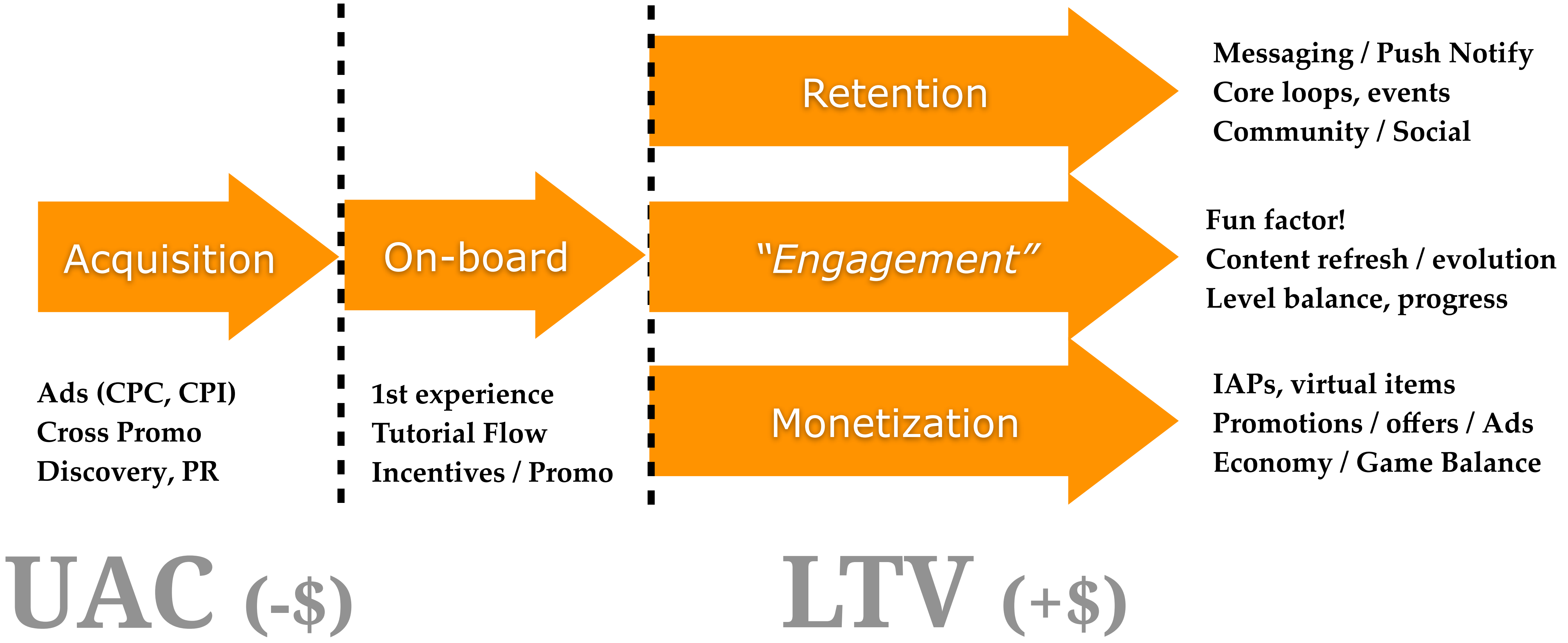
A/B Testing

Push Notifications

In-App Messaging

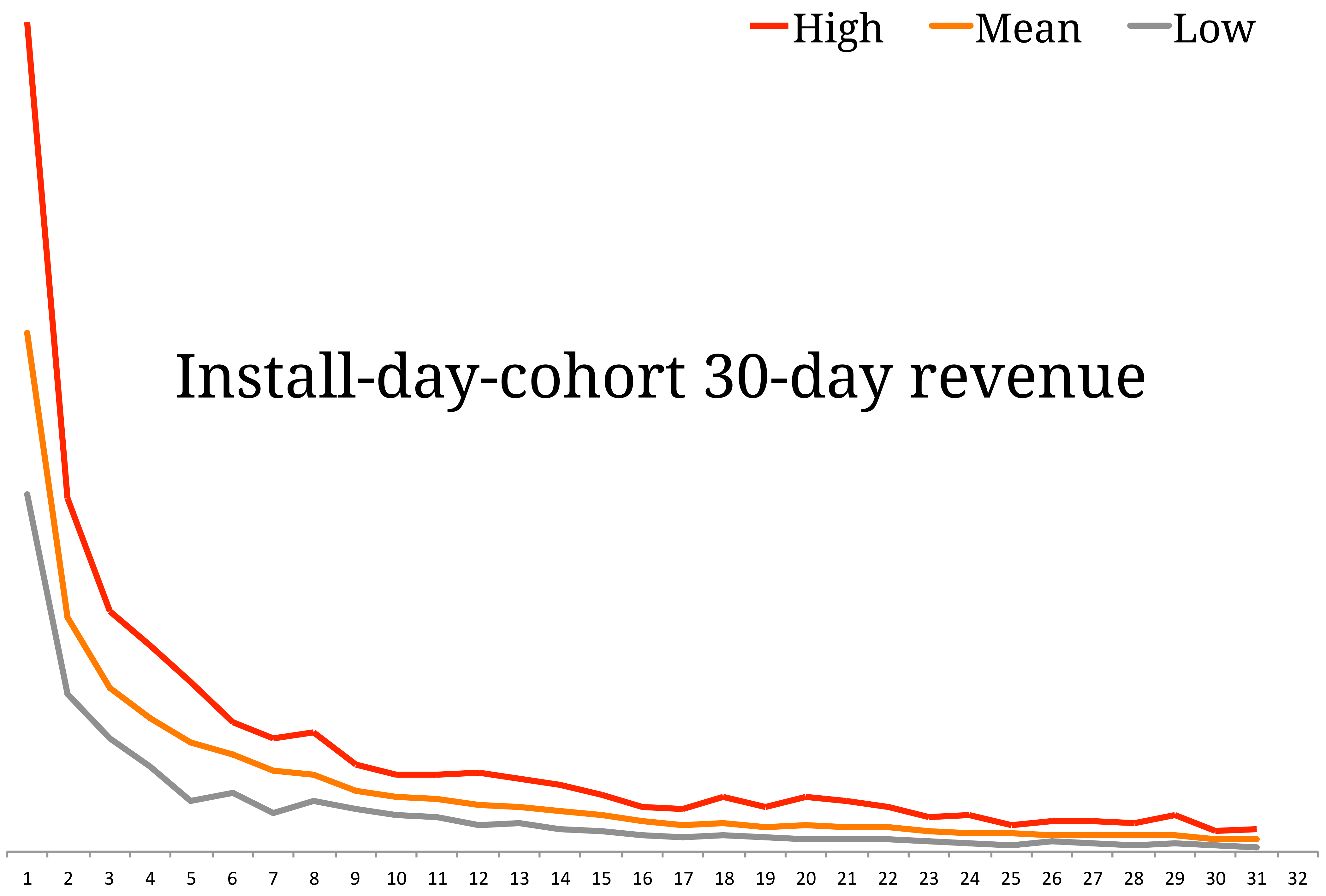
Introduction to A/B Testing


"Soft" Launch
Launch
Development
Beta
Test
Dev
LAUNCH
Test
Dev
Update
Test
Dev
Update
Test
Dev
Update
Test
Dev
Update
...
Game Service


Acquisition
On-board
Retention
"Engagement"
Monetization
Messaging / Push Notify
Core loops, events
Community / Social
Fun factor!
Content refresh / evolution
Level balance, progress
IAPs, virtual items
Promotions / offers / Ads
Economy / Game Balance
Ads (CPC, CPI)
Cross Promo
Discovery, PR
1st experience
Tutorial Flow
Incentives / Promo
UAC (-$)
LTV (+$)

$$ROI = LTV - UAC$$

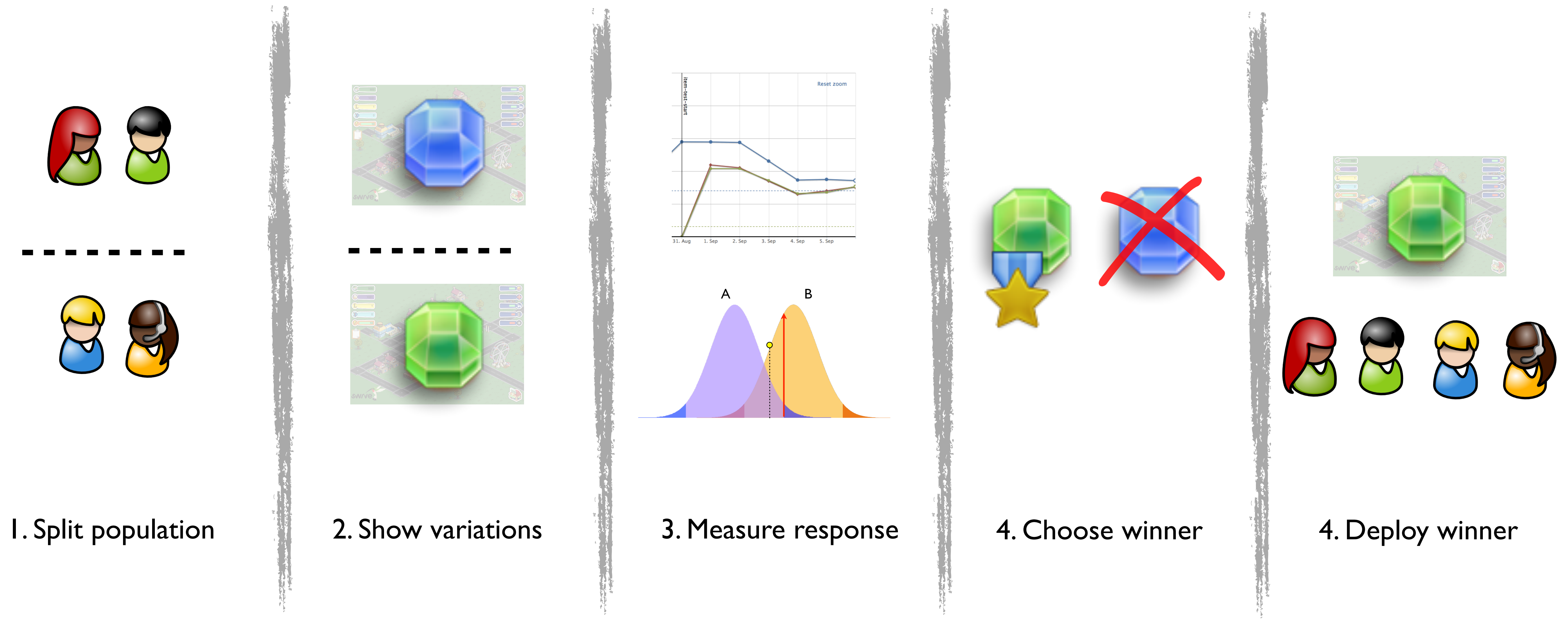
$$\mathbf{ROI} = \underbrace{\sum_{\mathbf{d}=1}^{\mathbf{lifetime?}} \mathbf{ARPU}_{\mathbf{d}}}_{\mathbf{LTV}} - \mathbf{UAC}$$

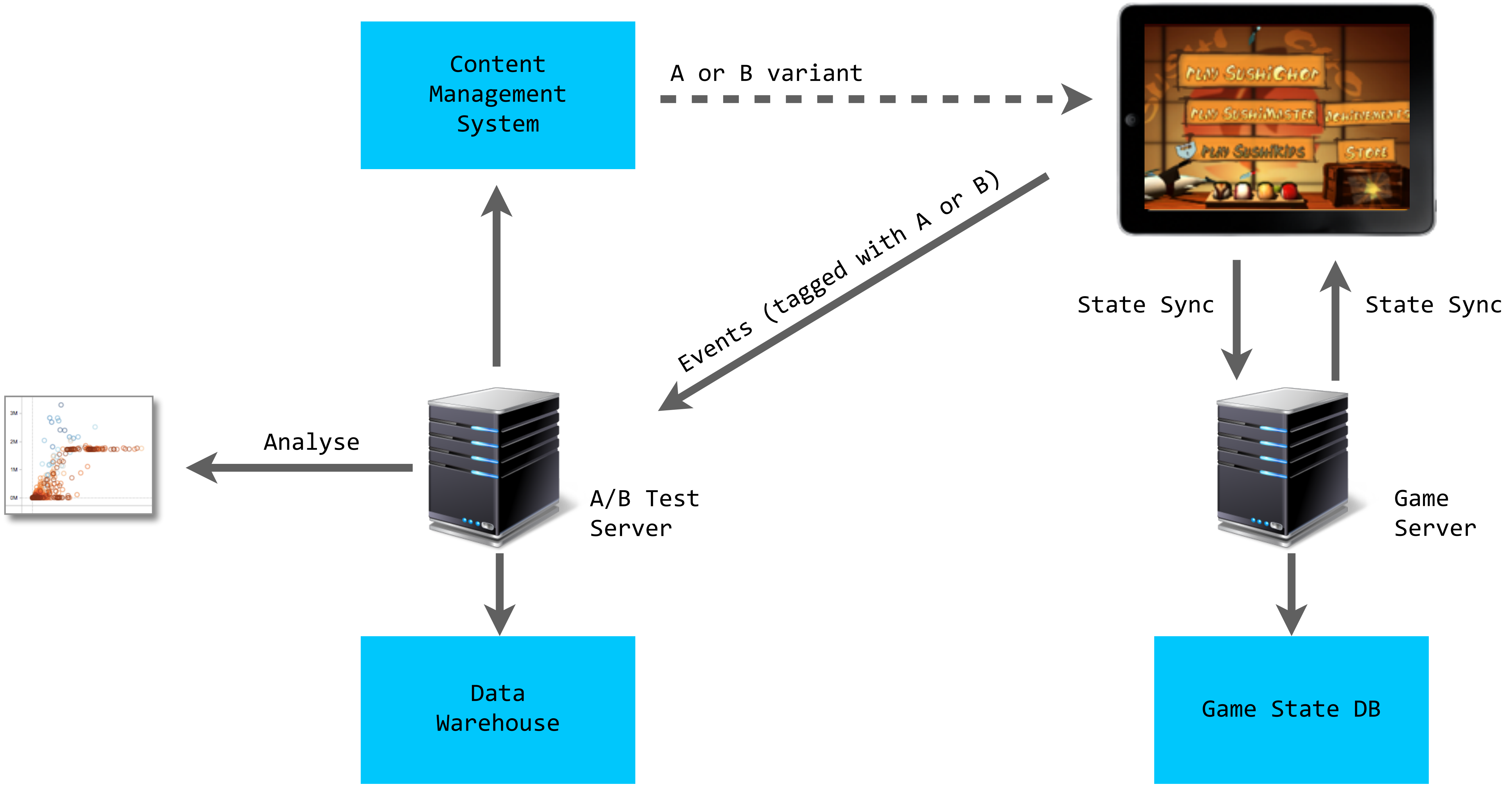

High
Mean
Low
Install-day-cohort 30-day revenue
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Understand
Metrics -> Analytics -> Insight

Test Hypotheses
Data driven

Take action
Iterate, fail-fast


A
B
1. Split population
2. Show variations
3. Measure response
4. Choose winner
4. Deploy winner

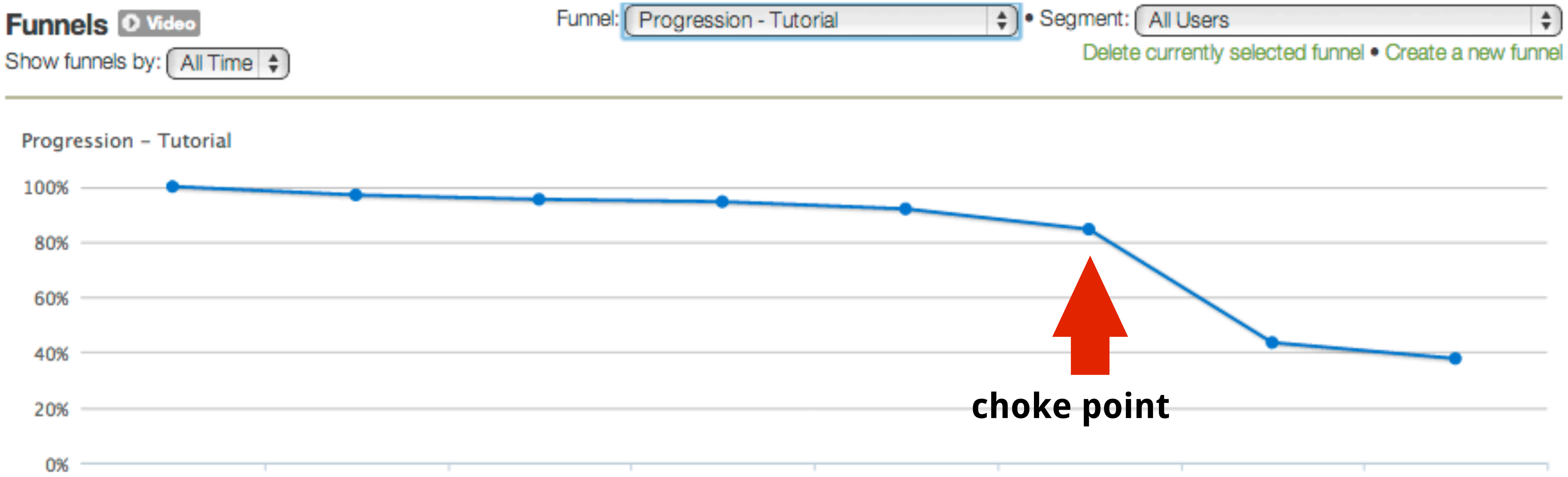

Content Management System
A or B variant
Events (tagged with A or B)
State Sync
State Sync
Analyse
A/B Test Server
Game Server
Data Warehouse
Game State DB

What to test?

