

Bayesian A/B-Testing

RecSys Summer School - June 2023
Copenhagen

Morten Arngren
Lead Data Scientist



WUNDERMAN



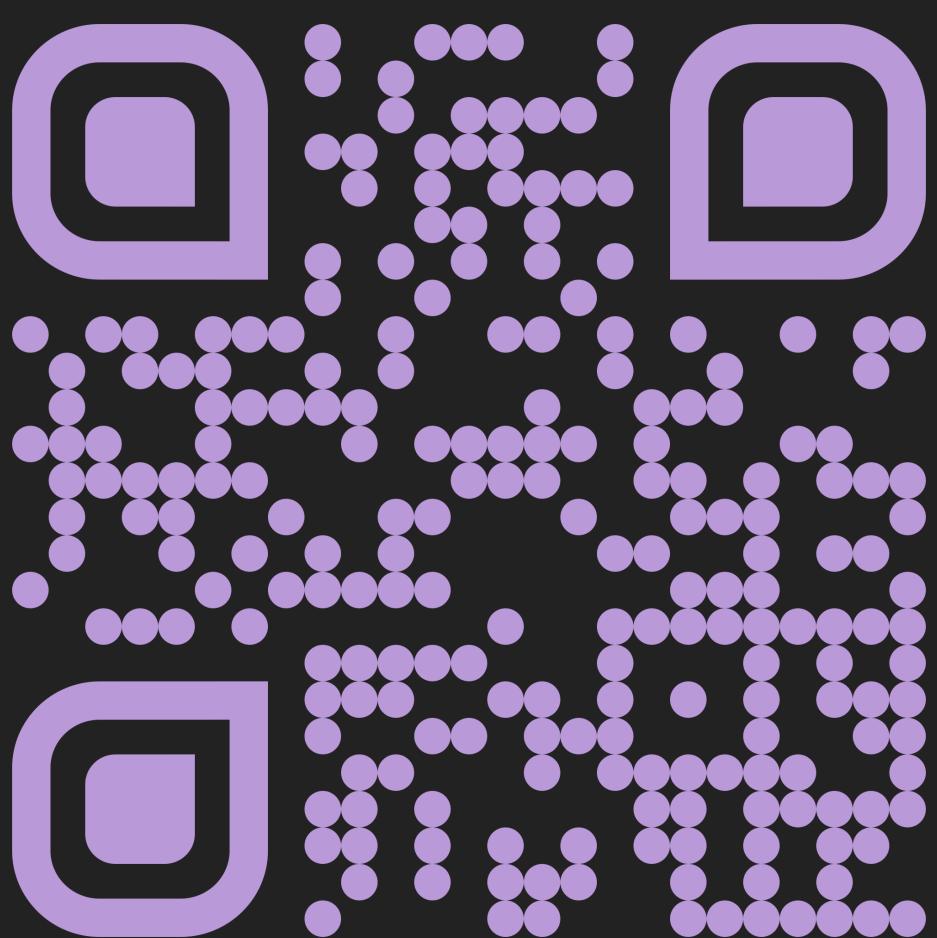
Code and keynote
available on GitHub

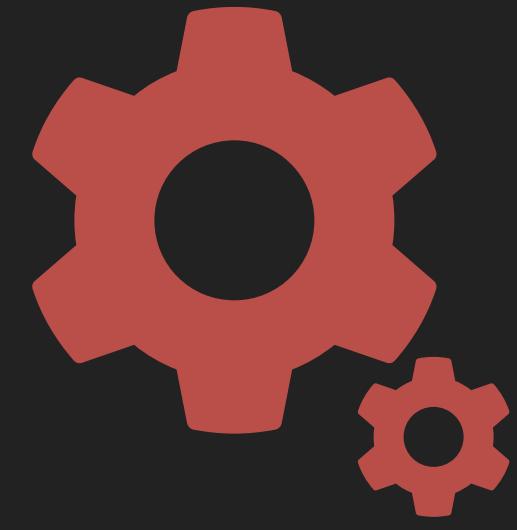
<https://github.com/Arngren>

Morten Arngren

PhD in Machine Learning in 2011
ex-Nokia | ex-Issuu | ex-Adform
(My first Rec. Engine)

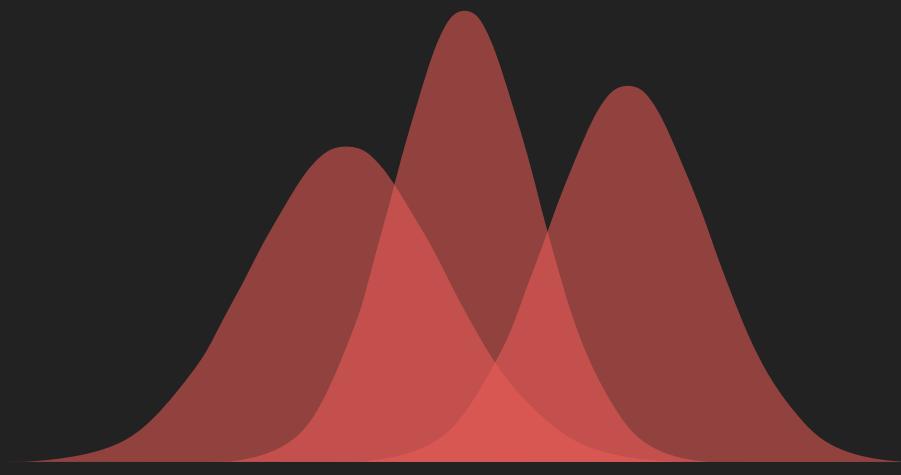
Lead Data Scientist at
WundermanThompson





Classic & Complex
A/B testing

Hypothesis testing
and the infamous
 p -value



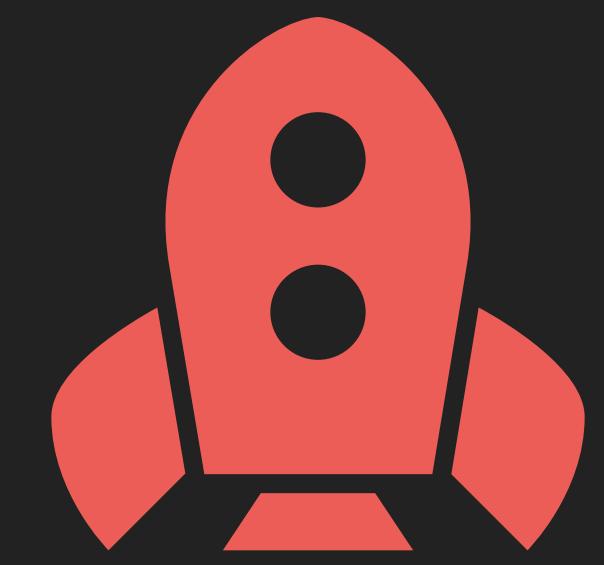
Bayesian A/B testing

A/A-testing
Bayesian evaluation



Campaign

Simulation



the stage

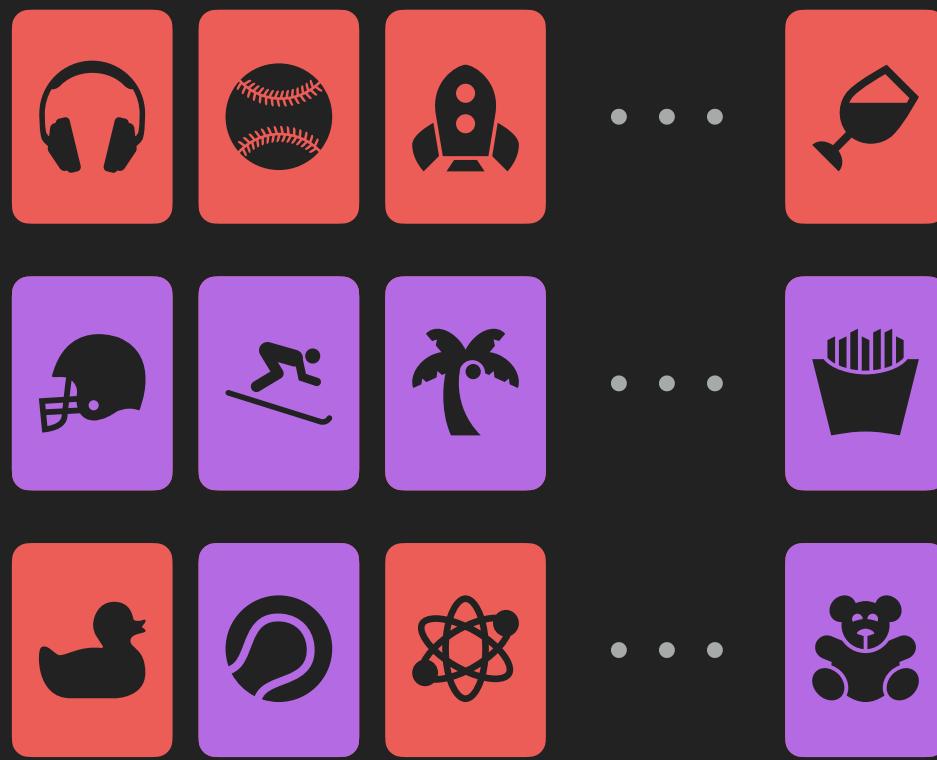
The stage

...

Two Rec. Engines

Movie Streaming

Recommendations



\

Slates



Running in production

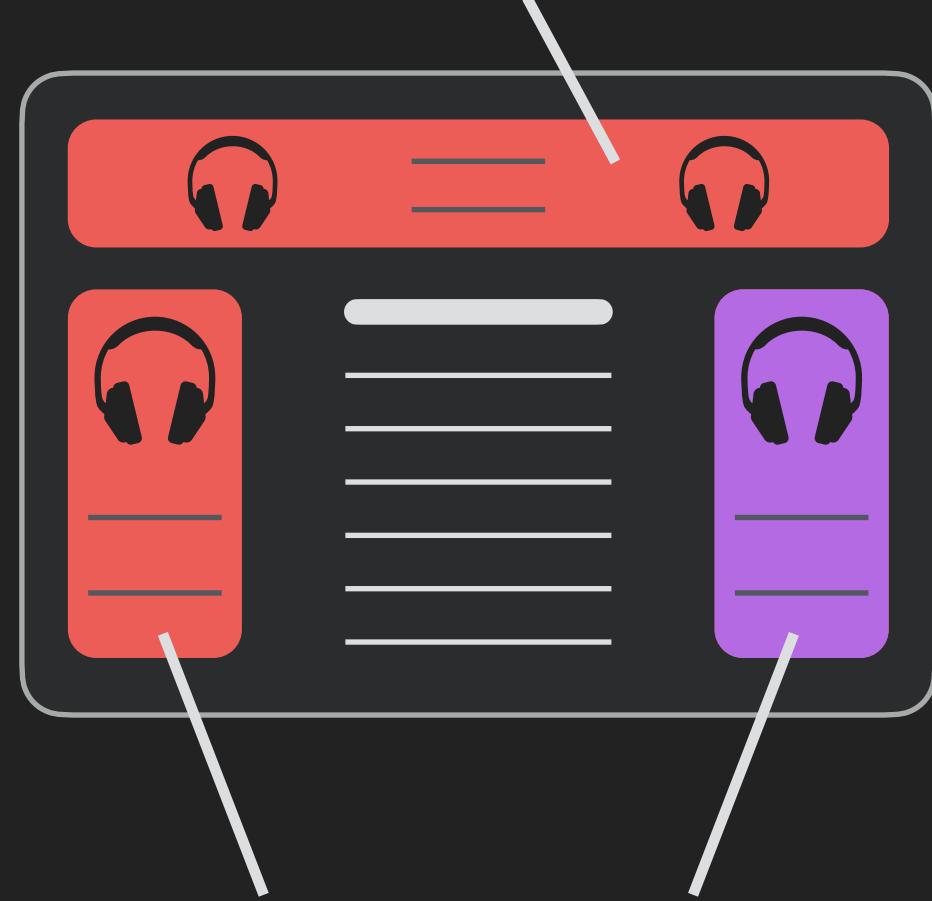


New developed Rec. Engine

which has
the best
performance
?