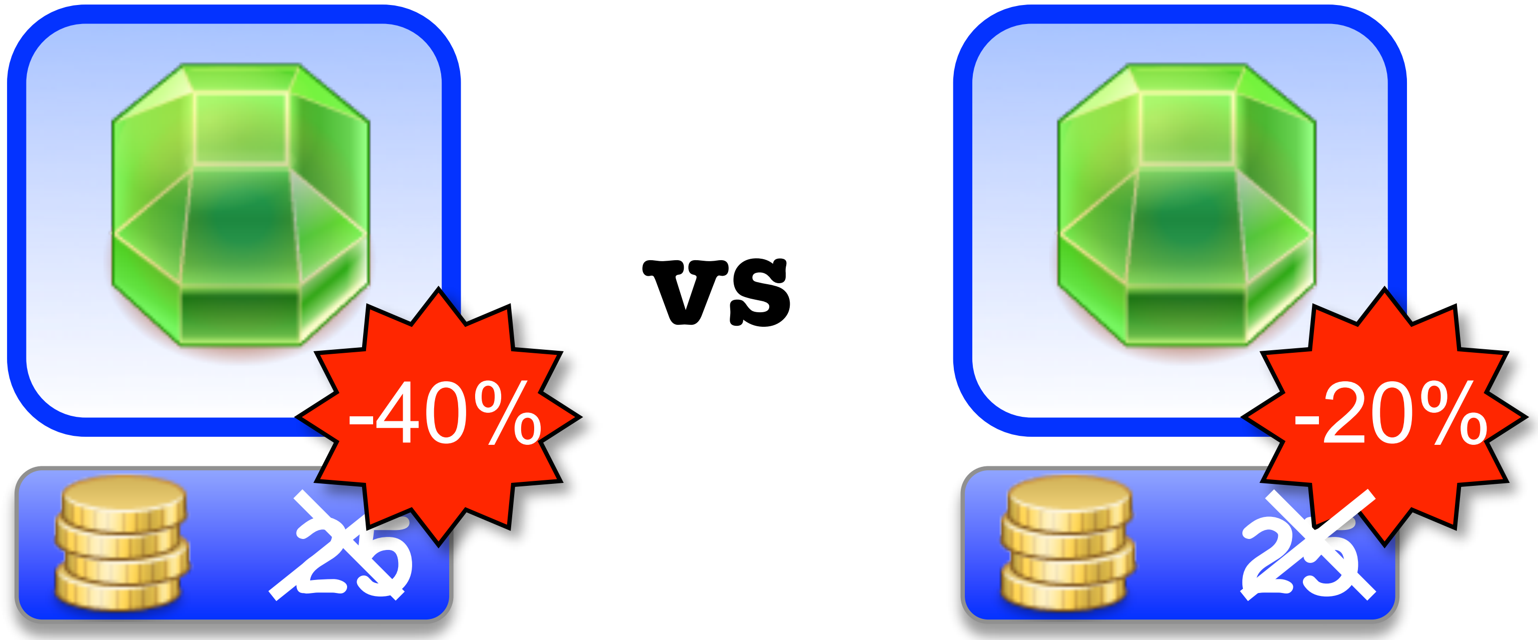
Message layouts / content

vs

Tutorial Flow



Promotion Discounts



Elasticity testing: exchange rate

vs

A

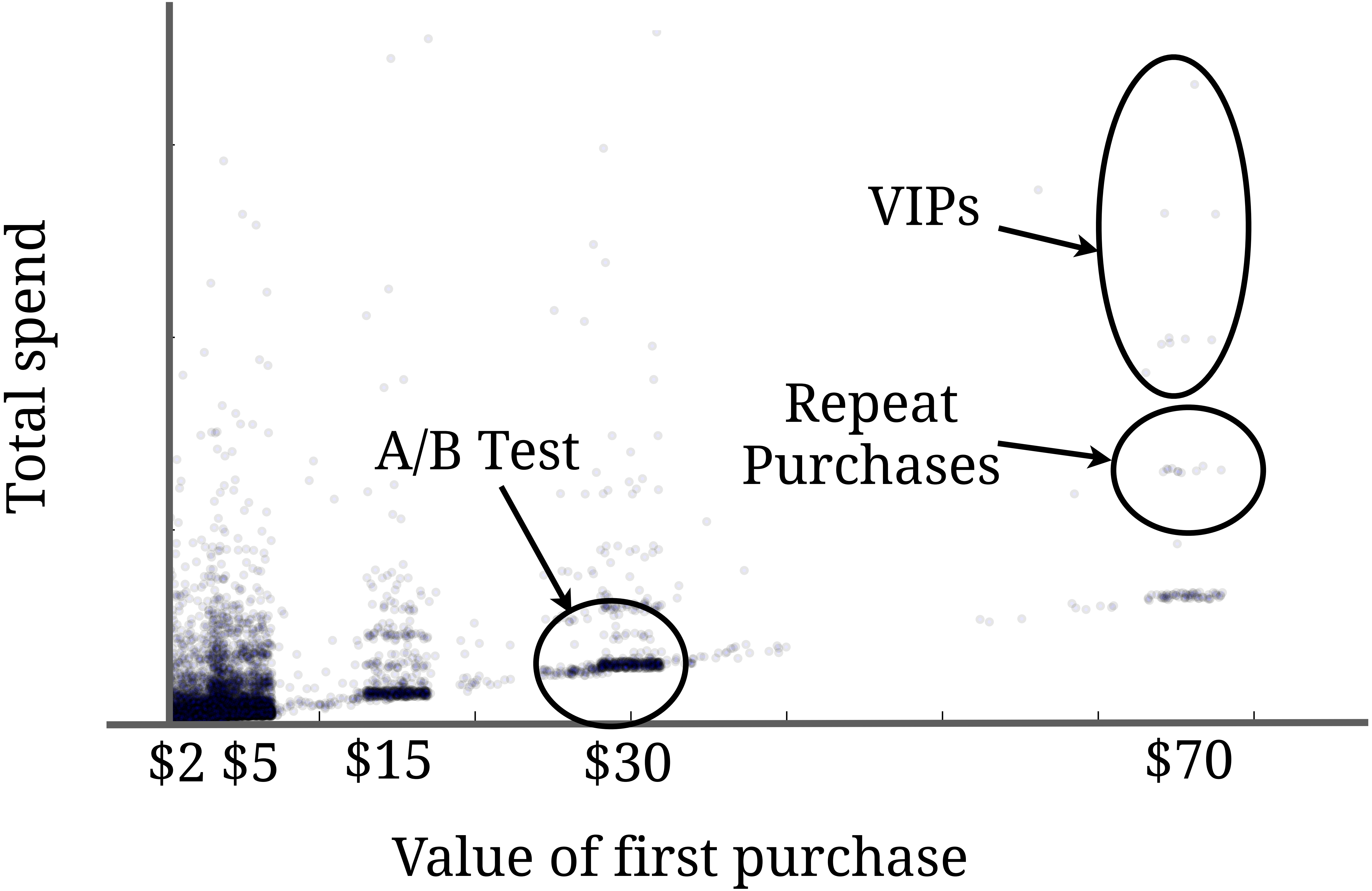
B

Store Inventory

Price set A

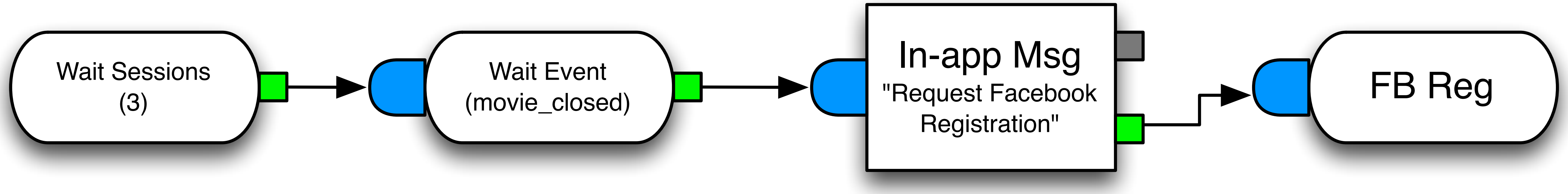
Price set B

Price set C



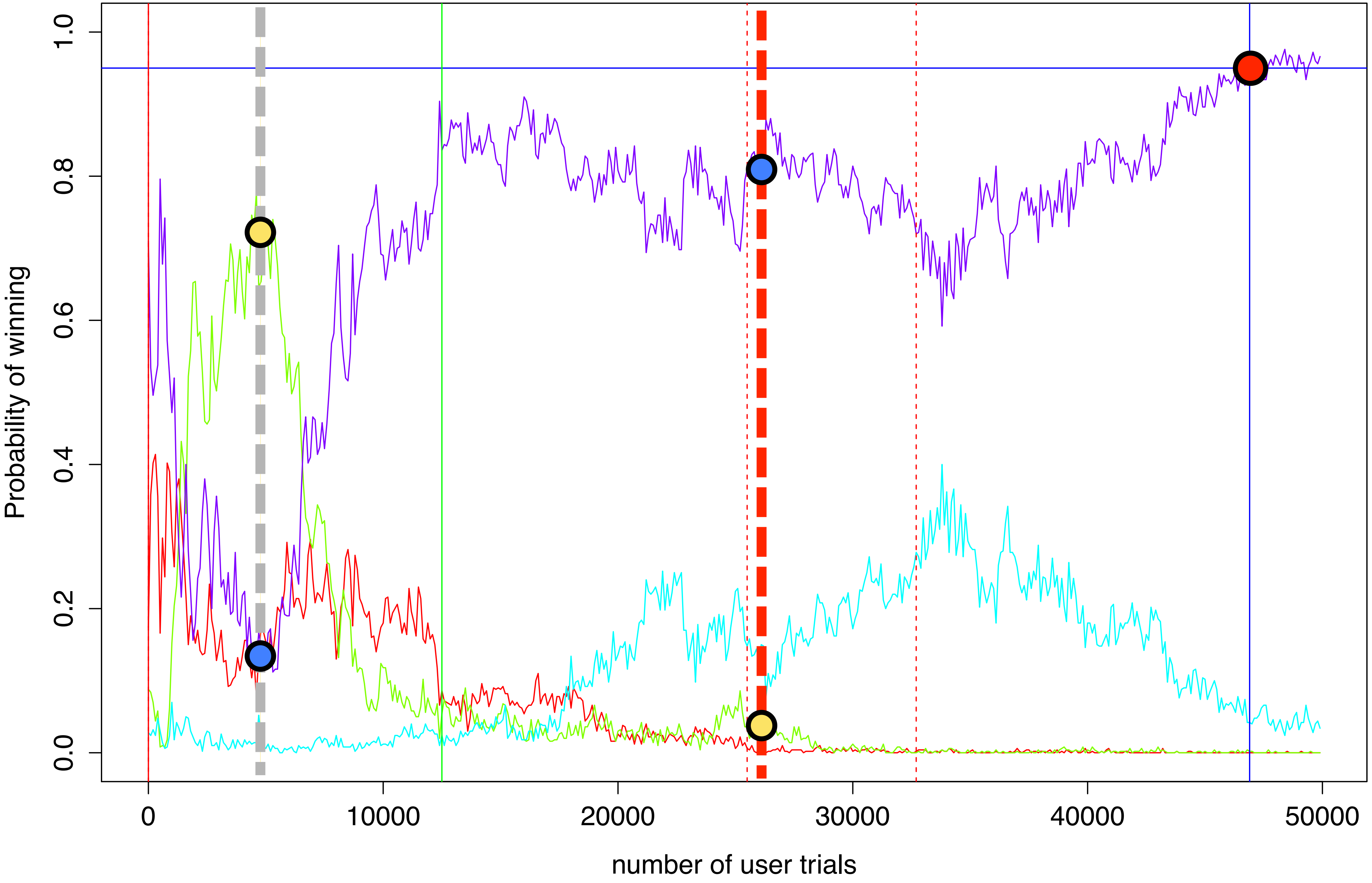
Total spend
VIPs
Repeat
Purchases
A/B Test
$2 $5
$15
$30
$70
Value of first purchase

Timing

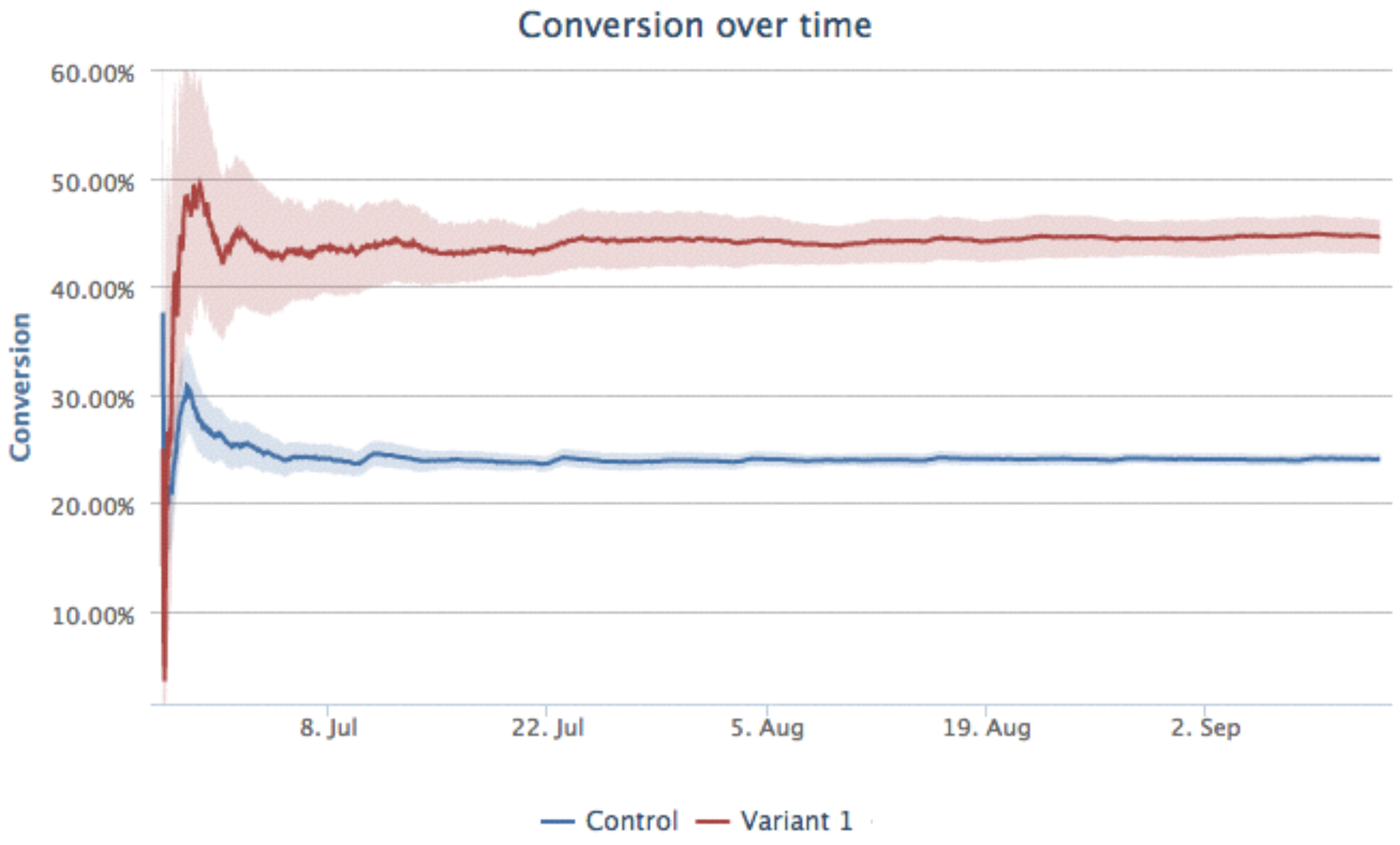


when

where



Probability of winning
0.0
0.2
0.4
0.6
0.8
1.0
0
10000
20000
30000
40000
50000
number of user trials



Conversion over time
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
Conversion
8. Jul
22. Jul
5. Aug
19. Aug
2. Sep
Control
Variant 1

Canacandycrushbalt*

Day 1 retention = 30%

* Apologies to King and Adam Saltsman

Beta Test

Original

New Version

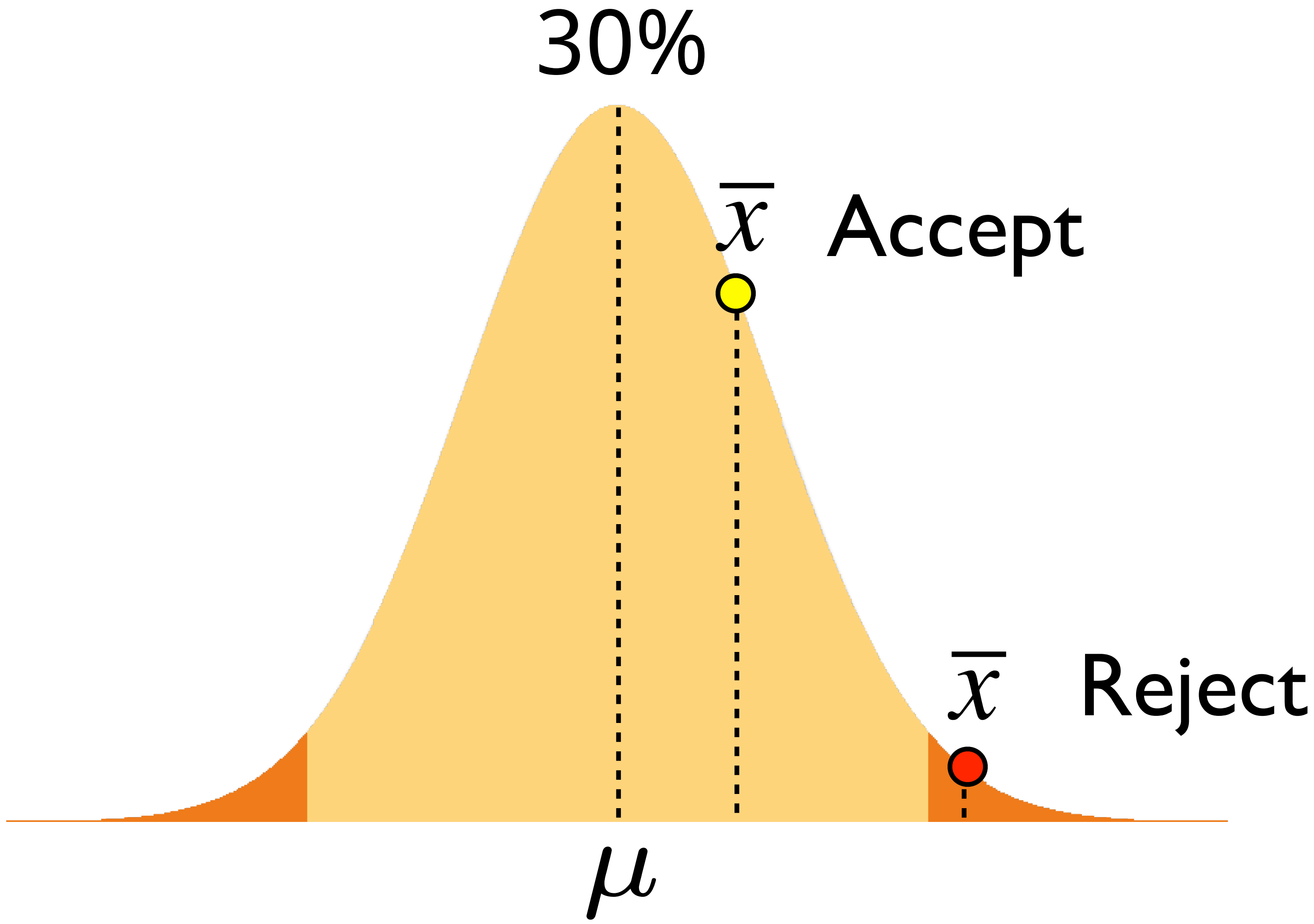
Expecting 30% day-1 retention

After 50 users, we see 0%

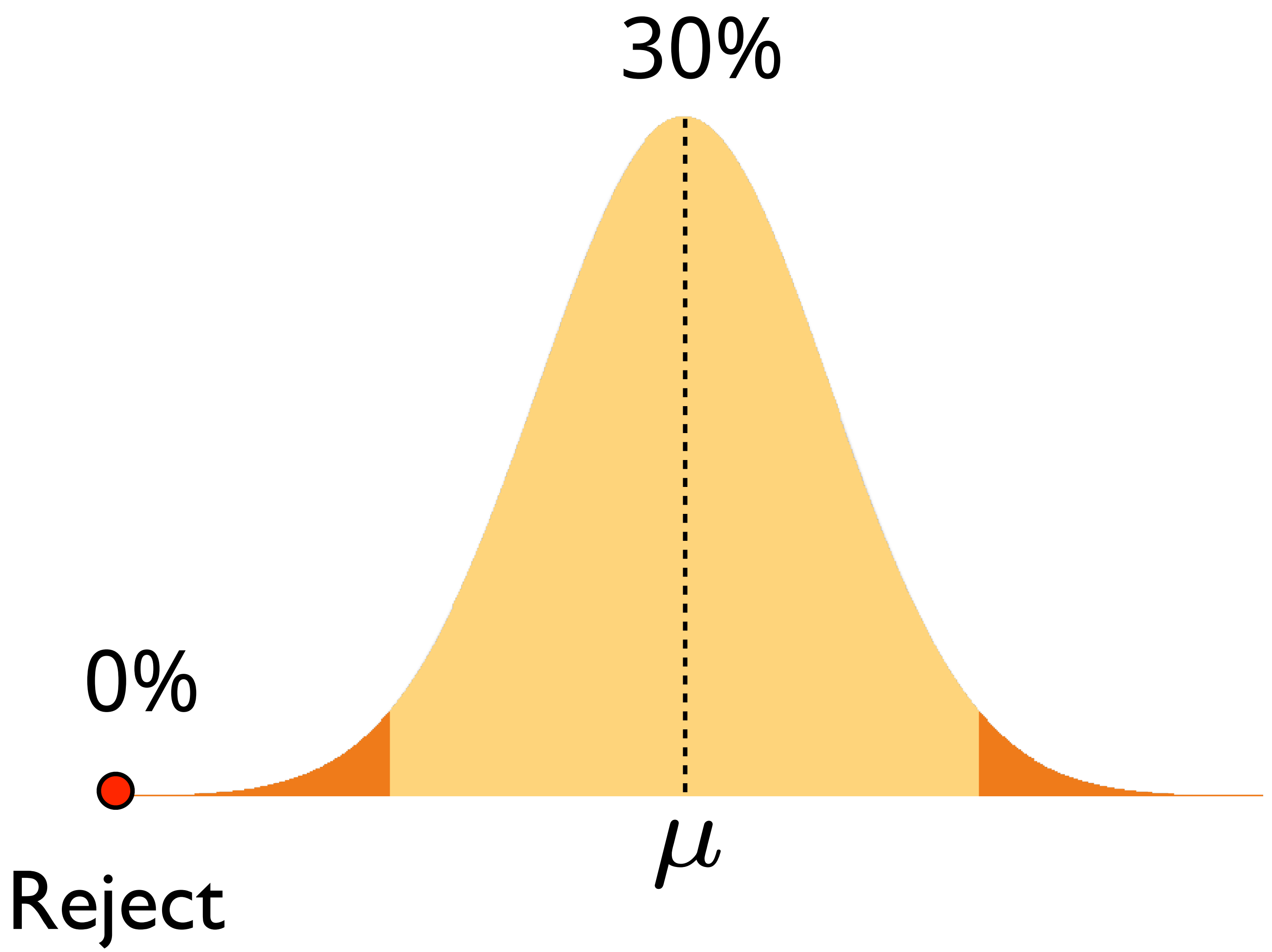
Is this bad?