News site

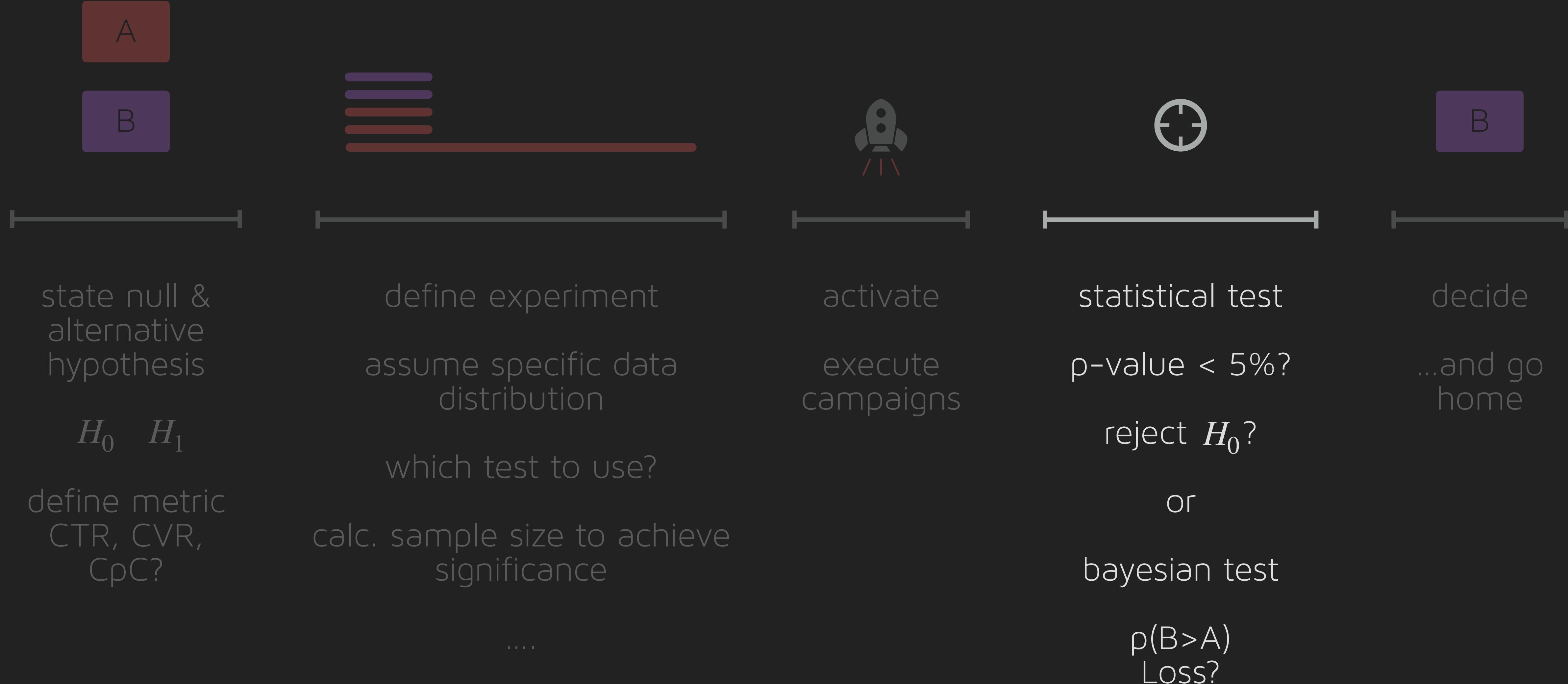
Advertisement



Advertisements

A/B testing

Classic approach

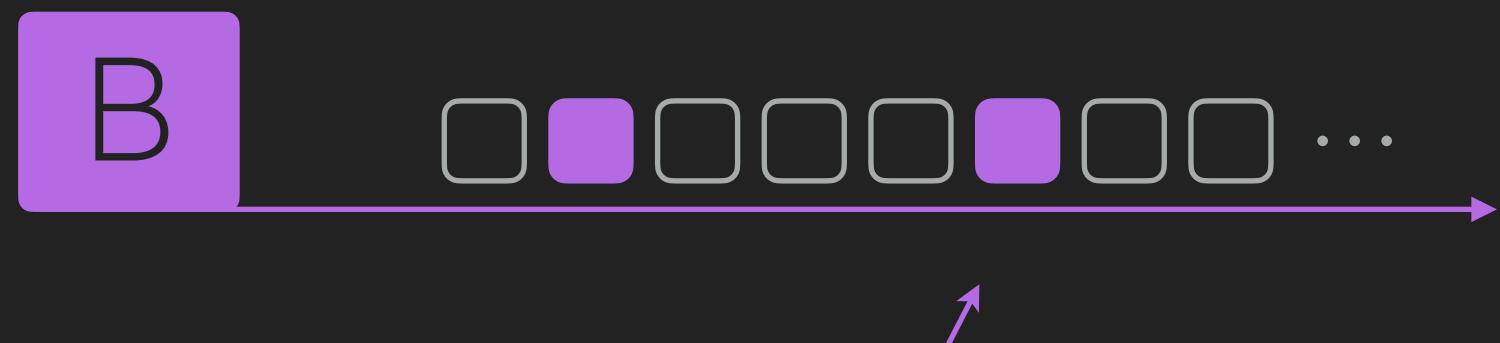
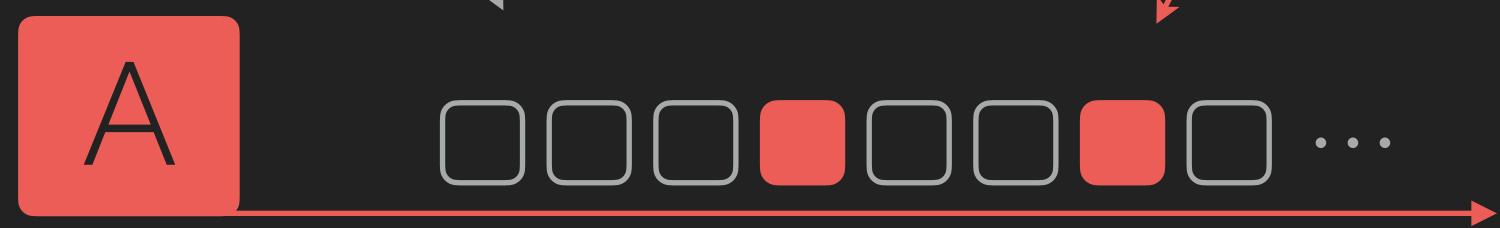


The stage

...
...
...

logs

impressions clicks / conversions



clicks / conversions

business metrics

CTR | Click-Through-Rate
Fraction of clicks? [%]

CVR | Conversion Rate
Product buys [%]

CpC | Cost-per-Click
Price of a click [\\$]

??? | Webshops margins
Earnings [\\$]

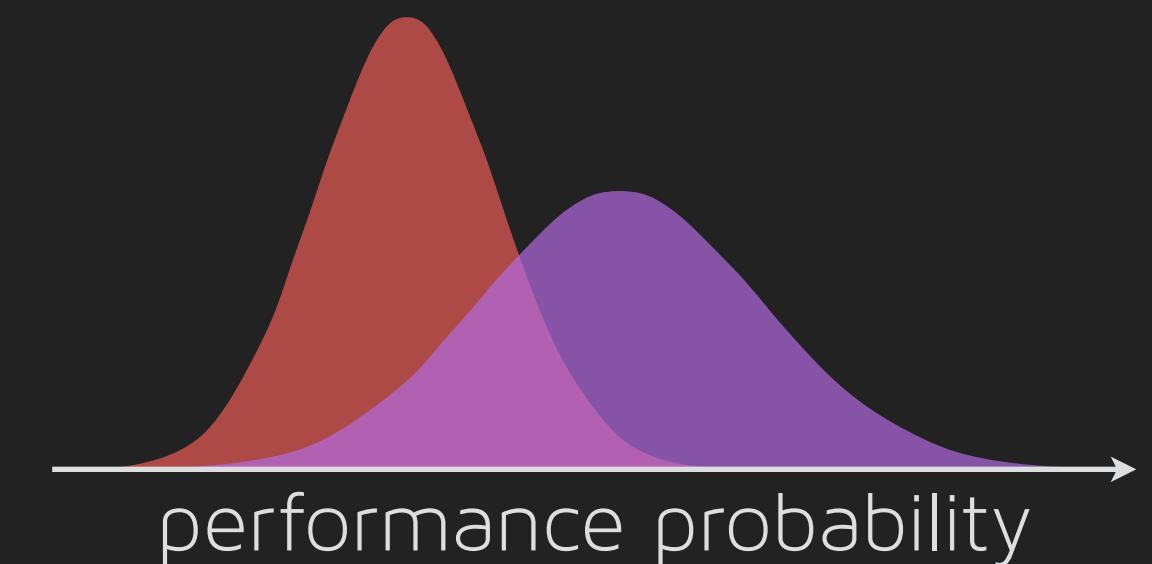
:

evaluation

Hypothesis testing
 $p\text{-value} < 5\%?$

or

Bayesian testing

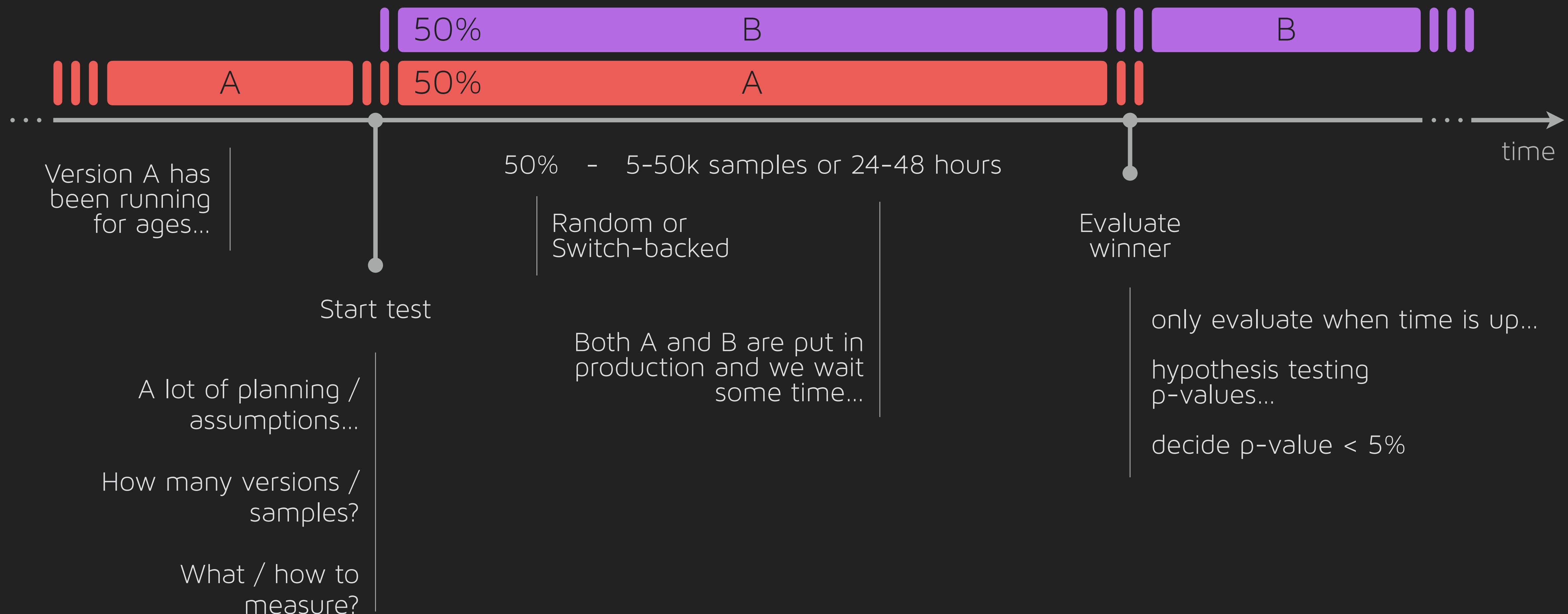




Introduction

Classic A/B-test

Timeline...



Complex A/B-test

...or the more right way...

$\frac{\sim 20\%}{}$

Reserve

don't spend it all on testing,
what if B sucks....

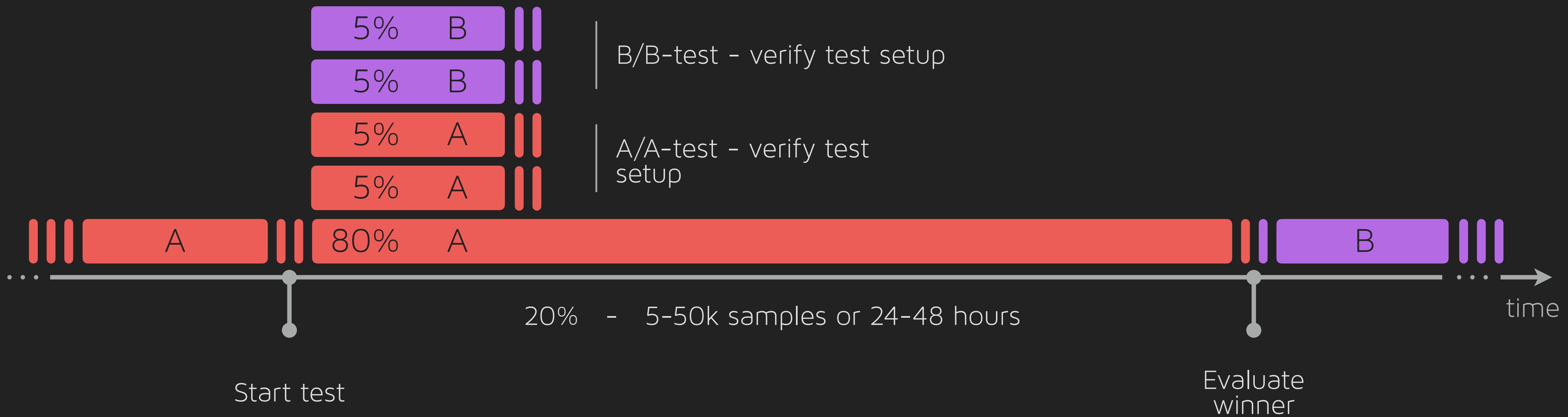
consider business requirements
and price...!

$\frac{A/A \text{ test}}{}$

Verify

split into several smaller tracks

conducting an A/A test to verify
the whole test setup



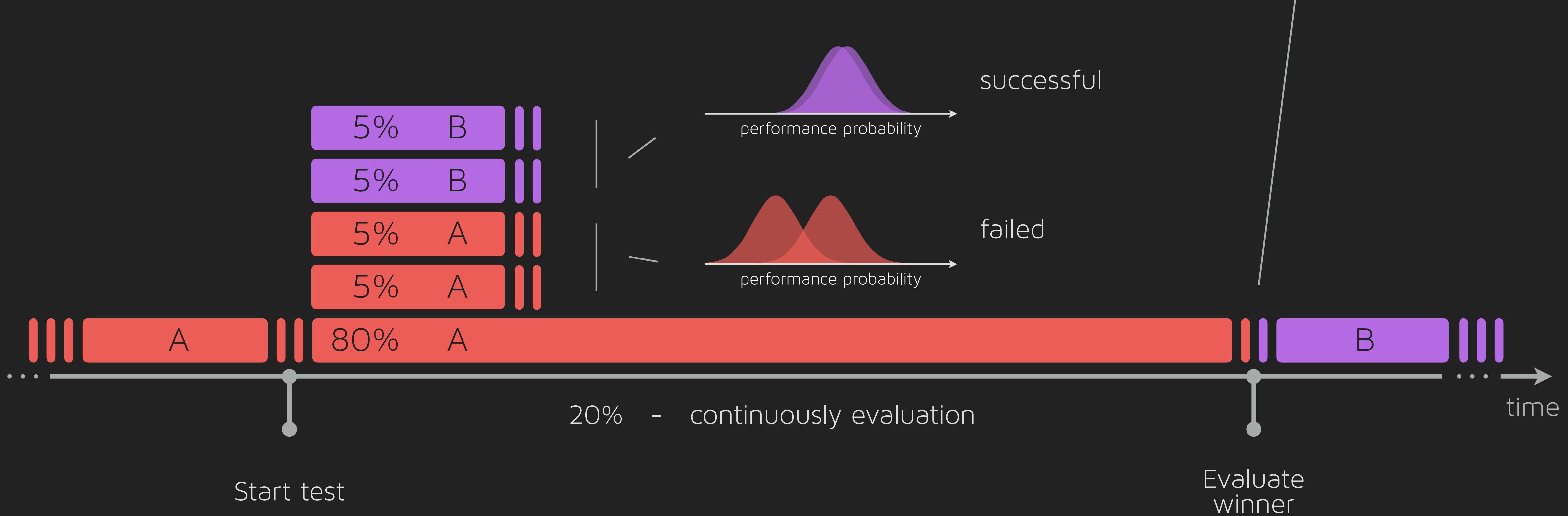
Bayesian A/B-test

Bayesian evaluating - probability of winner

models performance as probability distributions

captures the epistemic uncertainty

allow us to quantify the confidence



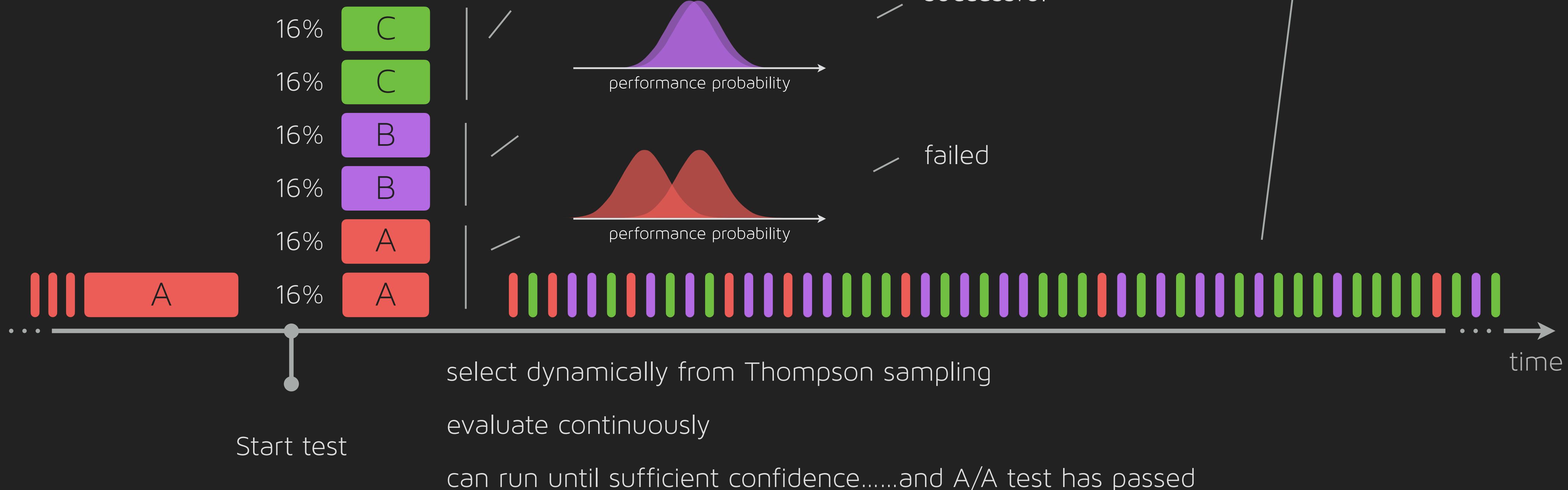
Bandit A/B-test

Dynamic architecture - runs self-contained - drop in new versions as cold start

models performance as probability distributions

captures the epistemic uncertainty

allow us to quantify the confidence



What we really want to know...