Null Hypothesis Testing (NHT) View

The Null Hypothesis



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



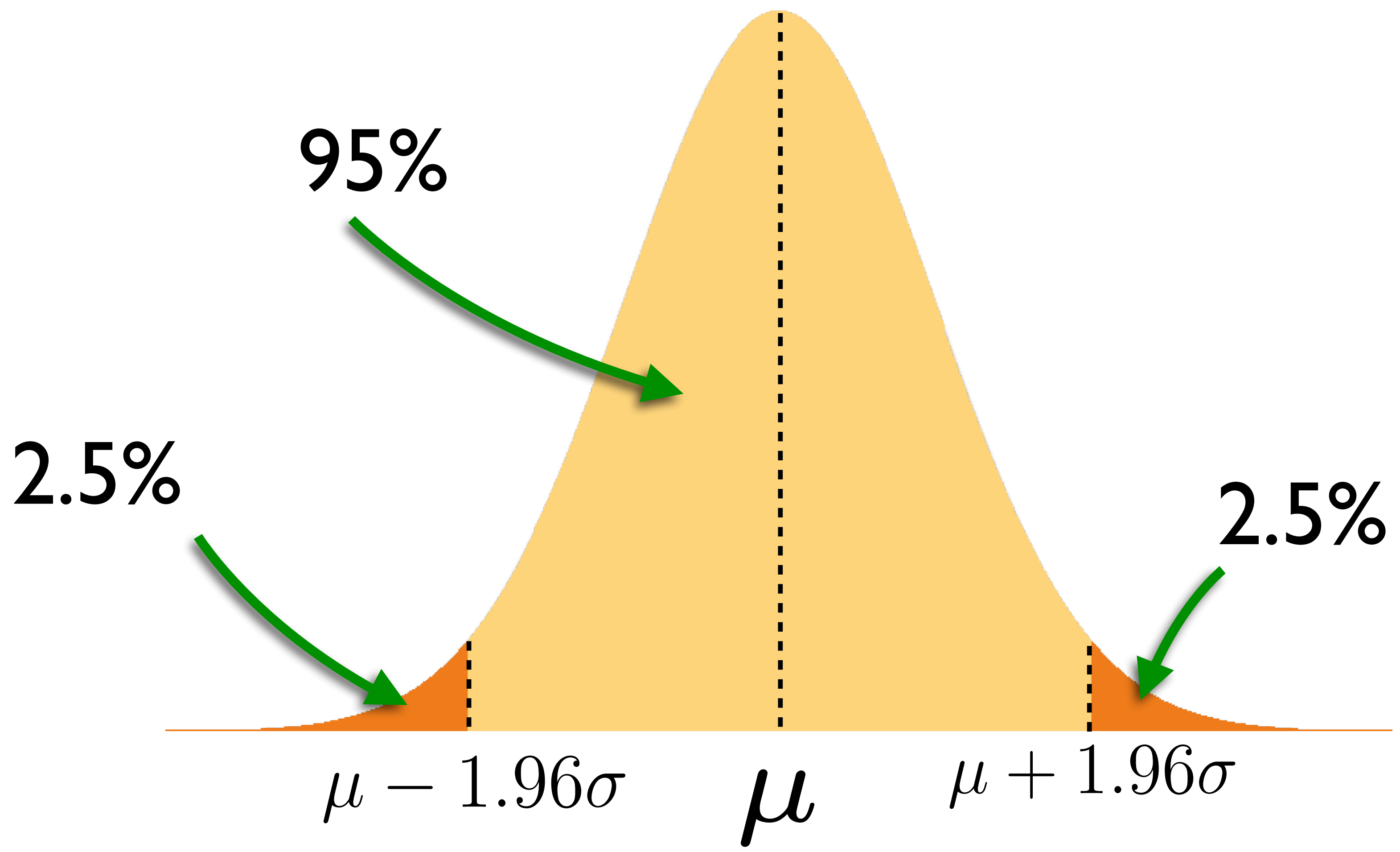
30%
0%
μ
Reject

Conclusion: 30% is <u>unlikely to be</u> the retention rate

Issue #1

p-value

p-Value



p-Value

The probability of observing <u>*as extreme a result*</u>
assuming the null hypothesis is true

OR

The probability of the data given the model

Null Hypothesis: H_0

		Truth	
		H	H
Observation	Accept H	✓	**Type-II** Error
	Reject H	**Type-I** Error	✓

False Positive

p-Value

$p < 0.05$

All we can ever say is either

- *not enough evidence that retention rates are the same*
- *the retention rates are different, 95% of the time*

actually…

$p < 0.05$

The evidence supports a rejection of the null hypothesis, i.e. the probability of seeing a result as extreme as this, assuming the retention rate is actually 30%, is less than 5%.

Issue #2

“Peeking”

Number of participants per group

Standard deviation

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

Required change

To get 5% false positive rate you need...

Peeks	5% Equivalent
1	2.9%
2	2.2%
3	1.8%
5	1.4%
10	1%

i.e. 5 times

http://www.evanmiller.org/how-not-to-run-an-ab-test.html

Issue #3
Family-wise Error

Family-wise Error

$$p = P(\text{ Type-I } Error) = 0.05$$

$$P(\text{ no Type-I } Error) = 0.95$$

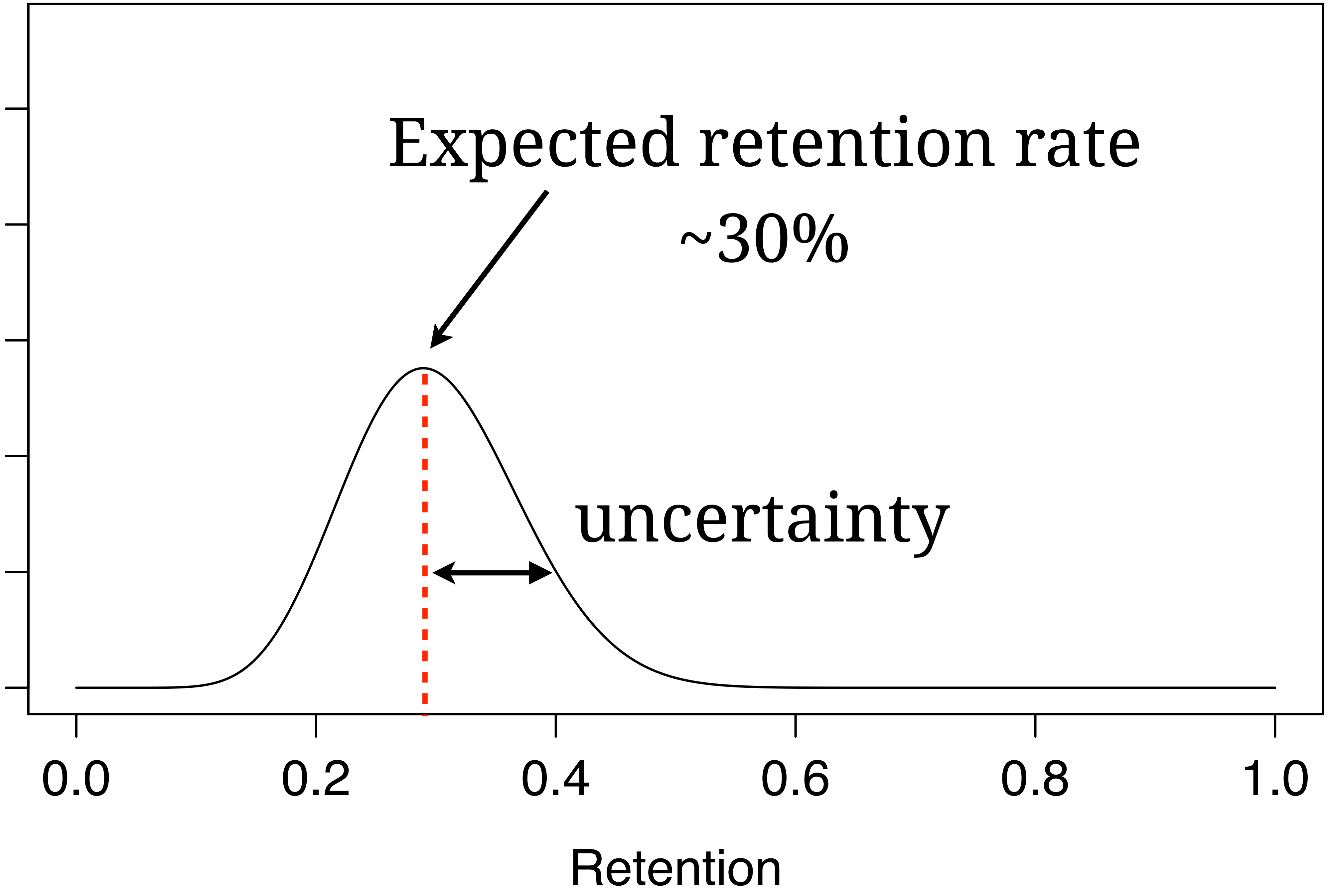
5% of the time we will get a false positive - for **one** treatment

$$P(\ no \text{ Type-I } Error\ for\ 2\ treatments) = (0.95)(0.95) = 0.9025$$

$$P(\ at\ least\ 1 \text{ Type-I } Error\ for\ 2\ treatments) = (1 - 0.9025) = 0.0975$$

Bayesian View

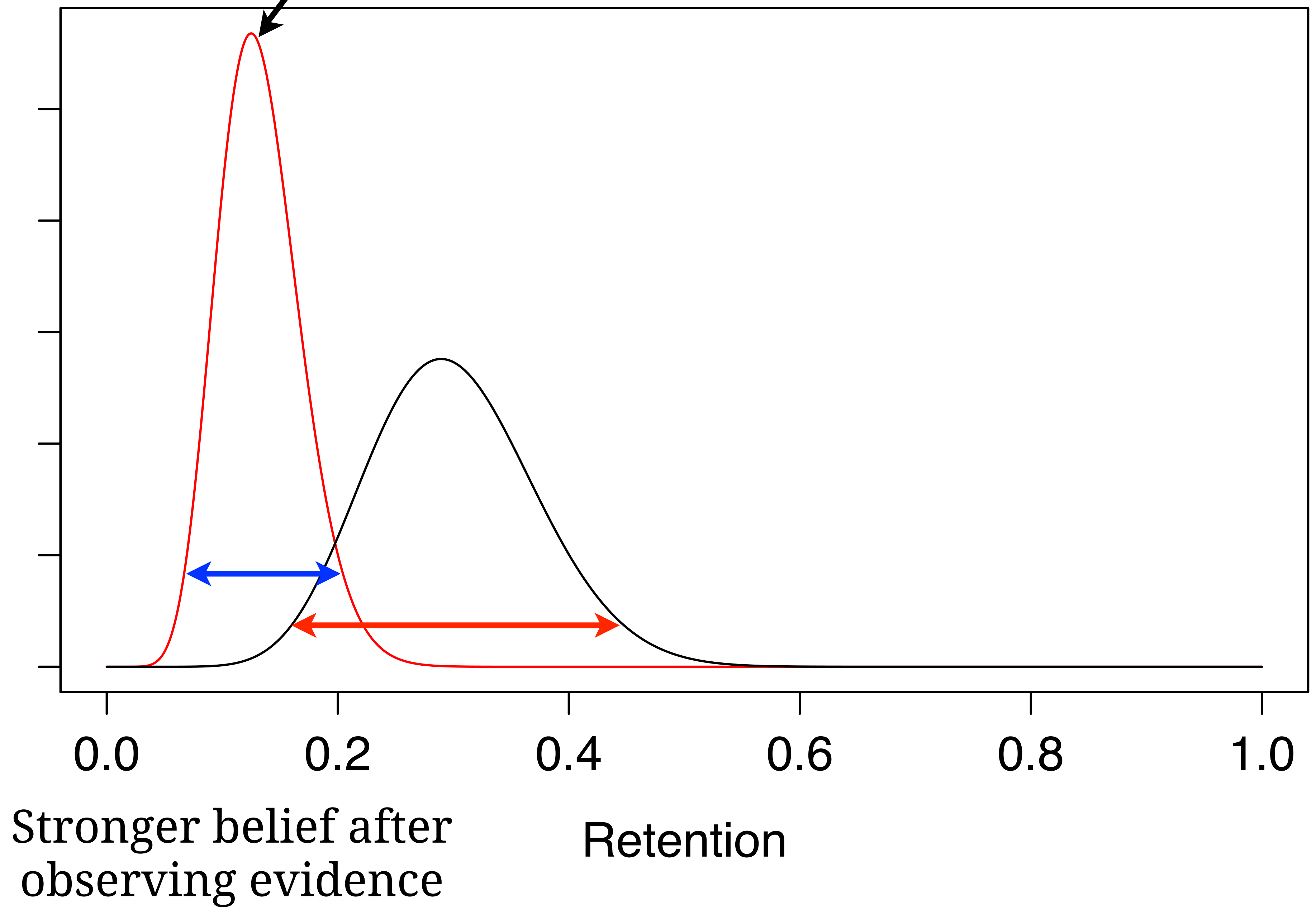
“Belief”



Expected retention rate
~30%
uncertainty
0.0
0.2
0.4
0.6
0.8
1.0
Retention

New “belief” after 0 retained users

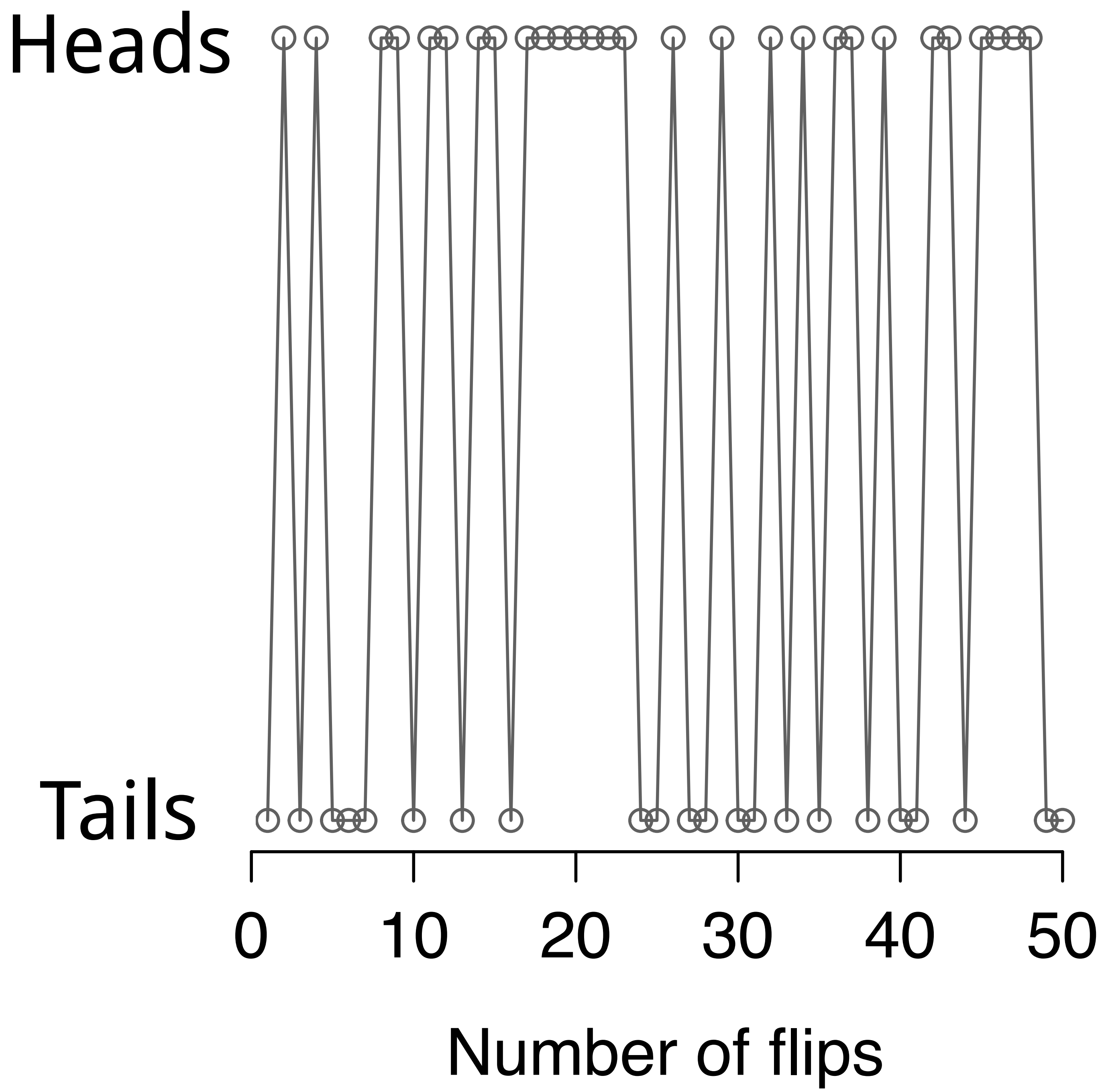
~15%



The probability of the model given the data

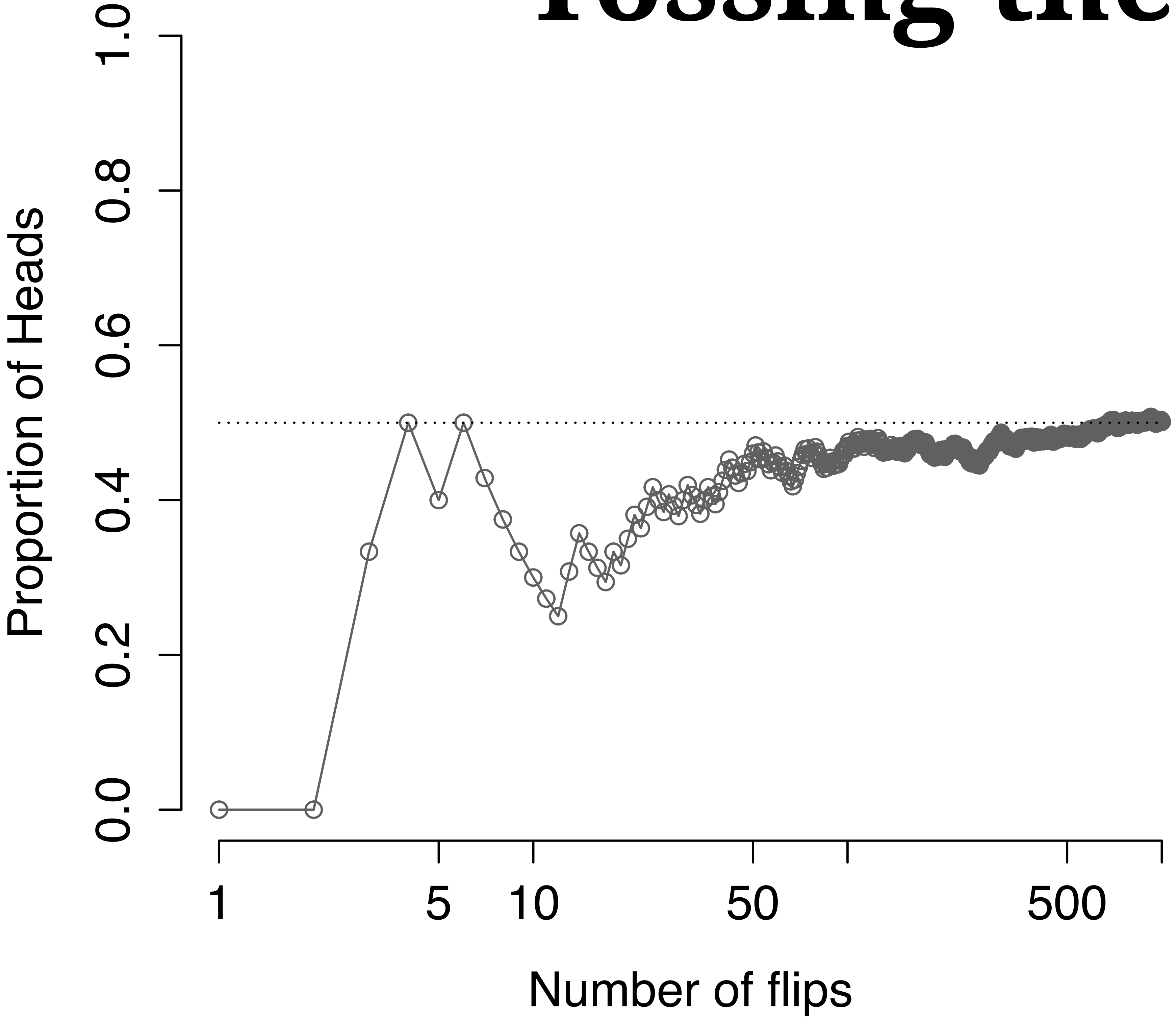
p(heads) = p(tails) = 0.5

Tossing the Coin



THTHTTTHHTHH…

Tossing the Coin



Long run average = 0.5

Terminology

$p(x)$ Probability of x

$p(x, y)$ Probability of x and y

(conjoint)

$p(x|y)$ Probability of x given y

(conditional)

The Bernoulli distribution

Head (H) = 1, Tails (T) = 0

A single toss: $$p(x|\theta) = \theta^x(1-\theta)^{(1-x)}$$

For a "fair" coin, θ = 0.5

$$p(heads = 1|0.5) = 0.5^1(1{-}0.5)^{(1-1)} = 0.5$$

$$p(tails = 0|0.5) = 0.5^0(1{-}0.5)^{(1-0)} = 0.5$$

The Binomial

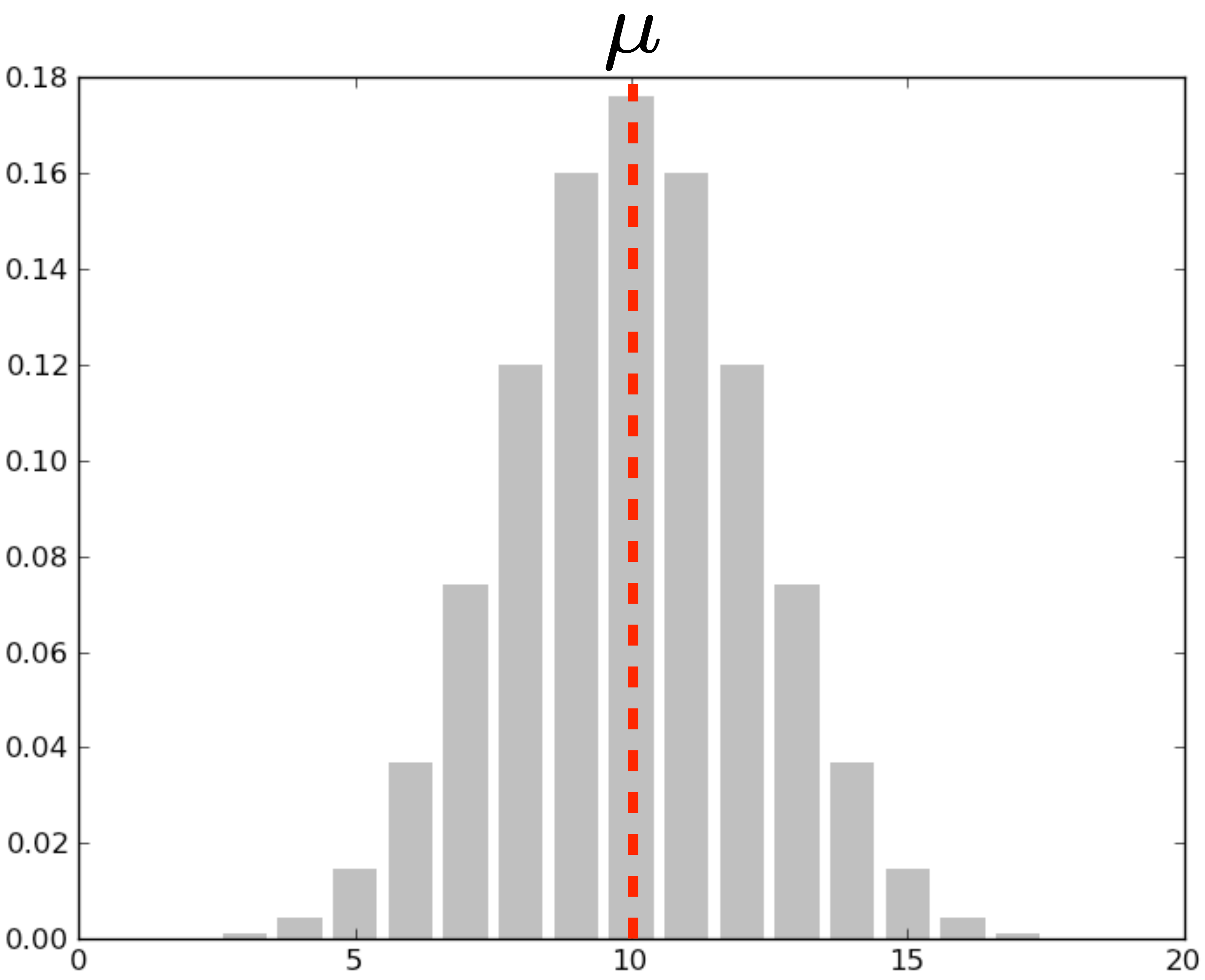
Probability of heads in *a single* throw:

$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Probability of x heads in n throws:

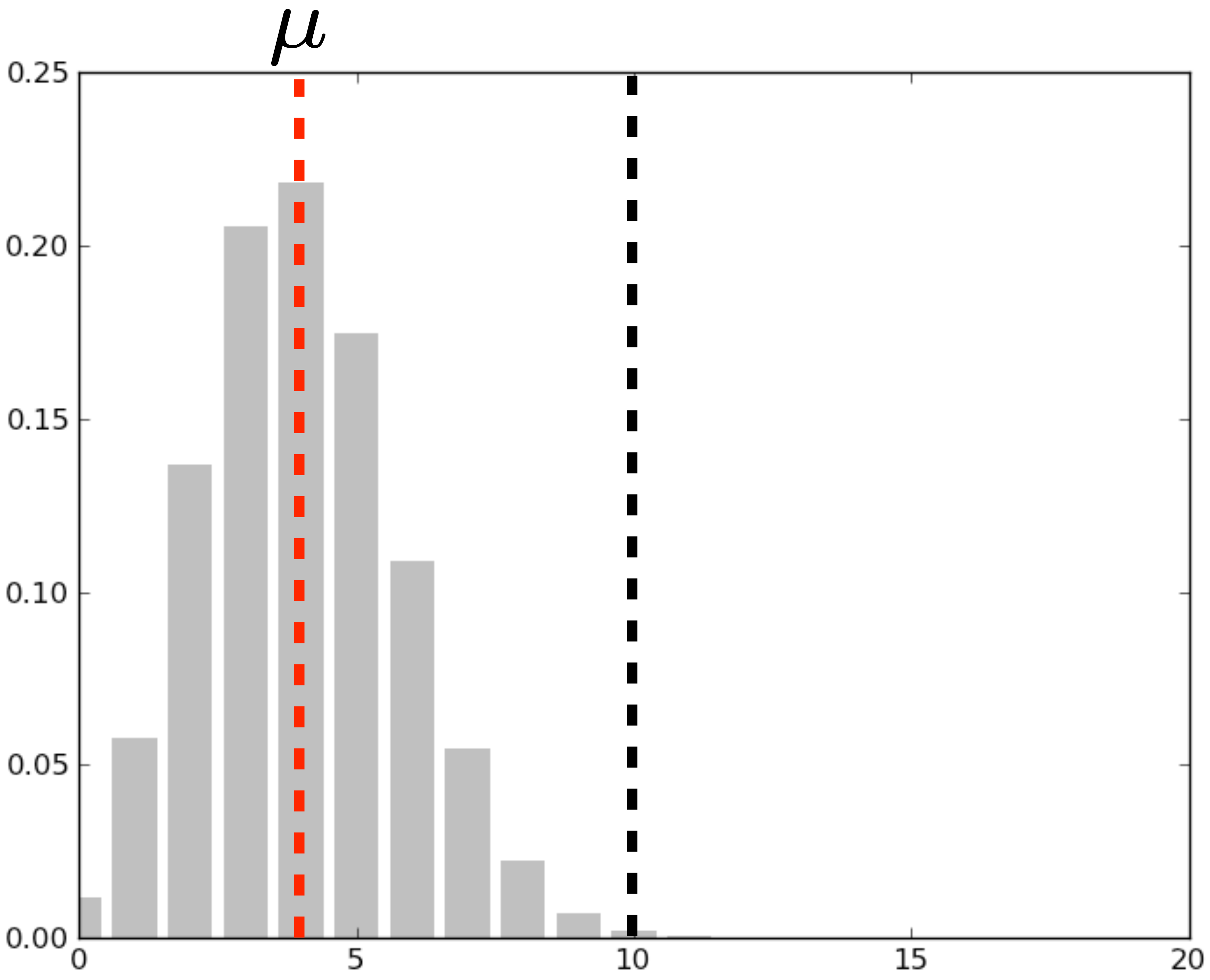
$$p(x|\theta, n) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{(n-x)}$$

$$p(x|0.5, 20)$$



20 tosses of a fair coin

$$p(x|0.2, 20)$$



20 tosses of an "un-fair" coin

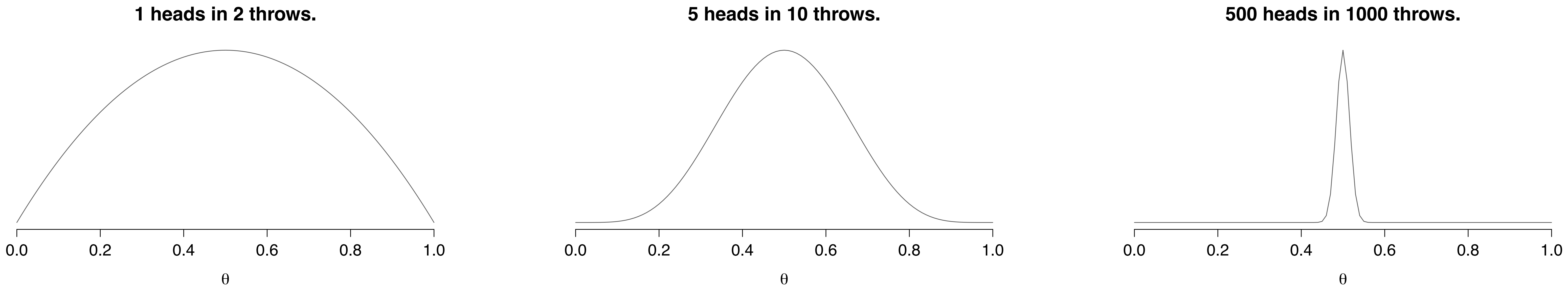
$$p(x|\theta, n) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{(n-x)}$$

Likelihood of θ given observation i of x heads in n throws:

$$L(\theta|x_i, n_i) = \left(\begin{array}{c} n_i \\ x_i \end{array} \right) \theta^{x_i} (1-\theta)^{(n_i - x_i)}$$

“Binomial Likelihood”

The Likelihood



Increasing likelihood of θ with more observations…

A recap…

x	The observations (#heads)
θ	The model parameter (e.g. fair coin)
$p(x\|\theta)$	Probability of data given model
$p(\theta\|x)$	We want to know this

Note that $p(x|\theta) \neq p(\theta|x)$

$$p(cloudy|raining) \neq p(raining|cloudy)$$

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \qquad p(x) = \sum_y p(x|y)p(y)$$

Bayes' Rule

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)}$$

discrete form

$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

continuous form

prob #heads given model

$p(x|\theta)$

prob model given #heads

$p(\theta|x)$

prob of model

$p(\theta)$

$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

$p(x)$

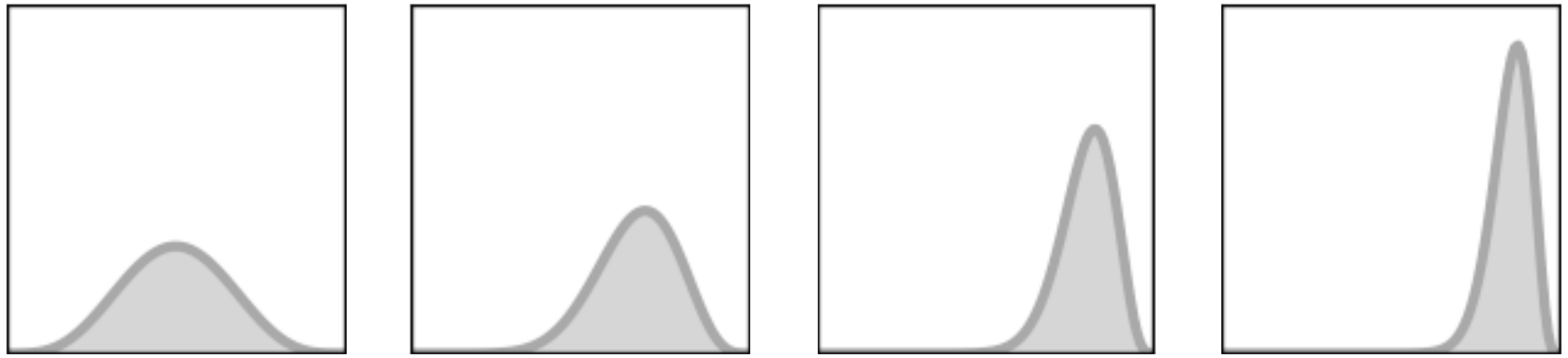
$$\underbrace{p(\theta|x)}_{\text{posterior}} = \underbrace{p(x|\theta)}_{\text{likelihood}} \; \underbrace{p(\theta)}_{\text{prior}} \; / \; \underbrace{p(x)}_{\substack{\text{normalizing}\\ \text{factor}}}$$

$$\text{normalizing factor} \quad p(x) = \int p(x|\theta)p(\theta)d\theta$$

The prior

$$p(\theta)$$

Captures our “belief” in the model
based on prior experience, observations or knowledge



$$p(\theta|x) = p(x|\theta)\ p(\theta)\ /\ p(x)$$

$$\hat{p}_0(\theta|x_0) = p(x_0|\theta)\ p(\theta)\ /\ p(x_0)$$

$$\hat{p}_1(\theta|x_1) = p(x_1|\theta)\ \hat{p}_0(\theta)\ /\ \hat{p}_0(x_1)$$

$$\vdots$$

$$\hat{p}_n(\theta|x_n) = p(x_n|\theta)\ \hat{p}_{n-1}(\theta)\ /\ \hat{p}_{n-1}(x_n)$$

Best estimate so far

Iterations with more data…

Selecting a prior

$$p(x|\theta, n) = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}$$

$$p(\theta|x) = \frac{\theta^x (1-\theta)^{(n-x)} p(\theta)}{\int \theta^x (1-\theta)^{(n-x)} p(\theta) d\theta}$$

We'd like the product of prior and likelihood to be "like" the likelihood

We'd like the integral to be easily evaluated

“Conjugate prior”

$$p(\bar{\theta}) = p(x|\theta)p(\theta)$$

Beta distribution

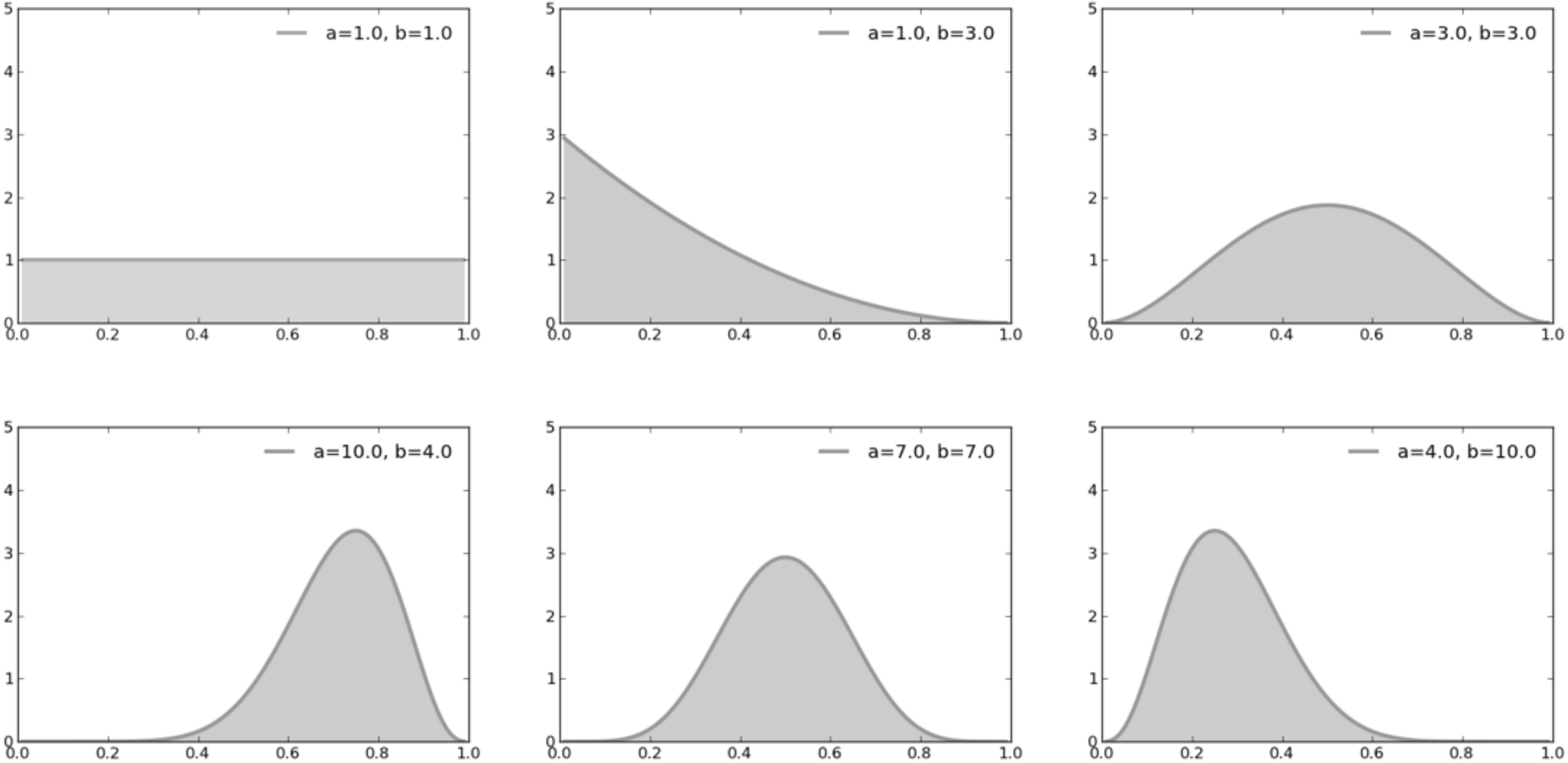
$$\mathrm{beta}(\theta|\mathrm{a},\mathrm{b}) = \theta^{(\mathrm{a}-1)}(1-\theta)^{(\mathrm{b}-1)} \ / \ \mathrm{B}(\mathrm{a},\mathrm{b})$$

number of heads + 1

number of tails + 1

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

Beta distribution



Putting it together...

binomial likelihood $\qquad$ beta prior

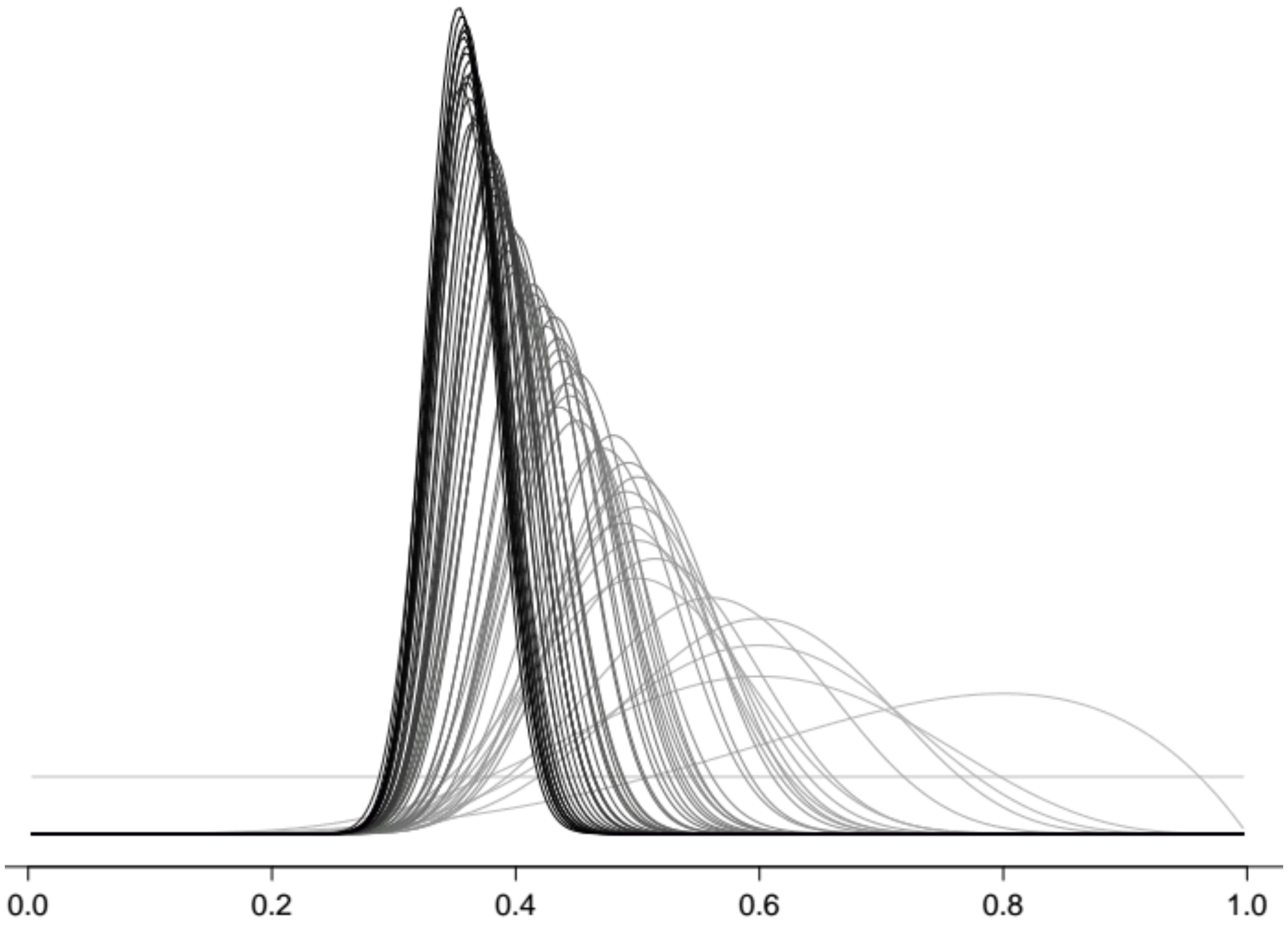
$$p(\theta|x,n) = \underbrace{\theta^{x}(1-\theta)^{(n-x)}}_{\text{binomial likelihood}}\ \underbrace{\theta^{(a-1)}(1-\theta)^{(b-1)}}_{\text{beta prior}}\ /\ B(a,b)\ p(x)$$