λ Performance - eg. CTR / CVR / CpC

$$P(\lambda_B > \lambda_A)$$

Probability of B
being better than A

Hypothesis test answers
another question

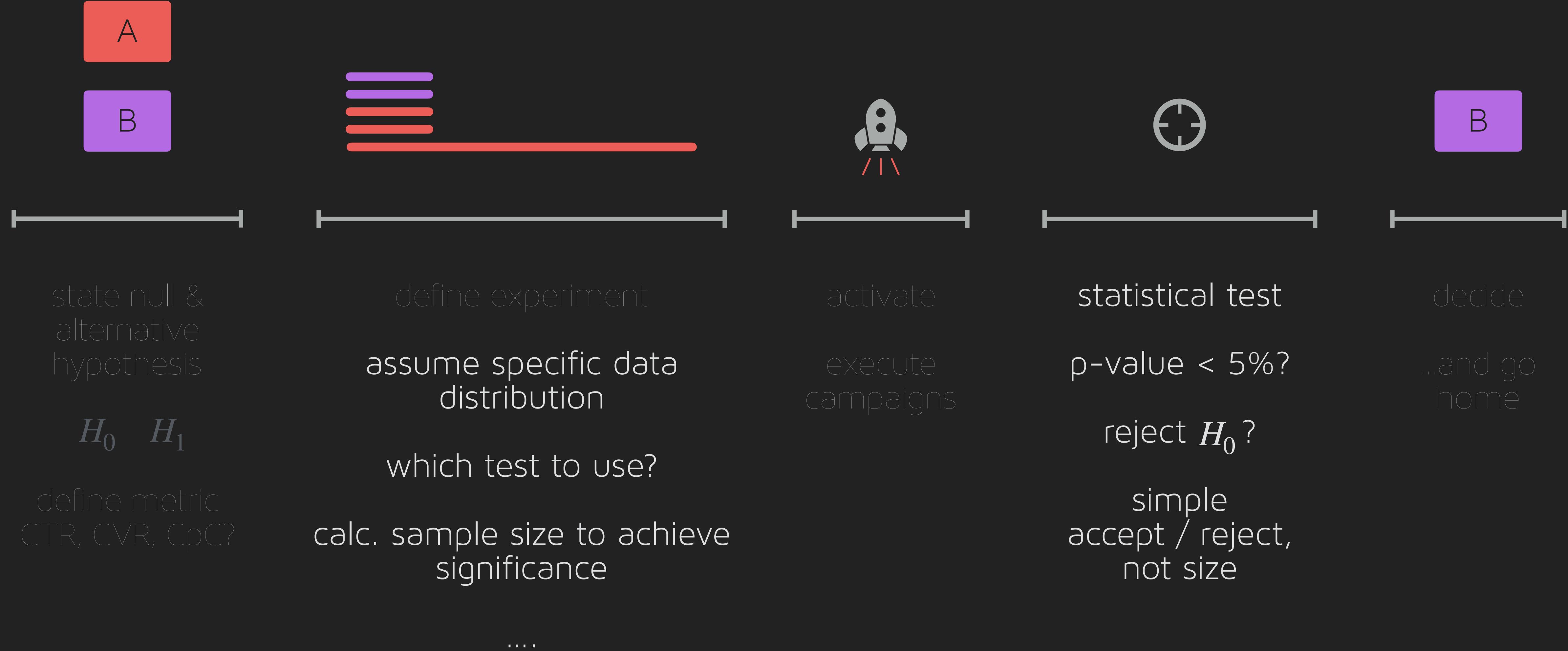
Bayesian approach
answers this question



Hypothesis Testing

Hypothesis significance testing

What's specific to it...?



Hypothesis significance testing

Popular hypothesis tests...

Students T-Test



Assumptions

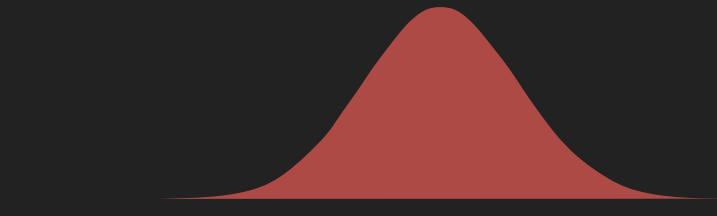
Student-t distributions

Very few samples

Mean of two identical distributions

https://en.wikipedia.org/wiki/Student%27s_t-test

Z-test



Assumptions

Gaussian distributions

Homogeneity of variance

Mean of two identical Gaussian distributions

<https://en.wikipedia.org/wiki/Z-test>

Fisher's exact test



Assumptions

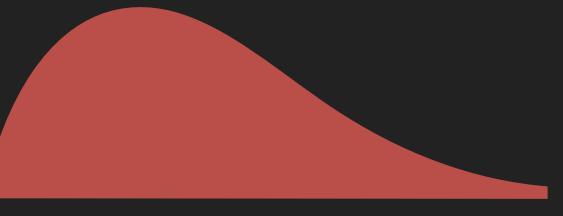
Contingency tables

Small sample sizes

Count data evaluating ratios

https://en.wikipedia.org/wiki/Fisher%27s_exact_test

Chi-squared test



Assumptions

Contingency table

Large sample sizes

Count data evaluating ratios

Chi-squared distribution

Generalisation of Fischer test

https://en.wikipedia.org/wiki/Chi-squared_test

Hypothesis significance testing

R. Fischer - "Lady tasting tea" - 1935

Classic approach

Test up against alternative hypothesis

Prove by rejecting null-hypothesis
being responsible for the observation

p-value

Probability of seeing the
observed if the null
hypothesis is true...

Reject null-hypothesis if p-value <5%

Milk or tea first?



Dr. Muriel Bristol claimed she could...

The null hypothesis is that she has no
ability to distinguish the teas.

Dr. Muriel Bristol nailed all of them...

4 successes has 1 chance
out of 70 ($\approx 1.4\% < 5\%$)

= p-value

Hypothesis significance testing

Power Analysis - how many samples?

Sample size to detect a significant difference?

$$n = \frac{(Z_\alpha + Z_\beta)^2 \cdot (p_1 \cdot [1 - p_1] + p_2 \cdot [1 - p_2])}{(p_2 - p_1)^2}$$

/
have to assume a certain lift - what if we are wrong in our assumptions?

Significance	$Z_\alpha = 1.96$	Reliability of 5%
	$Z_\beta = 0.84$	Power of 20% (80% to reject H_0 if false)
Expected CTR values	$p_1 = 0.1$	CTR of 10% for A - expected...
	$p_2 = 0.11$	CTR of 11% for B - expected...
Sample size	$n = 14.752$	Samples required

Single-sided or double sided test?



What are we testing for?

$$H_1 \quad \text{A} \quad < \quad \text{B}$$

$$H_1 \quad \text{A} \quad <> \quad \text{B}$$

Hypothesis testing challenges



General

can't really interpret probabilities of belief

if p-value > 5%,
then whole campaign is thrown out

blackbox to most....
tells nothing of the difference between A and B



Assumptions

data distribution?
Gaussian, t-student? etc.

one-sided or double sided?

a lot of assumptions - MUST be correct



Sample Size

hypothesis testing requires to estimate sample size first

must guess the performance of each variant. Not easy for B!

relies on estimating the significance when samples is reached

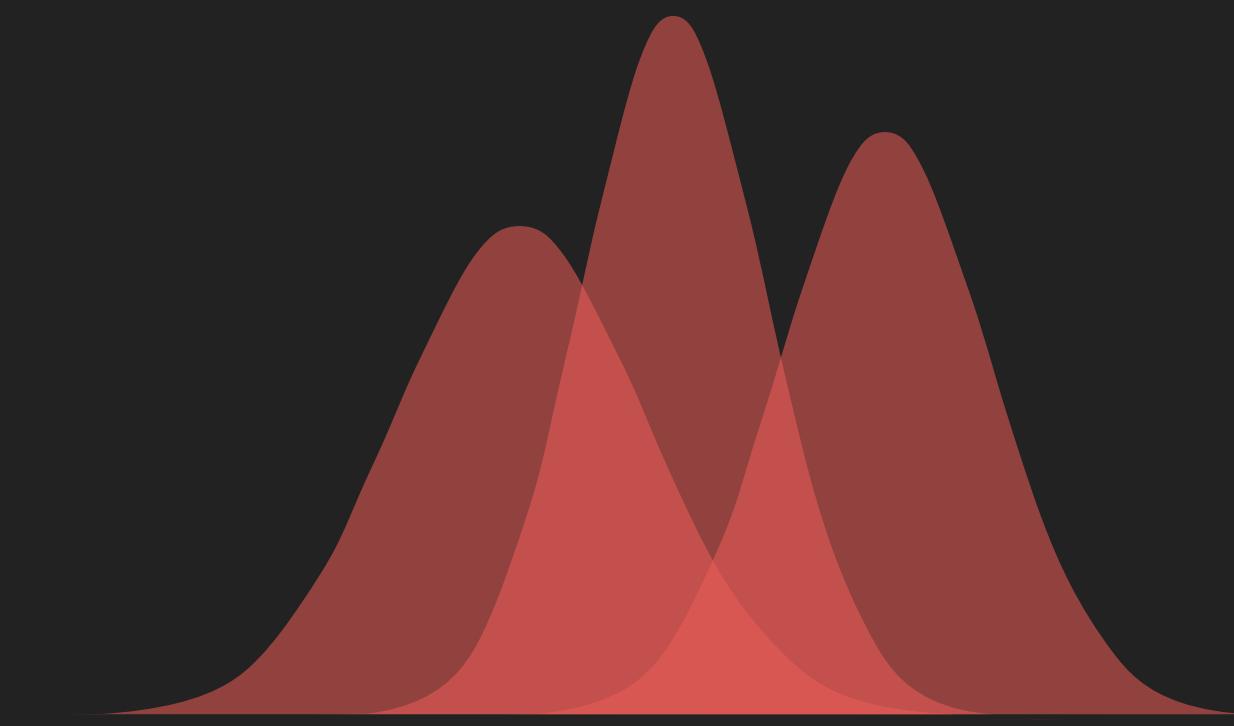


Peeking

"...checking A/B results before the test is over..."

very tempting...

significance assessment only occurs when samples size is reached...