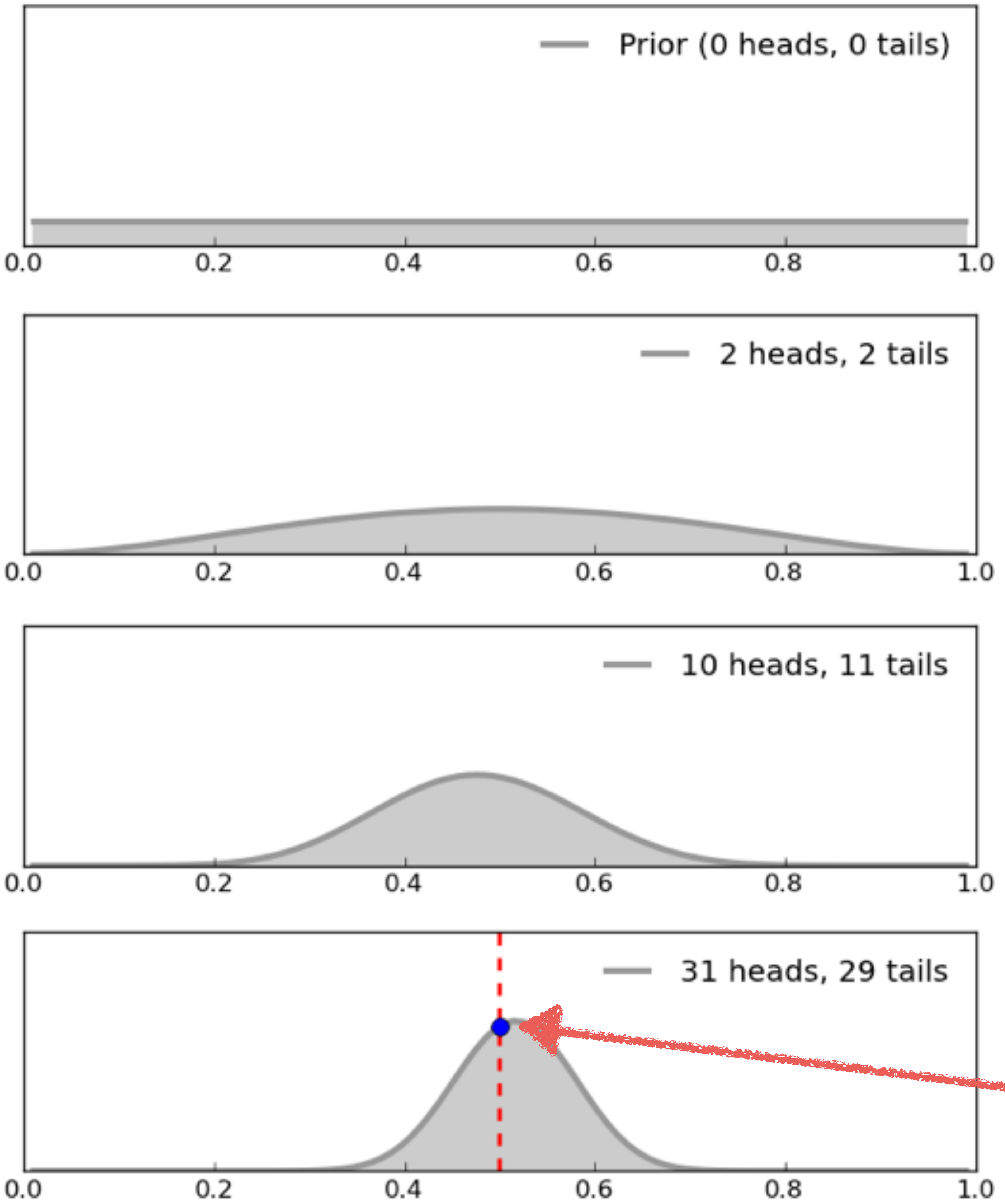
$$= \theta^{x+a-1}(1-\theta)^{(n-x+b-1)}\ /\ B(x+a, n-x+b)$$

number of heads x $\qquad$ number of tails (n-x)

Putting it together…

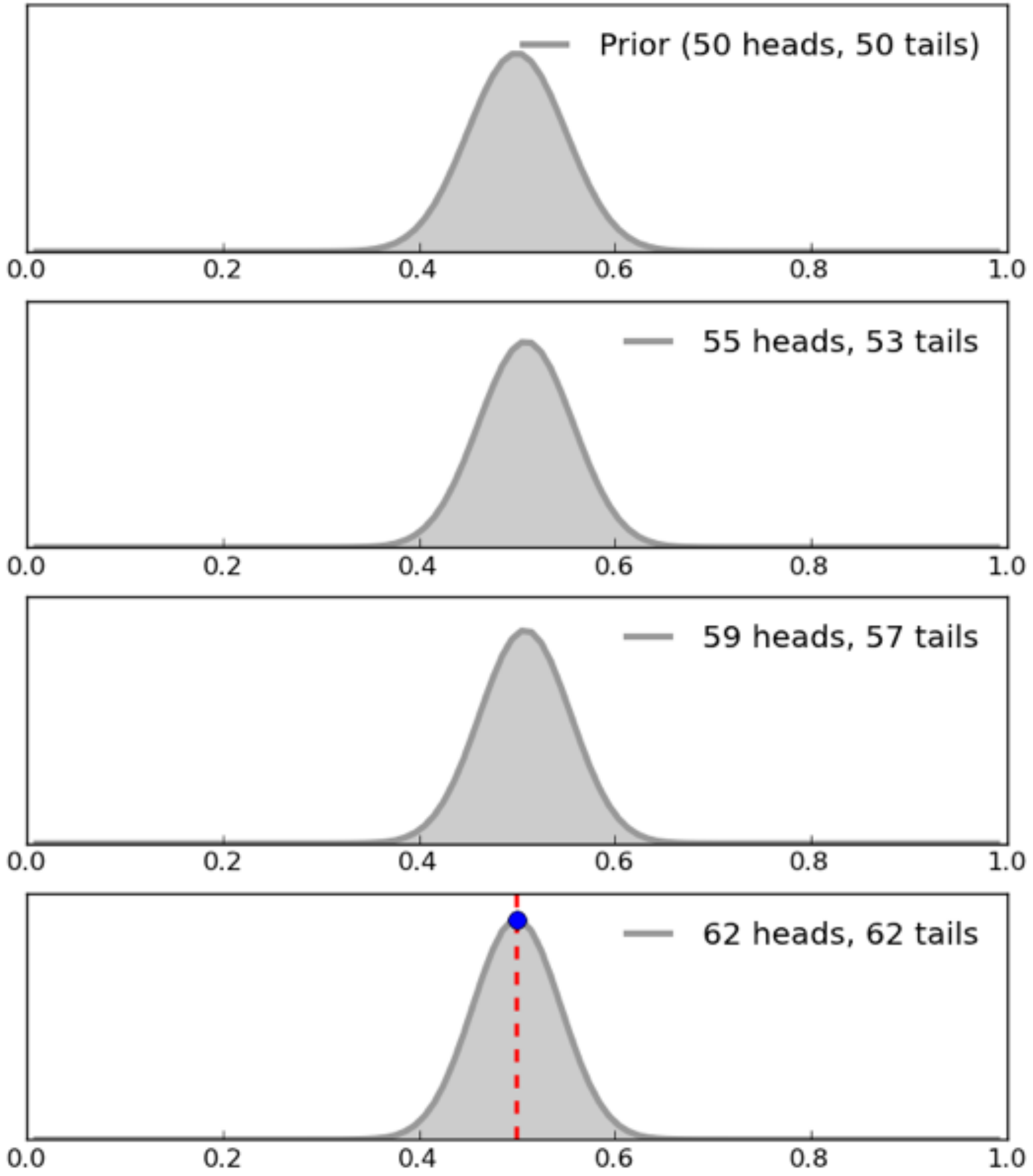
1. Decide on a **prior**, which captures your **belief**
2. Run experiment and **observe data** for heads, tails
3. Determine your **posterior** based on the data
4. Use posterior as your **new belief** and re-run experiment
5. Rinse, repeat until you hit an actionable **certainty**


0.0
0.2
0.4
0.6
0.8
1.0



Uniform prior
“Fair” coin

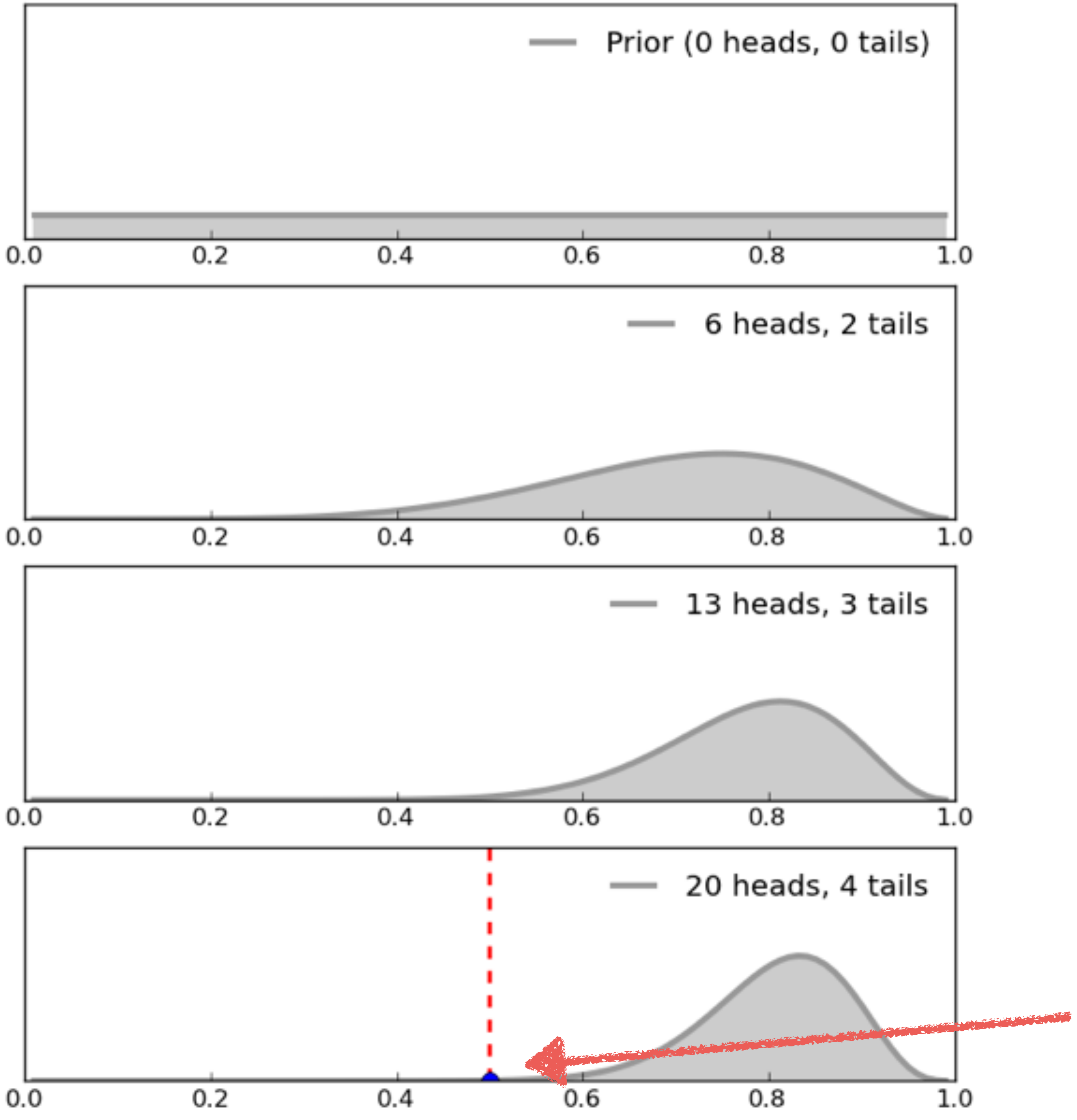
Pretty sure coin is fair



“Coin is fair” prior

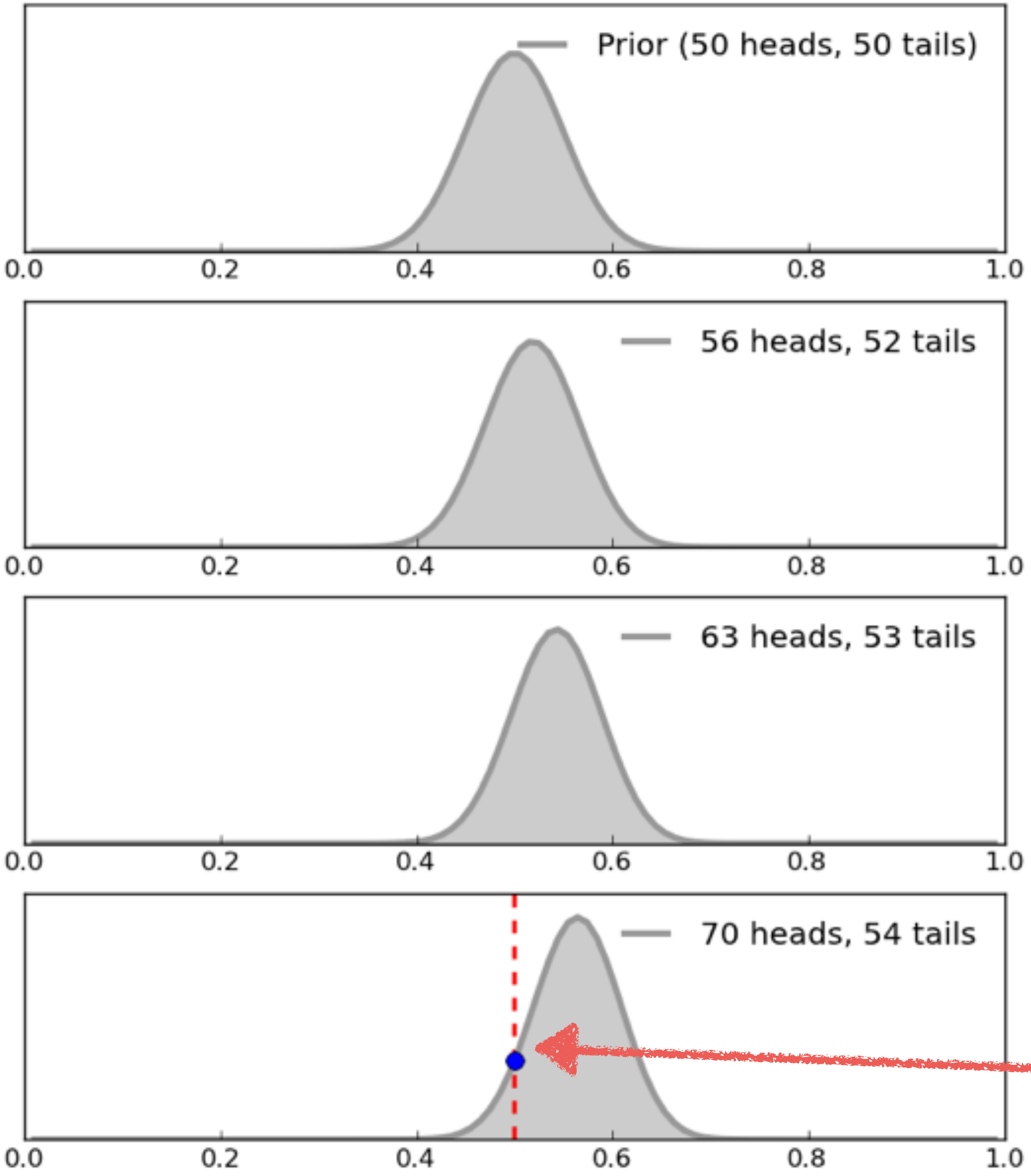
“Fair” coin

Very sure coin is fair



Uniform prior
"Biased" coin

Pretty sure coin is unfair

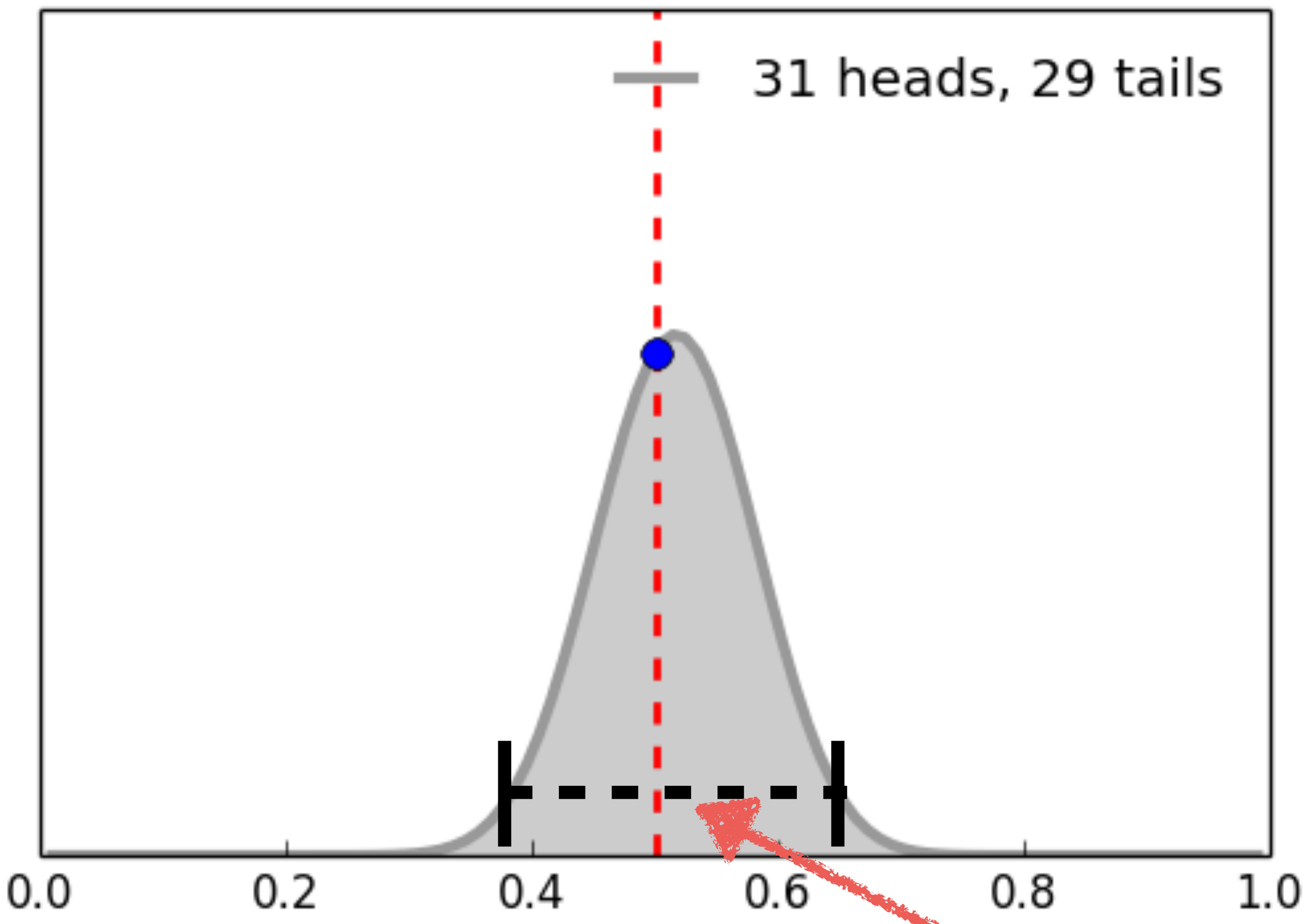


“Coin is fair” prior
“biased” coin

Not sure of anything yet!

When to reject?