Bayesian A/B testing

Types of Uncertainty

aleatoric uncertainty

natural stochasticity in observations (noise)

can't be reduced with more data

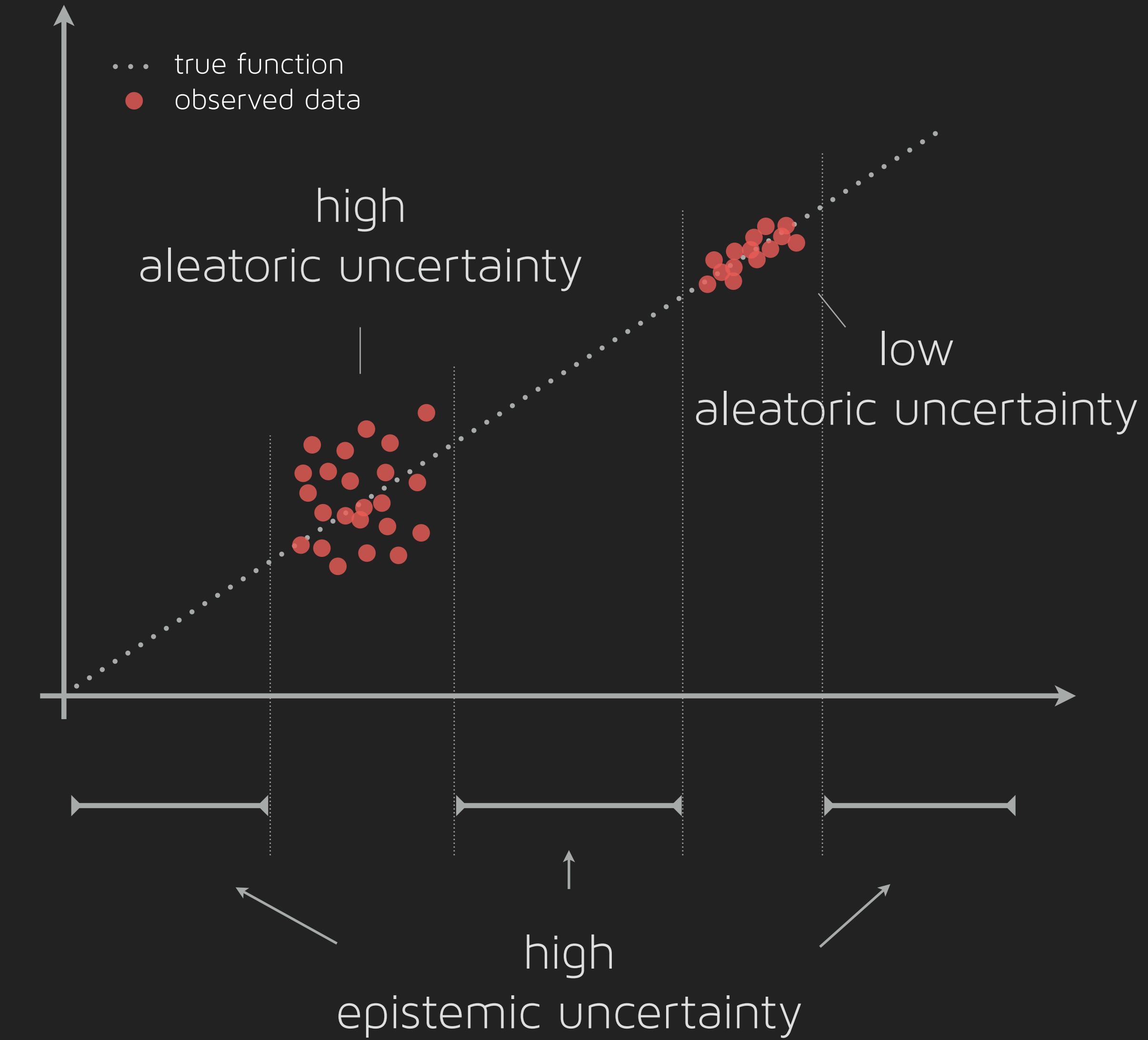
epistemic uncertainty

uncertainty around too few data points

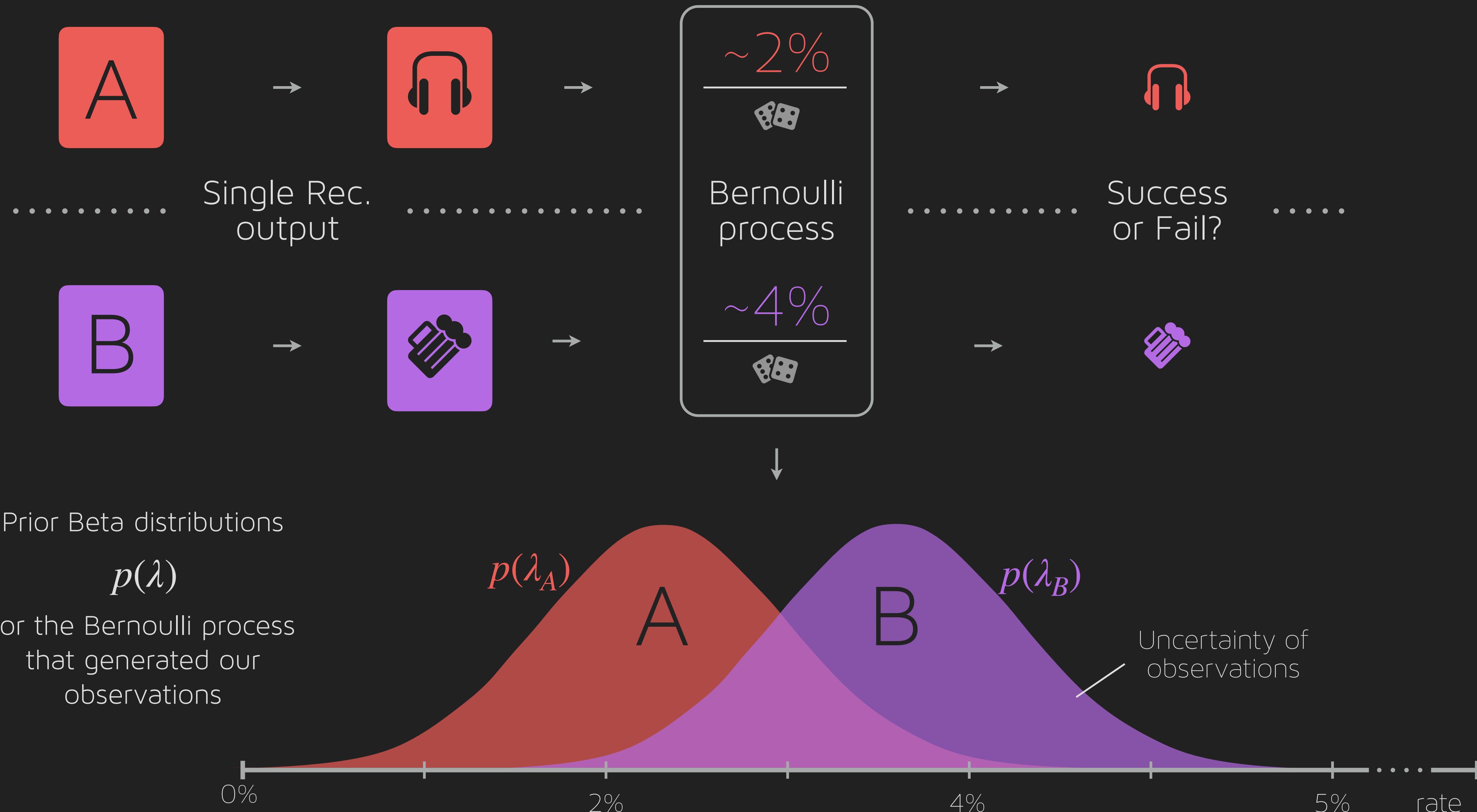
out-of-distribution outliers

two main types

example



Stochastic process behind the scenes...



Bayesian stochastic modeling / testing

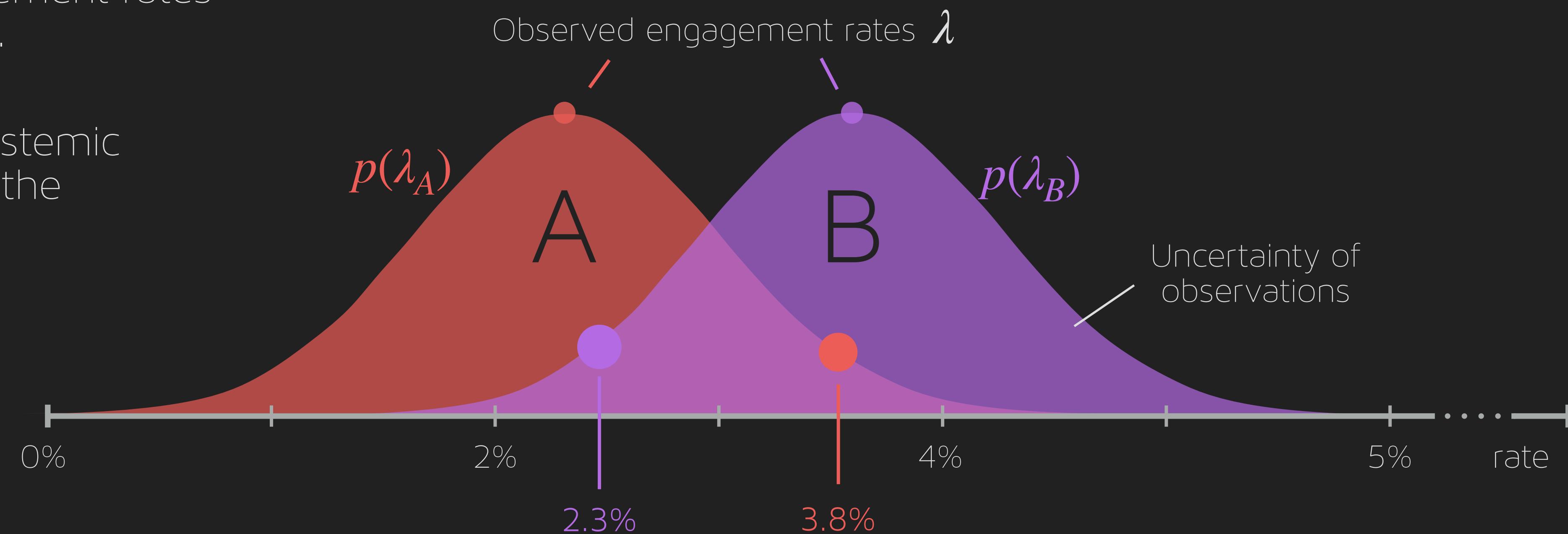
Model events as stochastic Bernoulli engagement events.

The engagement rate for each candidate can then be modelled as a probability distribution to capture the uncertainty.

Example of two candidates and their engagement rates with uncertainty.

Captures the epistemic uncertainty, not the aleatoric in this case.

The uncertainty allows us to quantify how much do we believe B is in fact the winner ...and evaluate the likelihood of many other scenarios.



Example: The unlikely case, where the true engagement rate dictates A is better than B.
These probabilities can be quantified....