The Credible Interval

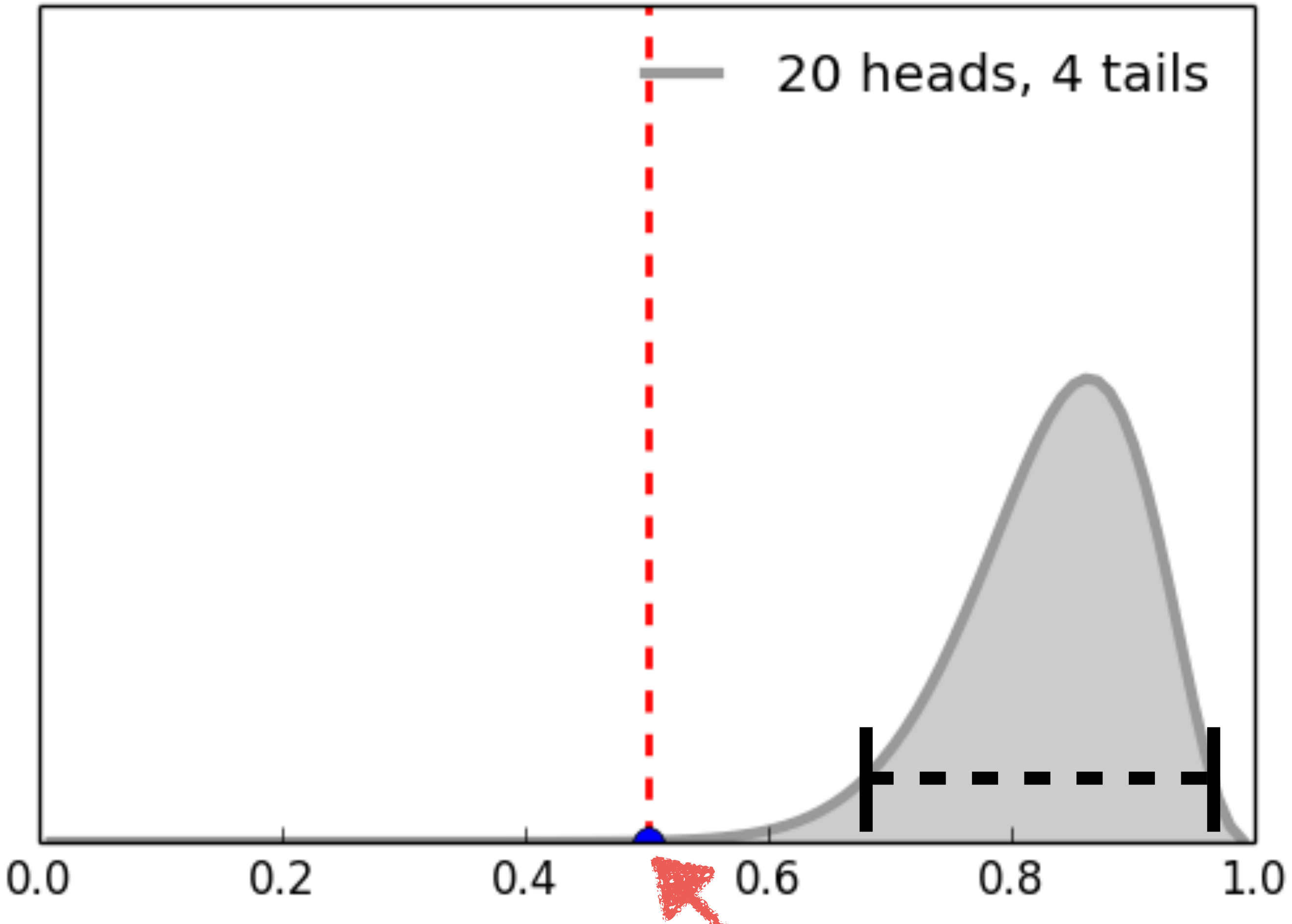


Uniform prior

“Fair” coin

95% credible interval

The Credible Interval



Uniform prior

“Biased” coin

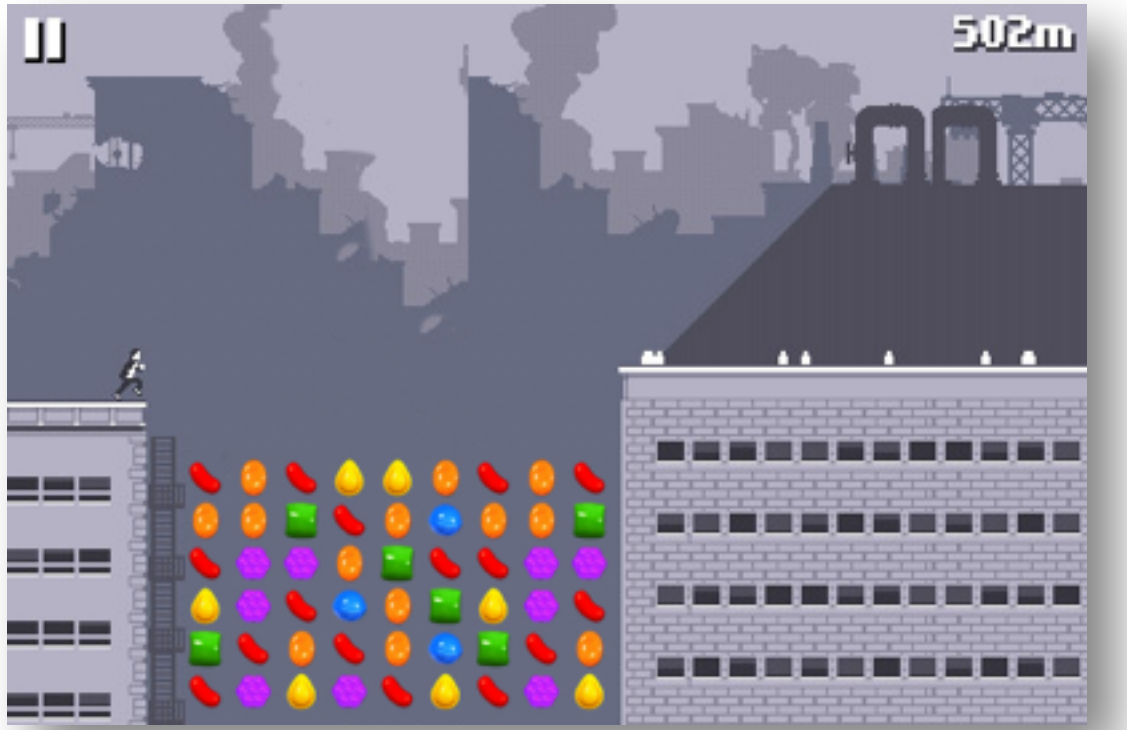
Outside credible interval

The Prior

- Captures our prior belief, expertise, opinion
- Strong prior belief means:
  - we need lots of evidence to contradict
  - results converge more quickly (if prior is “relevant”)
- Provides inertia
- With enough samples, prior’s impact diminishes, rapidly

Running a test…

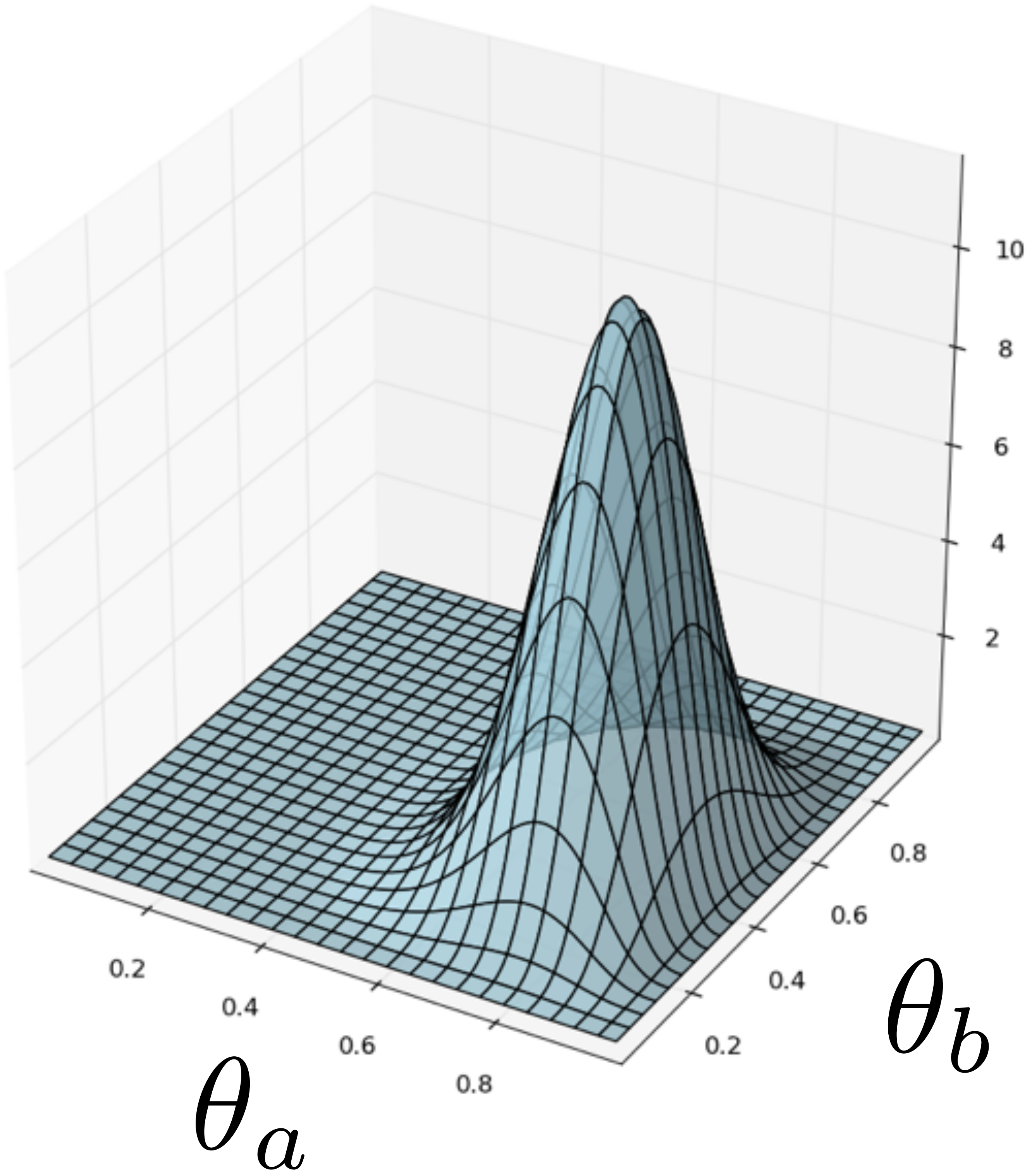
A



B

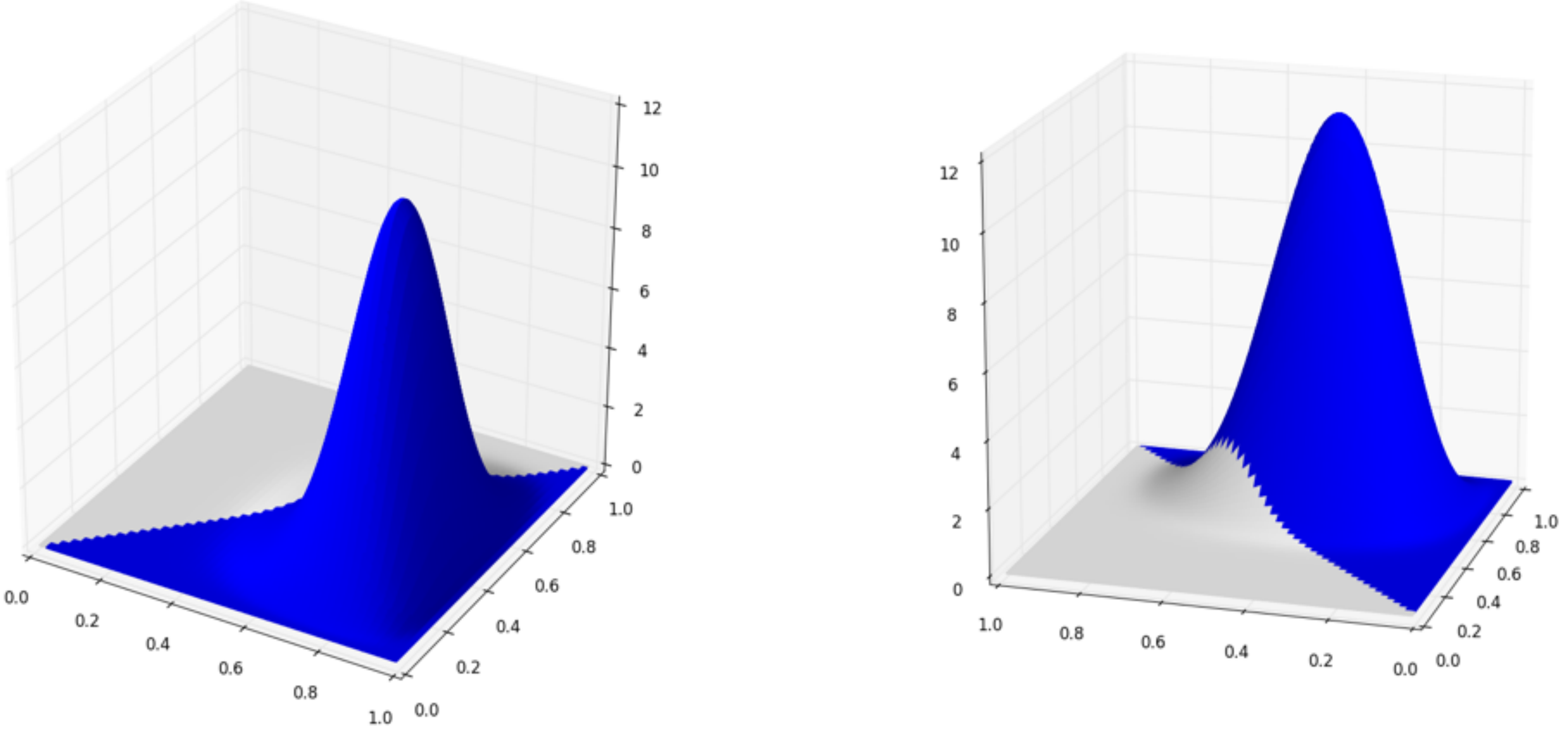
Multiple variant tests

- With 1 or more variants we have a multi-dimensional problem
- Need to evaluate volumes under the posterior
- In general requires numerical quadrature = Markov Chain Monte-Carlo (MCMC)


10
8
6
4
2
0.2
0.4
0.6
0.8
0.2
0.4
0.6
0.8
θ_a
θ_b

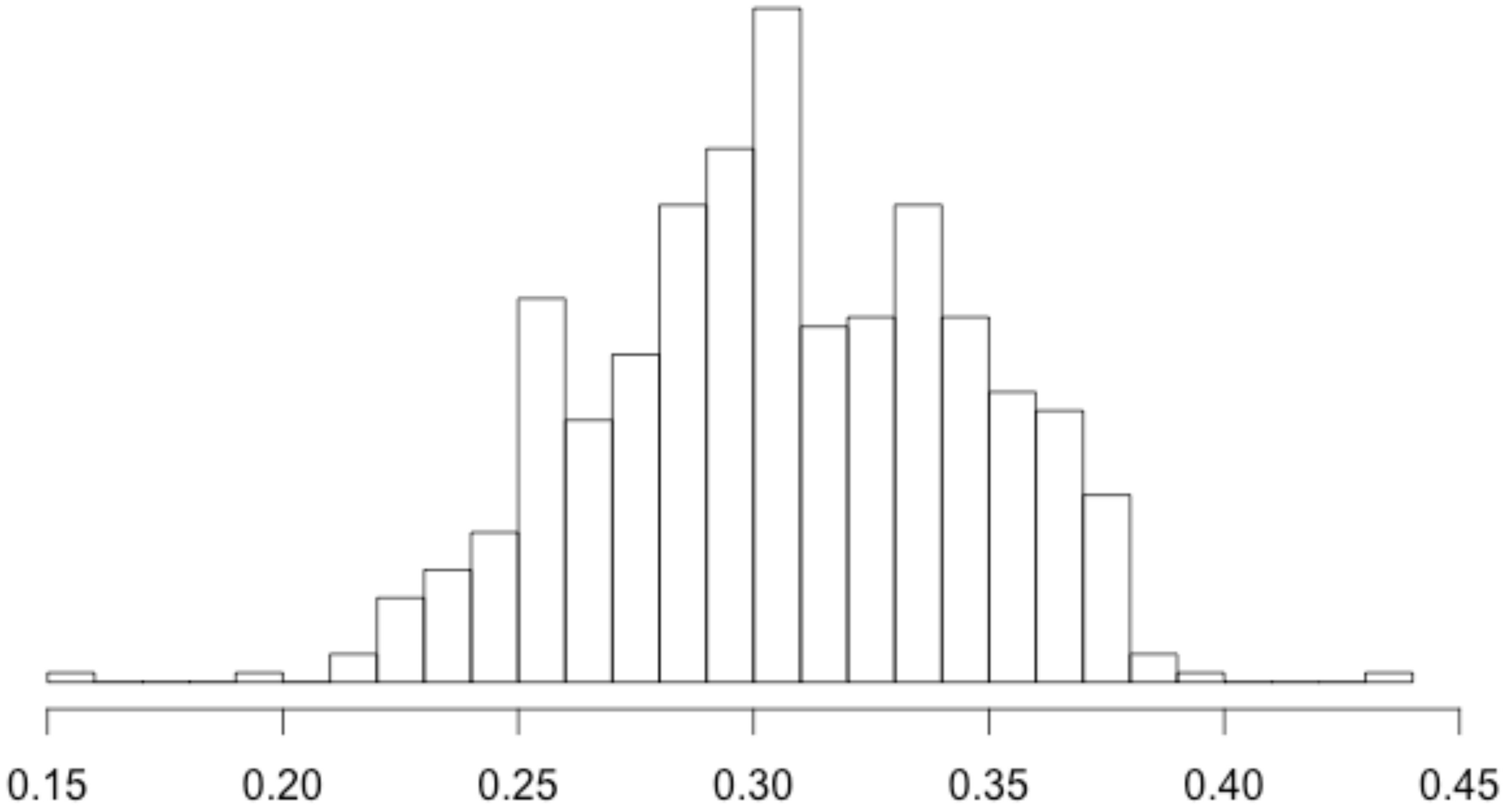
Probability of Winning

$$p(\theta_a > \theta_b) = \int_{\theta_a > \theta_b} p(\theta_a | x_a) p(\theta_b | x_b) \; d\theta_a d\theta_b$$