Bayesian stochastic modeling / testing

Practical probability distributions to model engagement

clicks / impressions

beta distribution

$$p(x|\alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

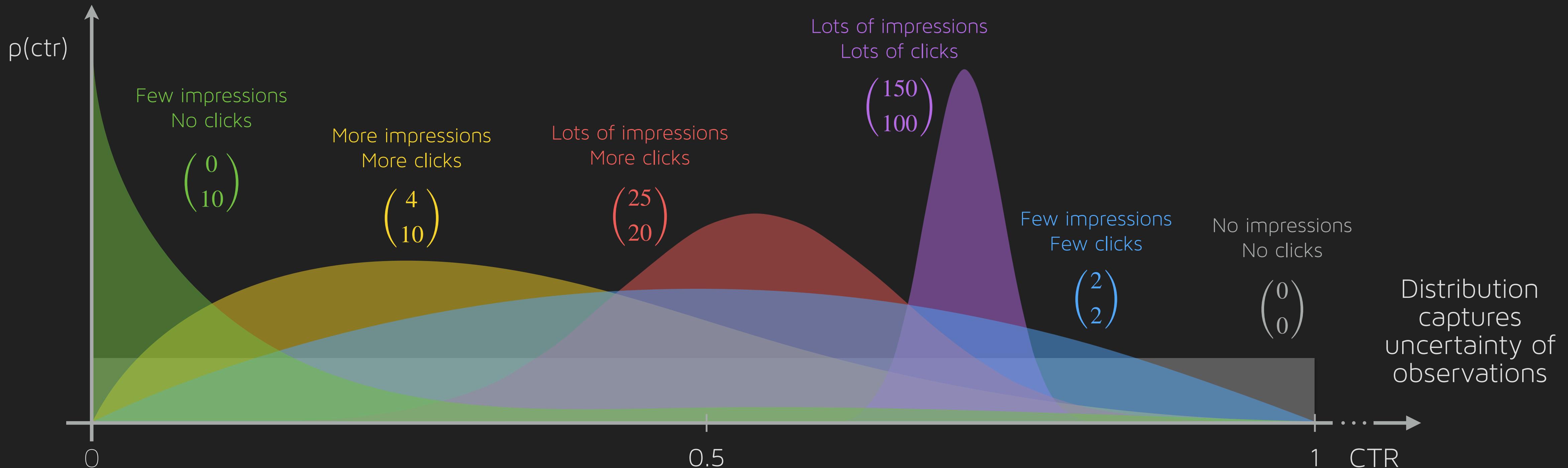
$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \text{clicks} \\ \text{impressions} - \text{clicks} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

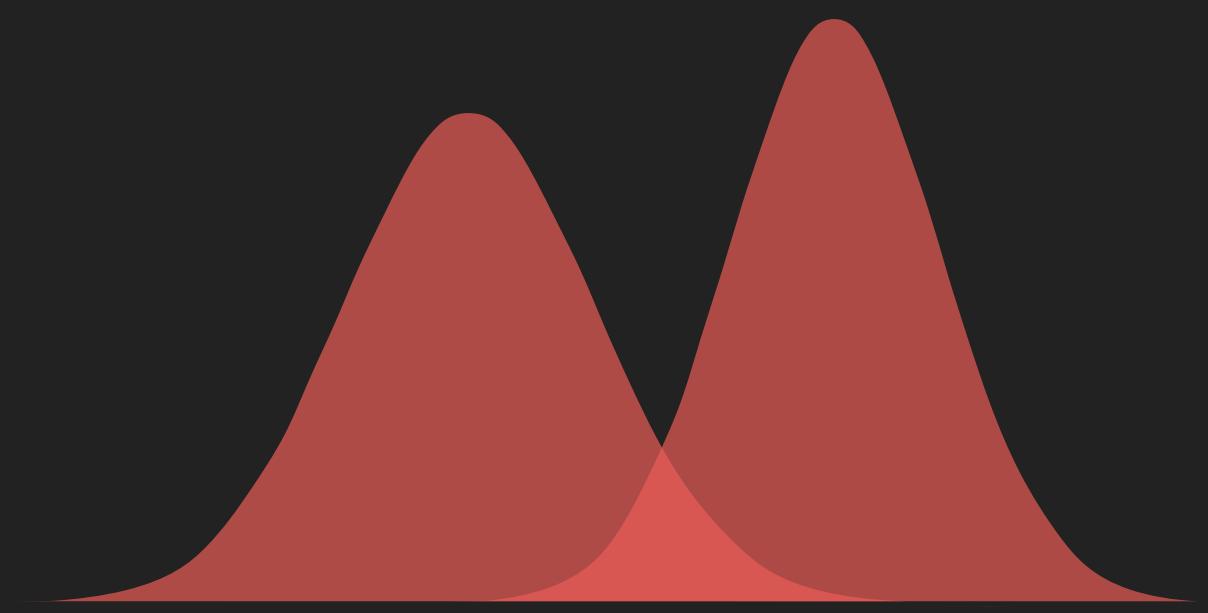
cost / click

gamma distribution

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \text{cost} \\ \text{clicks}^{-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$





Bayesian Metrics
A/A testing

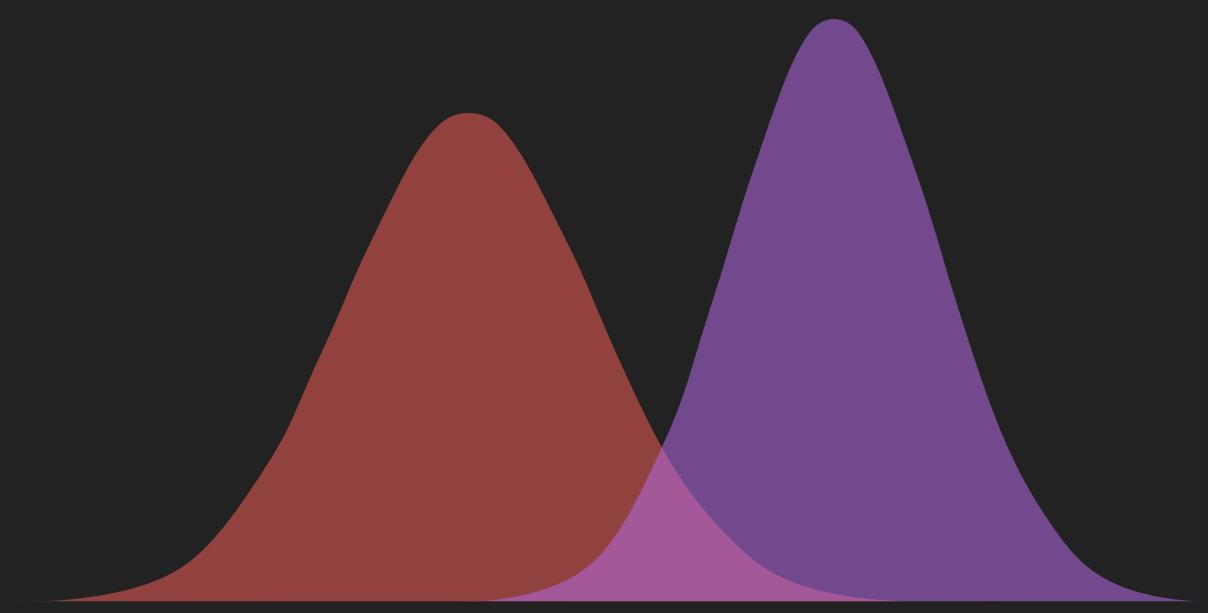
Bayesian same/same test

Selected metrics for measuring differences
between two probability distributions

Are A/A the same
distribution?



Example: A1 and A2 are close enough to each other....or what?



Bayesian Metrics
A/B testing

Bayesian Metrics - Probability of winner

Bayesian testing by calculating

the probability of selecting the right winner

ie. confidence score / certainty

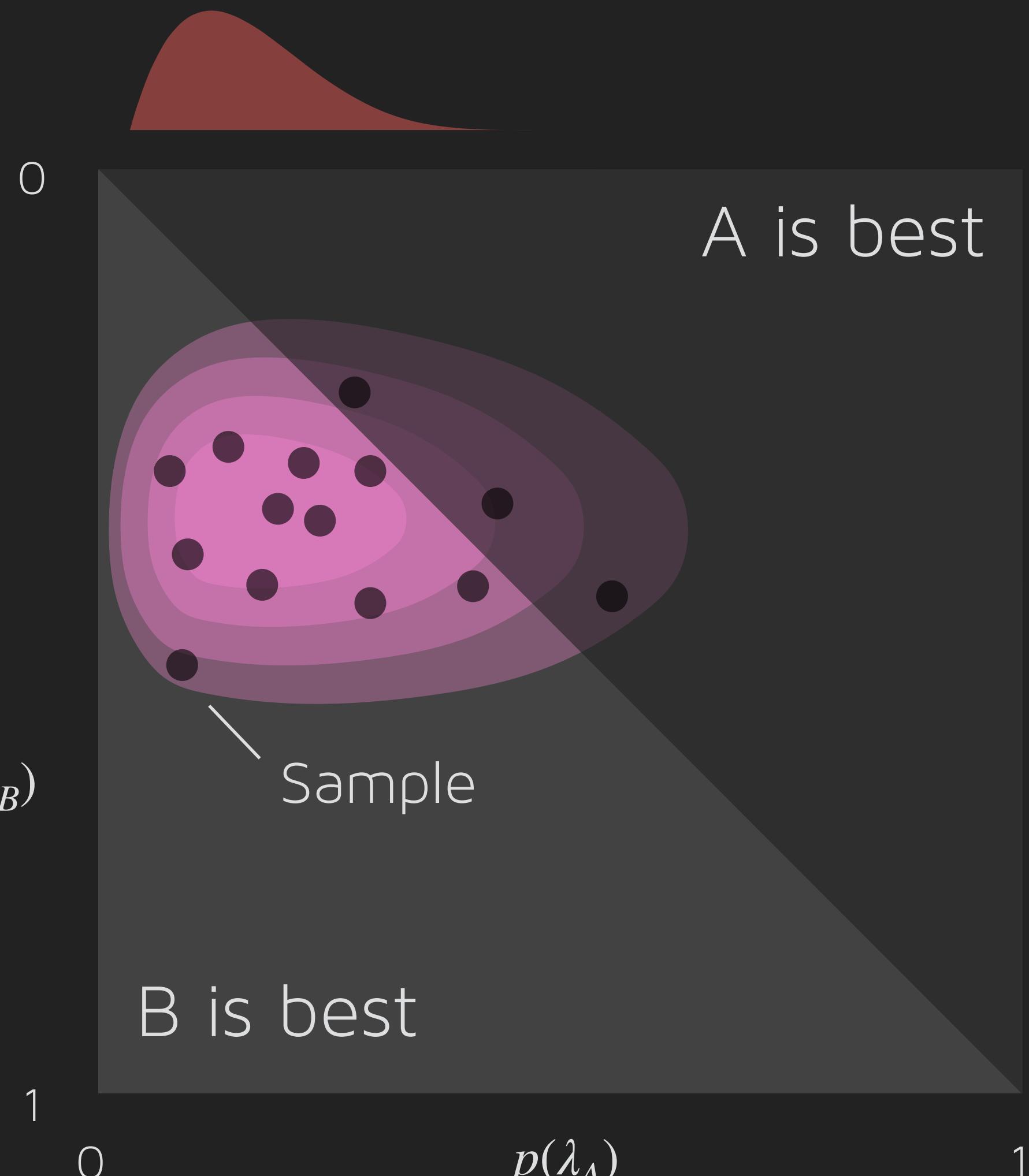
This can be deducted from the joint probability
...and corresponds to the area, where $p_{\lambda_B} > p_{\lambda_A}$

This marginalisation can be estimated via
Monte Carlo sampling

....or calculated analytically from for the
Beta distribution

$$\begin{aligned} P(p_B > p_A) &= \int_0^1 \int_0^{\lambda_A} P(\lambda_A, \lambda_B) d\lambda_B d\lambda_A \\ &= \sum_{i=0}^{\alpha_B-1} \frac{B(\alpha_A + i, \beta_A + \beta_B)}{(\beta_B + i)B(1 + i, \beta_B)B(\alpha_A, \beta_A)} \end{aligned}$$

Joint probability of engagement-rates



Bayesian Metrics - Expected Loss

Average performance loss for wrong choice

Evaluation of loss if choosing wrong

$$\mathcal{L}(\lambda_A, \lambda_B, A) = \max(\lambda_B - \lambda_A, 0)$$

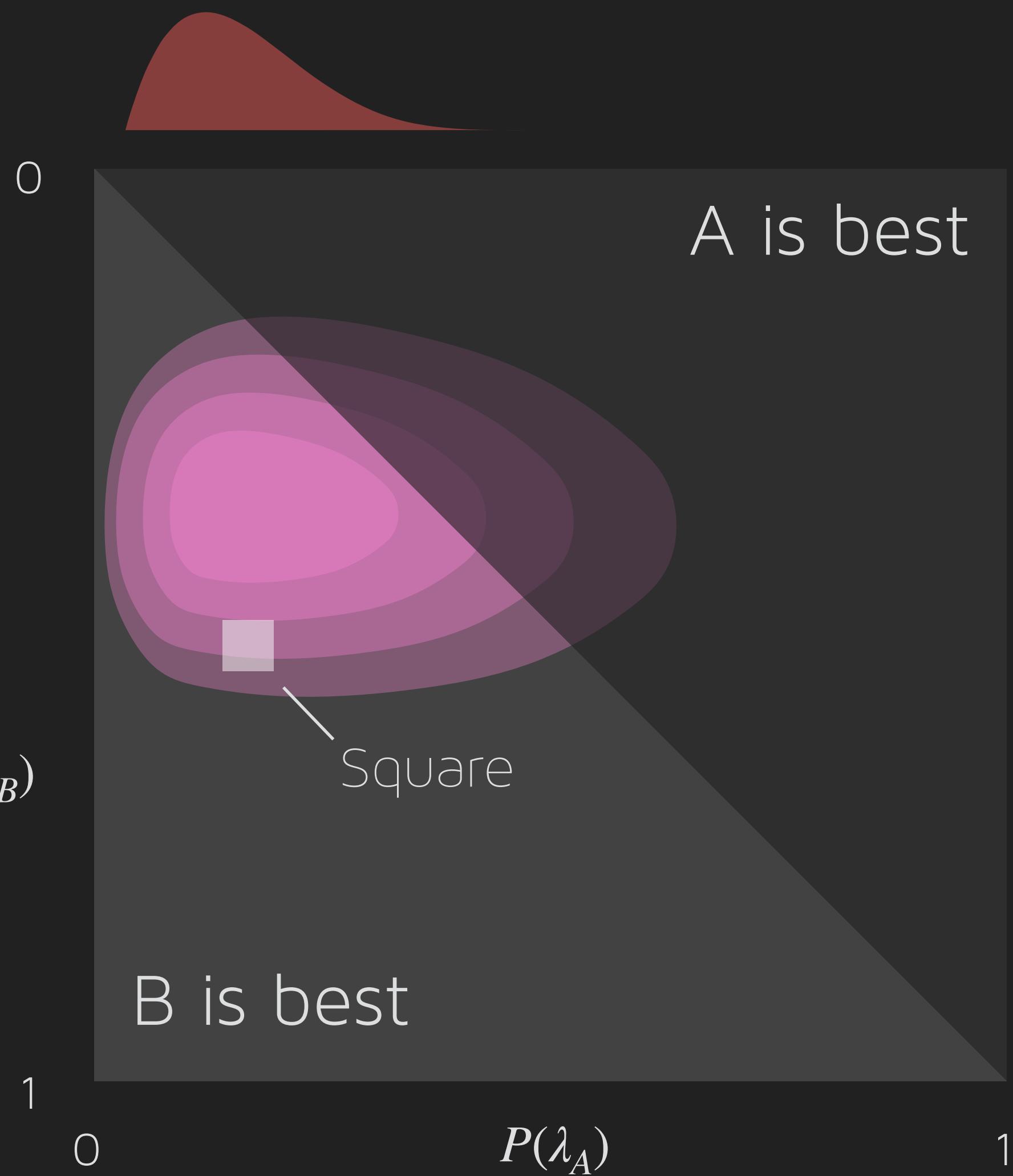
$$\mathcal{L}(\lambda_A, \lambda_B, B) = \max(\lambda_A - \lambda_B, 0)$$

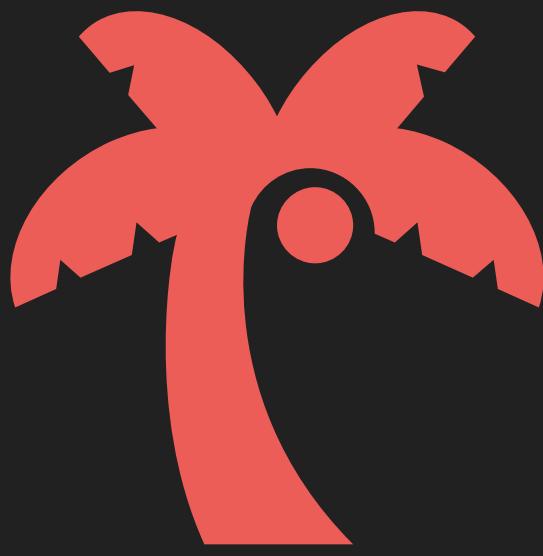
Expectation of loss by integrating across joint distribution

$$E[\mathcal{L}](\lambda_A, \lambda_B) = \int_0^1 \int_0^1 P(\lambda_A, \lambda_B) \mathcal{L}(\lambda_A, \lambda_B, ?) d\lambda_B d\lambda_A$$

Unit will be λ - in some cases close to \$, as we shall see...

Joint probability of engagement-rates





Single Example

Example....

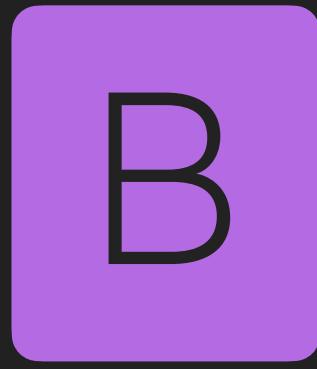
Observations



$\frac{1000}{\text{impressions}}$

$\frac{100}{\text{clicks}}$

$\frac{\$11}{\text{Cost}}$



$\frac{1000}{\text{impressions}}$

$\frac{120}{\text{clicks}}$

$\frac{\$19}{\text{Cost}}$

Hypothesis testing - p-values....

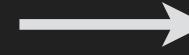
Let's see what the numbers are after 1000 impressions...

Example

Impressions	1000
Clicks A	100
Clicks B	120

$$\lambda_A = \frac{100}{1000} = 10\%$$

$$\lambda_B = \frac{120}{1000} = 12\%$$



B
winner ?

1000
—
#samples

12%
—
p-value

Uuh, this is not good...
...Peeking

3840
—
#samples
actually required

0.2%
—
p-value

Aaaah, much better....

Bayesian Metrics - Probability of winner & lift analysis

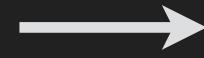
The uncertainty allows for reasoning and exploring the probability of various scenarios.

Example

Impressions
Clicks A
Clicks B

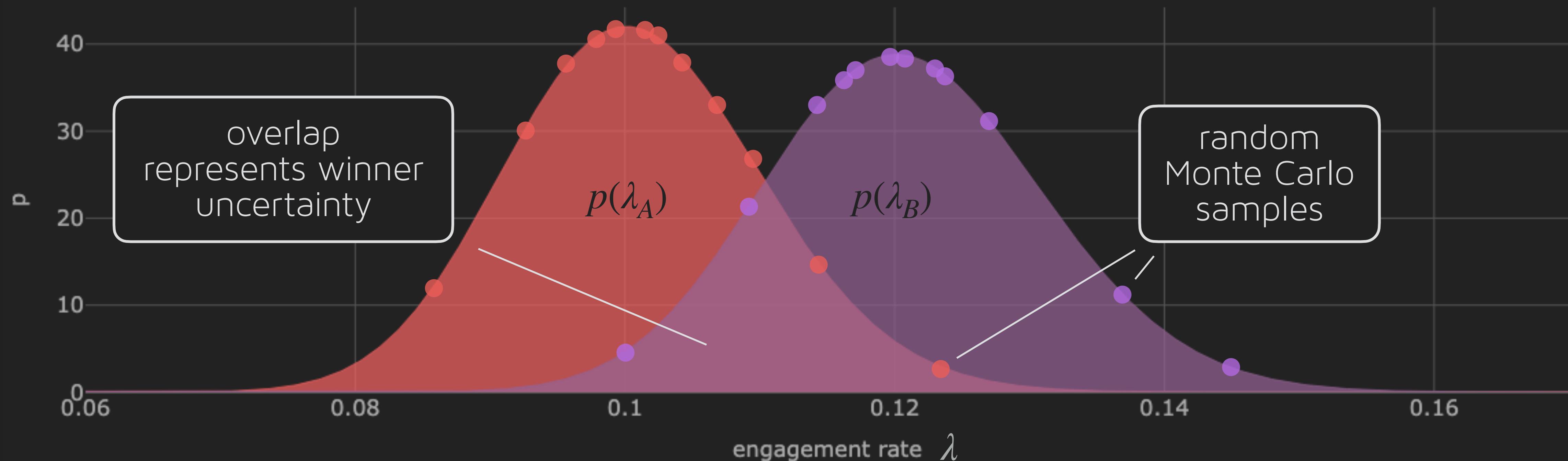
1000
100
120

$$\lambda_A = \frac{100}{1000} = 10\%$$
$$\lambda_B = \frac{120}{1000} = 12\%$$



B
winner ?