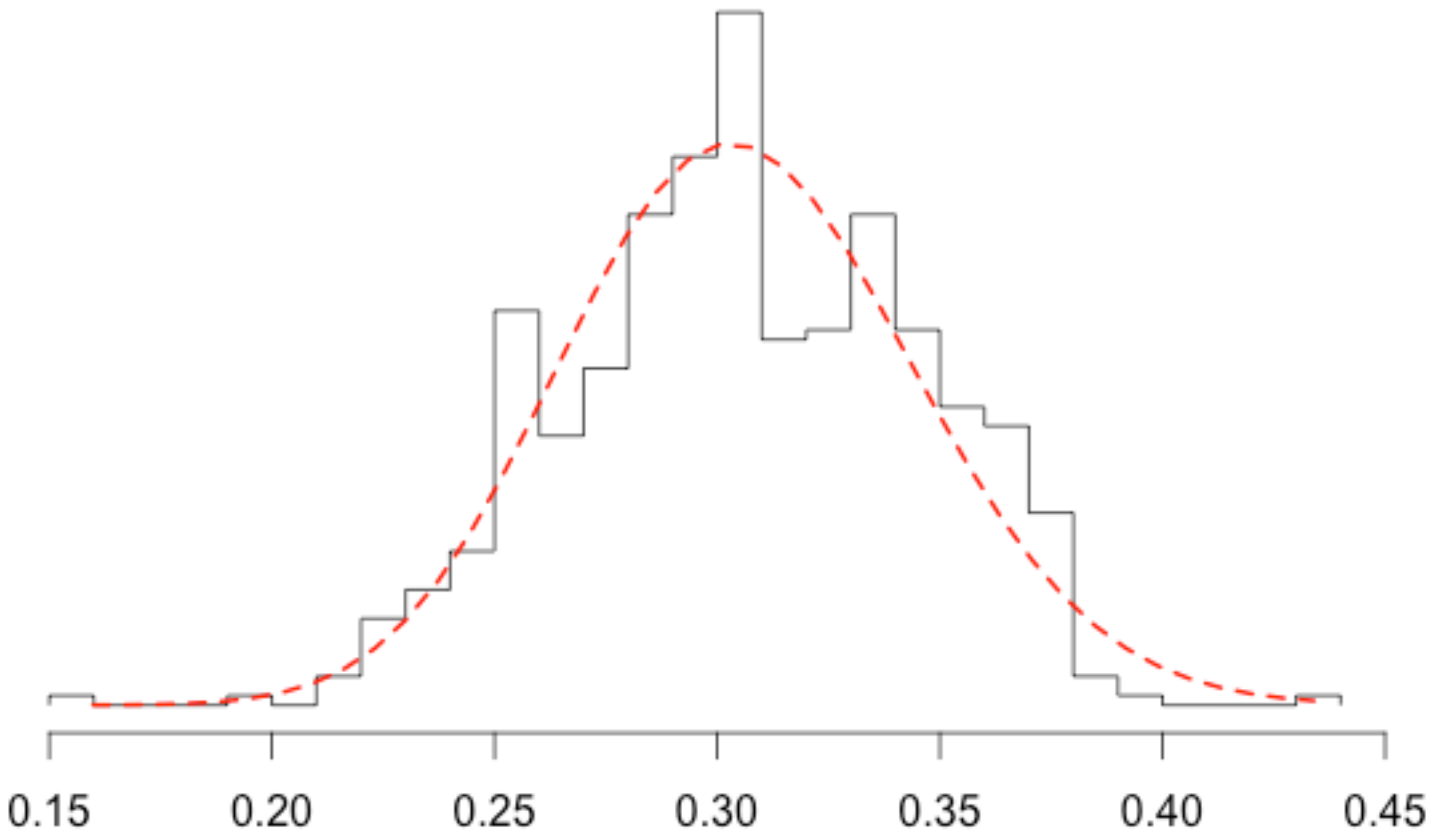


$\theta_a > \theta_b$

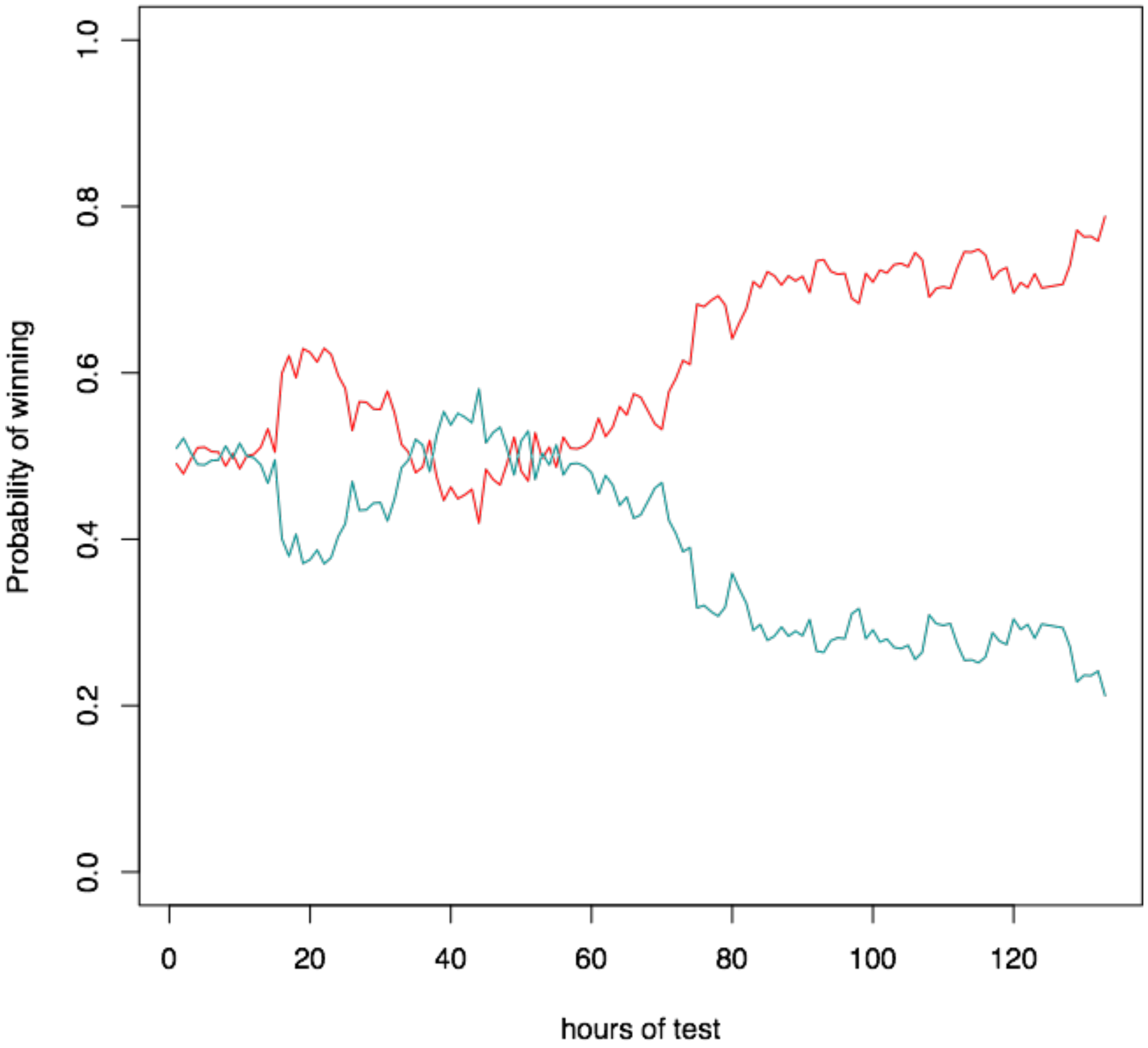
What’s the prior?



Fit a beta



a=42, b=94

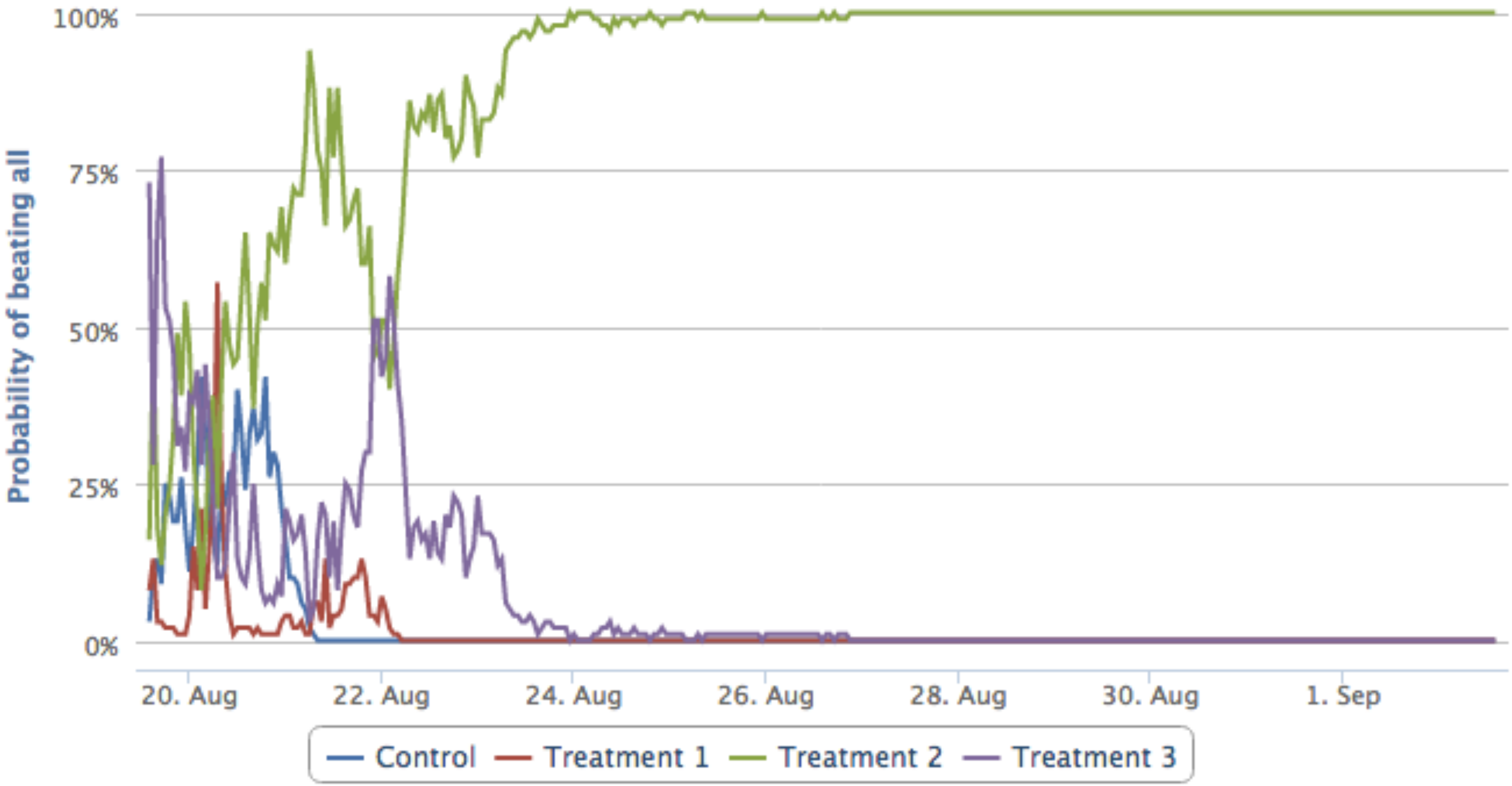


Probability of winning
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
hours of test

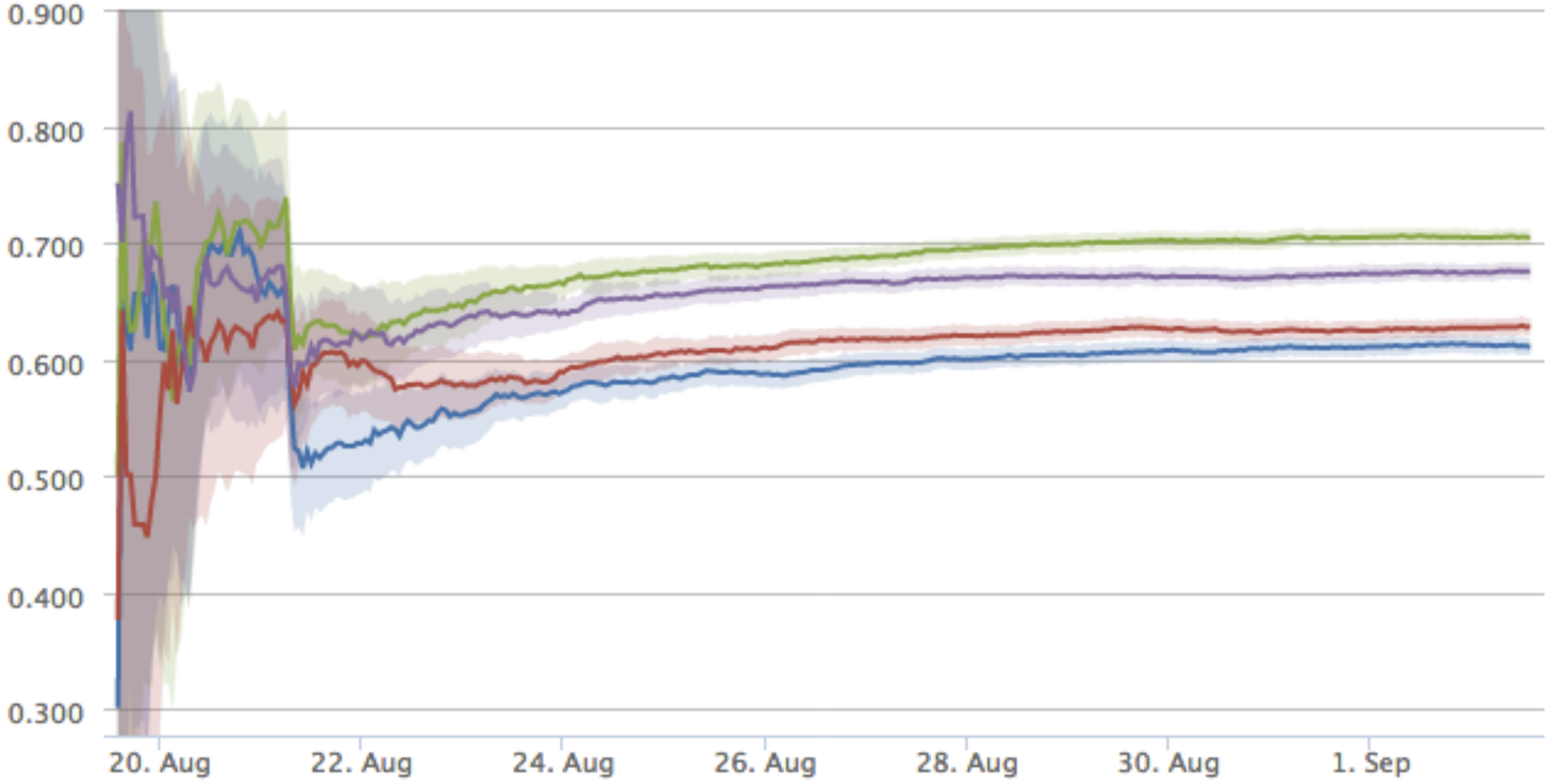
Some examples…

	Variant	Score	Change	Probability of beating control	Probability of beating all	Conversions / Participants
	Control	0.611			0%	6,870 / 11,243
	Treatment 1	0.6276	+2.71%	100%	0%	7,037 / 11,212
	Treatment 2	0.7044	+15.27%	100%	100%	7,955 / 11,294
	Treatment 3	0.6755	+10.55%	100%	0%	7,616 / 11,274

A successful test



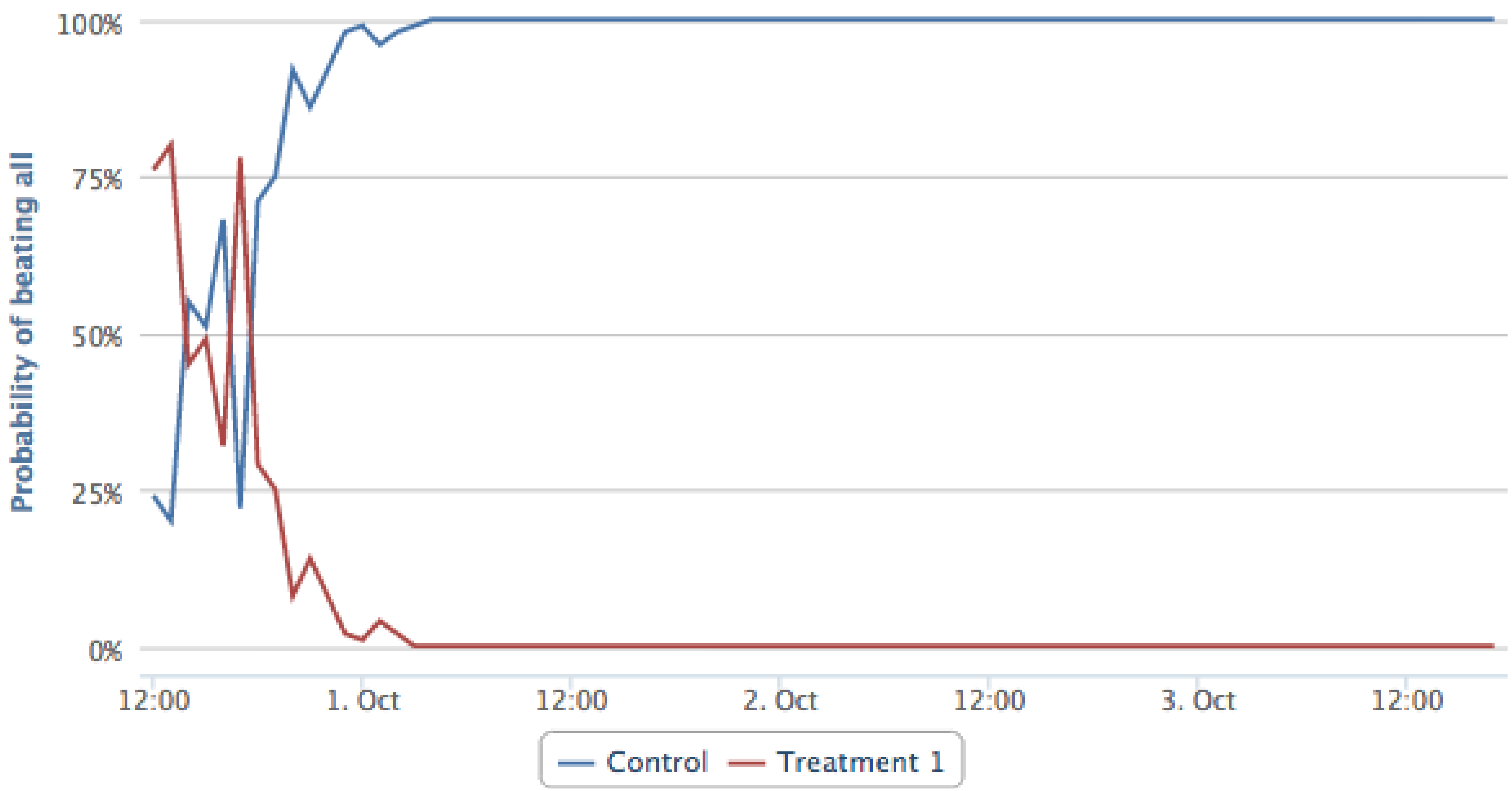
Probability of beating all



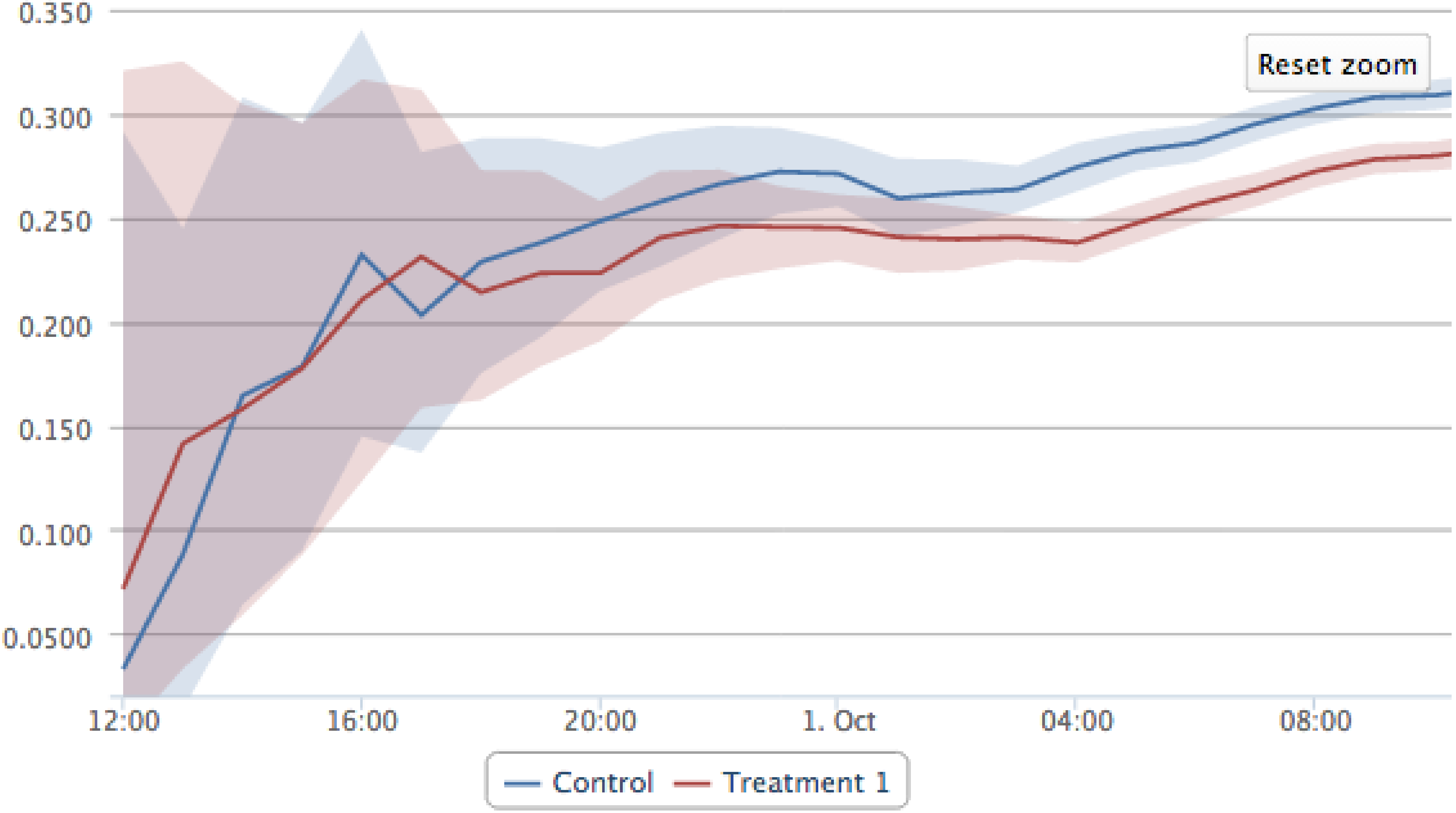
Observed conversion rates (with CI bounds)

	Variant ⓘ	Score ⓘ	Change ⓘ	Probability of beating control ⓘ	Probability of beating all ⓘ	Conversions / Participants ⓘ
⭐	Control	0.3774			100% ✅	15,567 / 41,244
	Treatment 1	0.3477	-7.88%	0% 🔴	0% 🔴	14,385 / 41,372

A ~~not so~~ successful test...



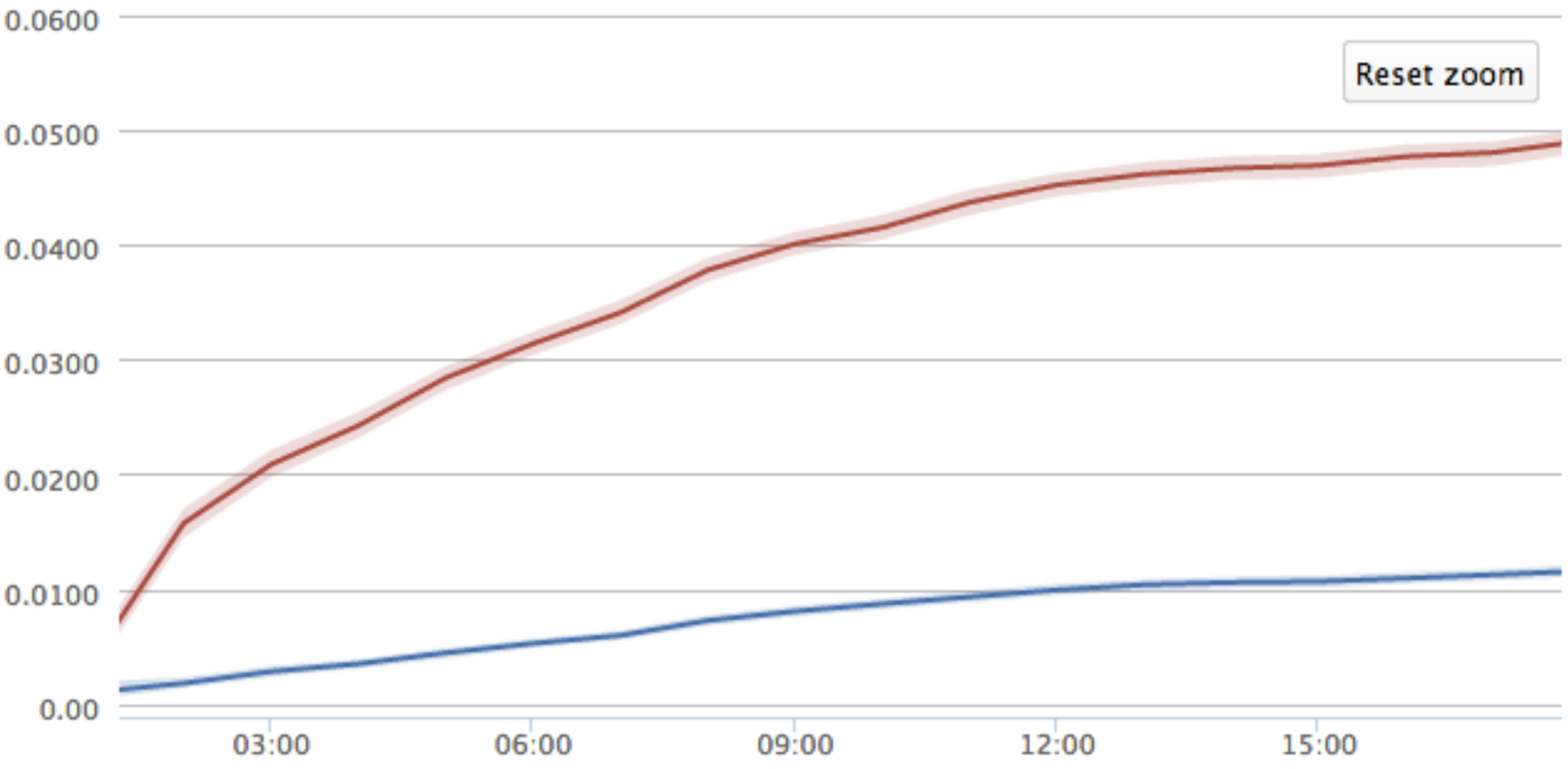
Probability of beating all



Observed conversion rate (posterior)

Assumptions

- Users are independent
- User’s convert quickly (immediately)
- Probability of conversion is independent of time



Un-converged conversion rate

Benefits / Features

- Continuously observable
- No need to fix population size in advance
- Incorporate prior knowledge / expertise
- Result is a "true" probability
- A measure of the difference magnitude is given
- Consistent framework for lots of different scenarios

Useful Links

- https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers
- "Doing Bayesian Data Analysis: A Tutorial with R and Bugs", John K. Kruschke
- http://www.evanmiller.org/how-not-to-run-an-ab-test.html
- http://www.kaushik.net - Occam's Razor Blog
- http://exp-platform.com - Ron Kovahi et al.

Thanks

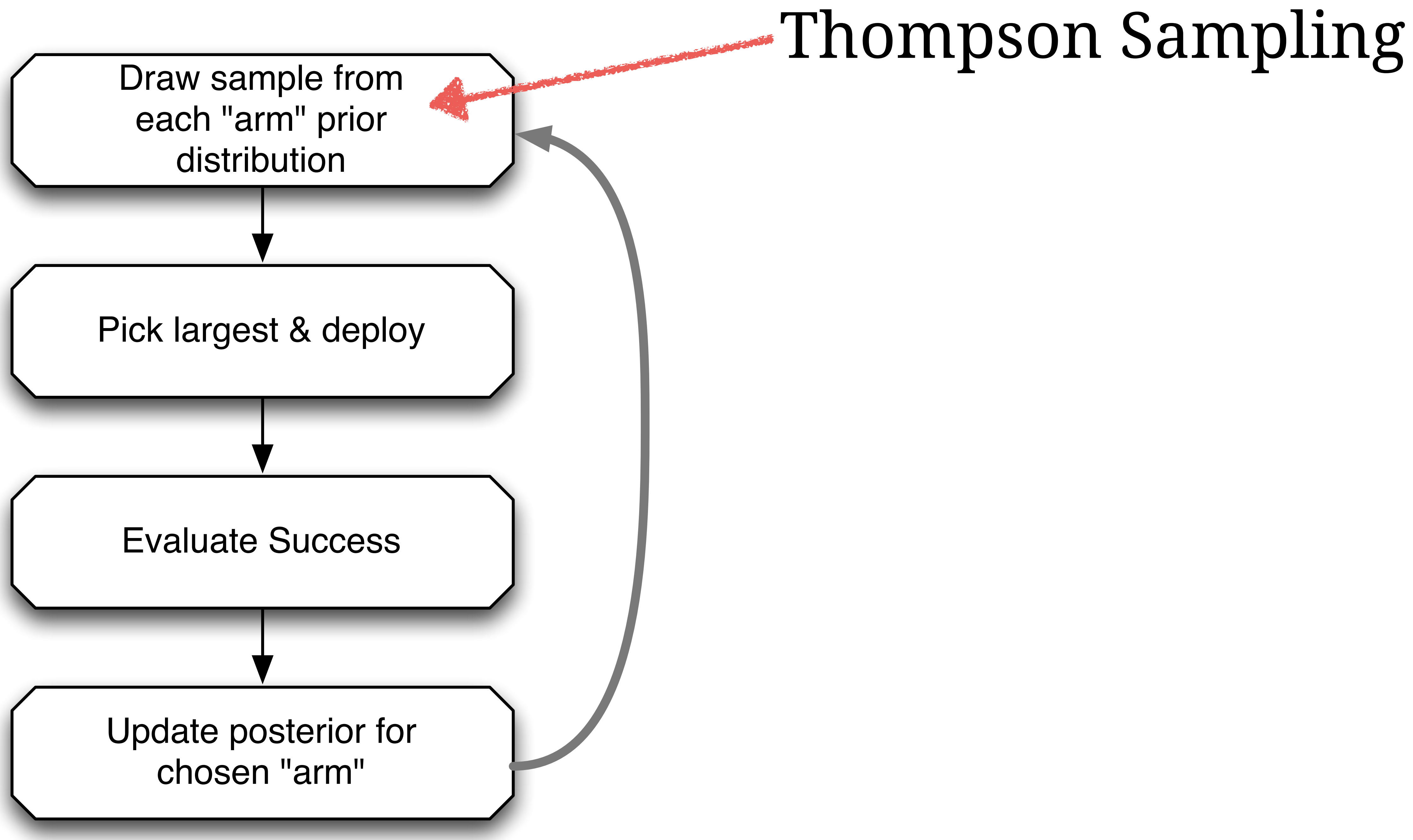
steve@swrve.com
@stevec64

Multi-arm bandits

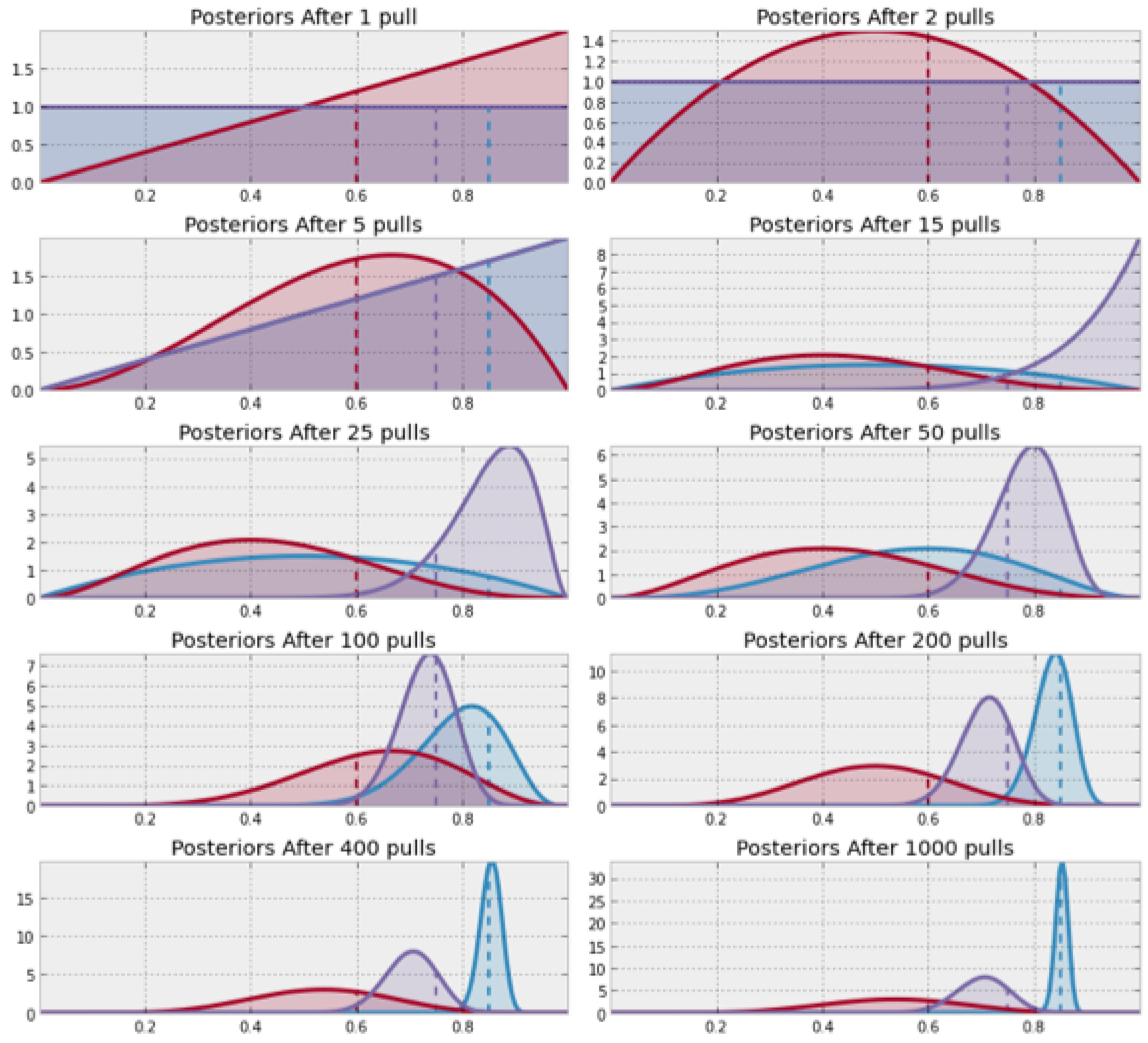
5%	6%	10%	4%

Pull 1	✗	✗	✗	💰
Pull 2	✗	✗	✗	✗
Pull 3	✗	✗	💰	✗

Multi-arm bandits



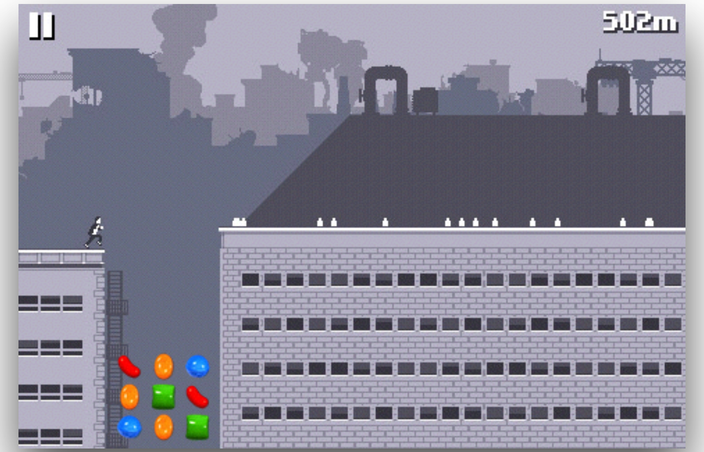
Thompson, William R. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". Biometrika, 25(3–4):285–294

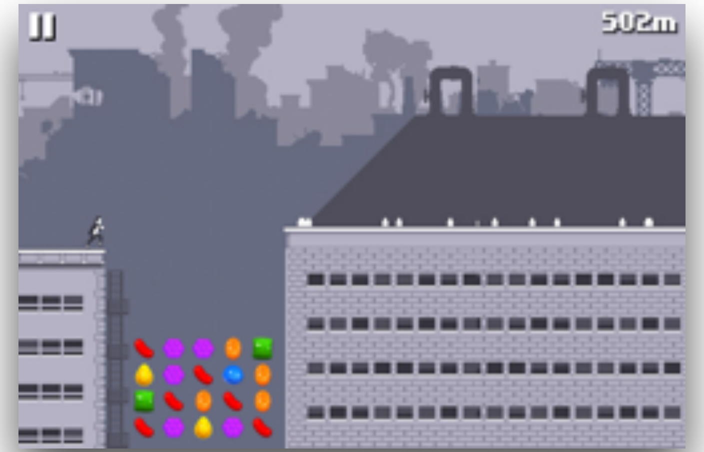


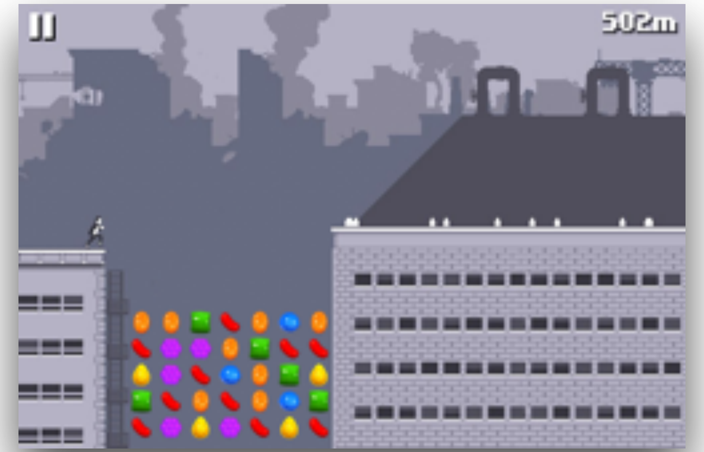
https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers

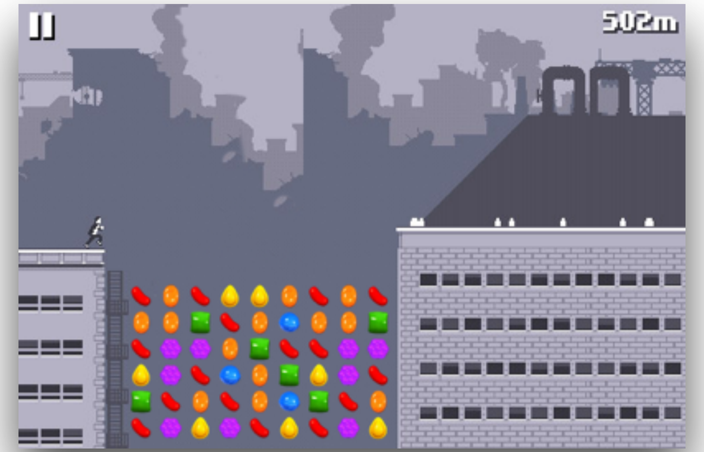
Difficulty tuning

a slight silly example...

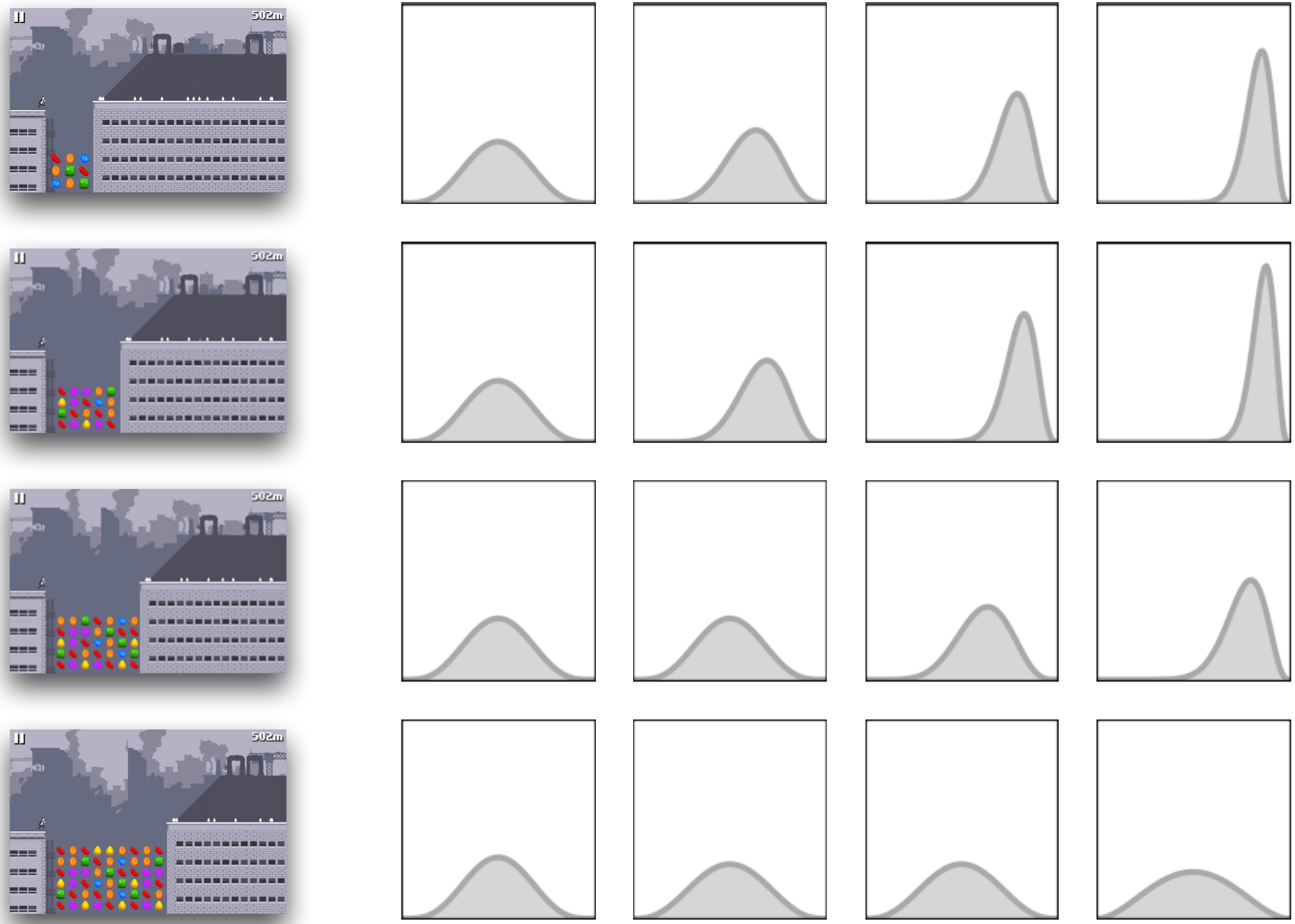








Canacandycrushbalt



Winner