Beta distributions of CTR



Bayesian Metrics - Probability of winner & lift analysis

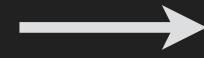
The uncertainty allows for reasoning and exploring the probability of various scenarios.

Example

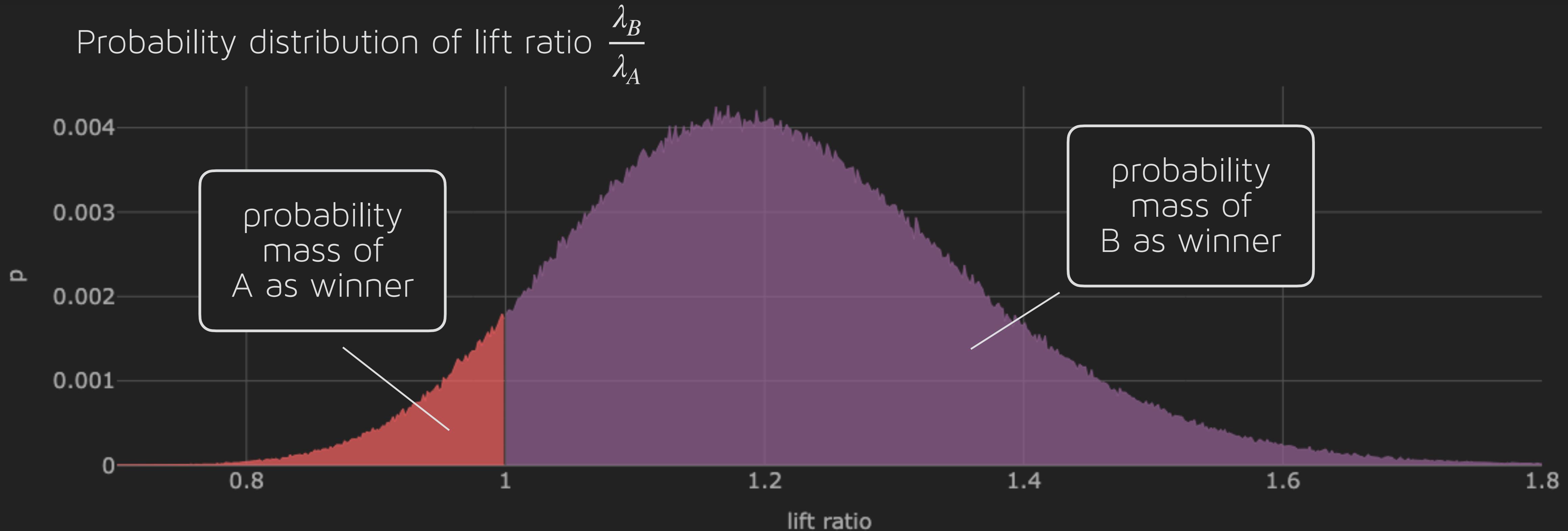
Impressions
Clicks A
Clicks B

1000
100
120

$$\lambda_A = \frac{100}{1000} = 10\%$$
$$\lambda_B = \frac{120}{1000} = 12\%$$



B
winner ?



Bayesian Metrics - Probability of winner & lift analysis

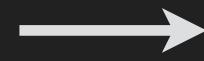
The uncertainty allows for reasoning and exploring the probability of various scenarios.

Example

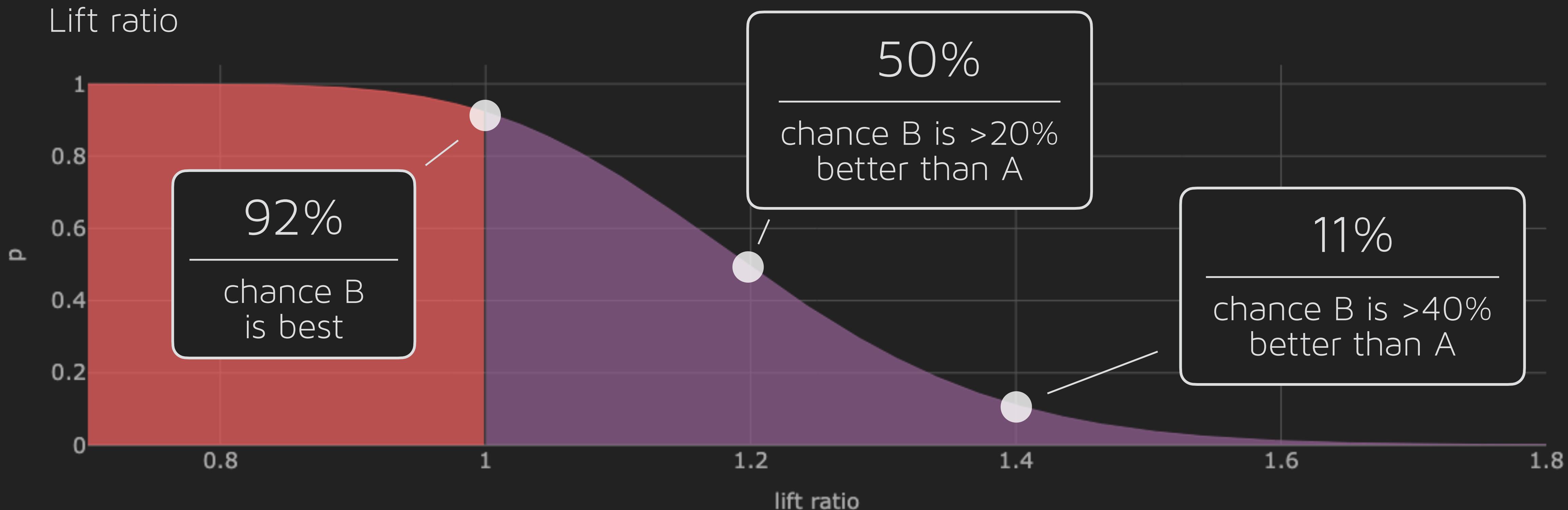
Impressions
Clicks A
Clicks B

1000
100
120

$$\lambda_A = \frac{100}{1000} = 10\%$$
$$\lambda_B = \frac{120}{1000} = 12\%$$



B
winner ?



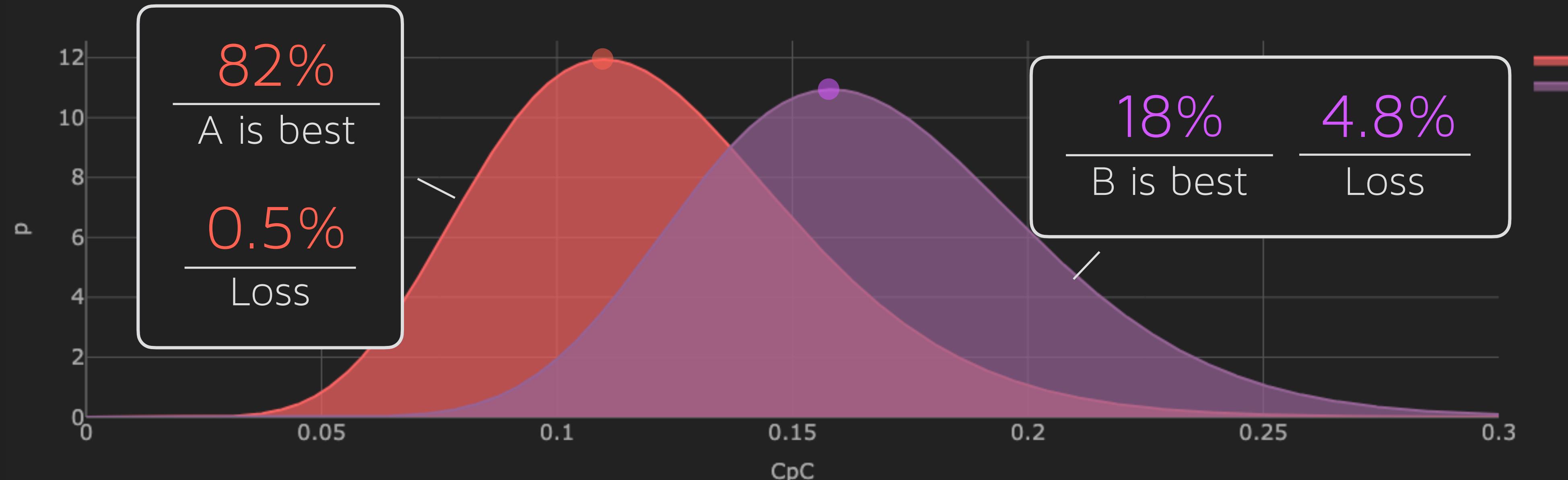
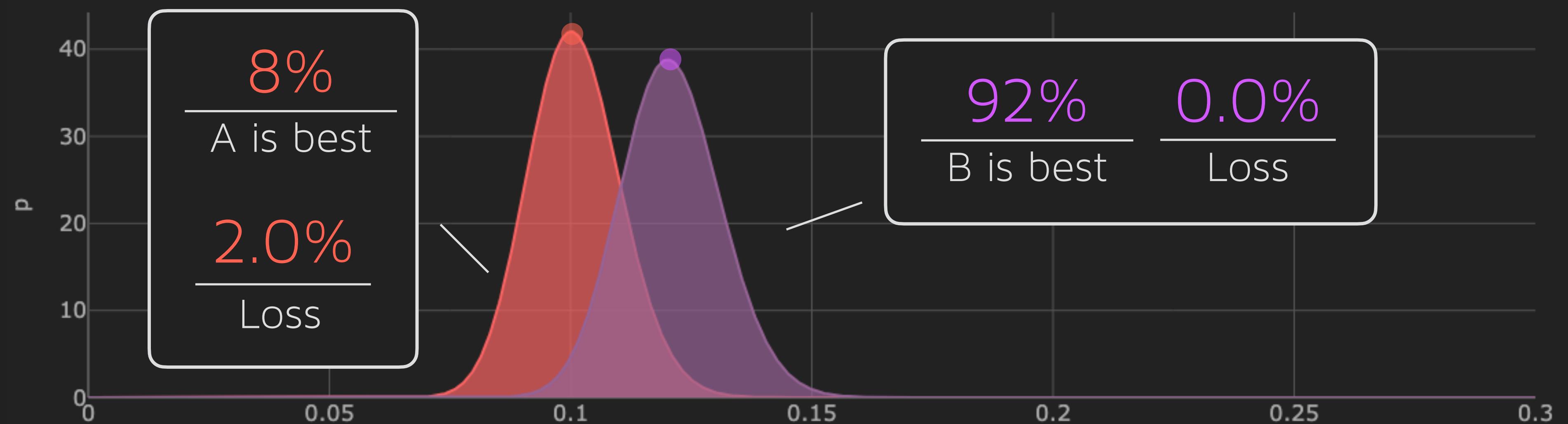
Bayesian A/B Testing - Decision

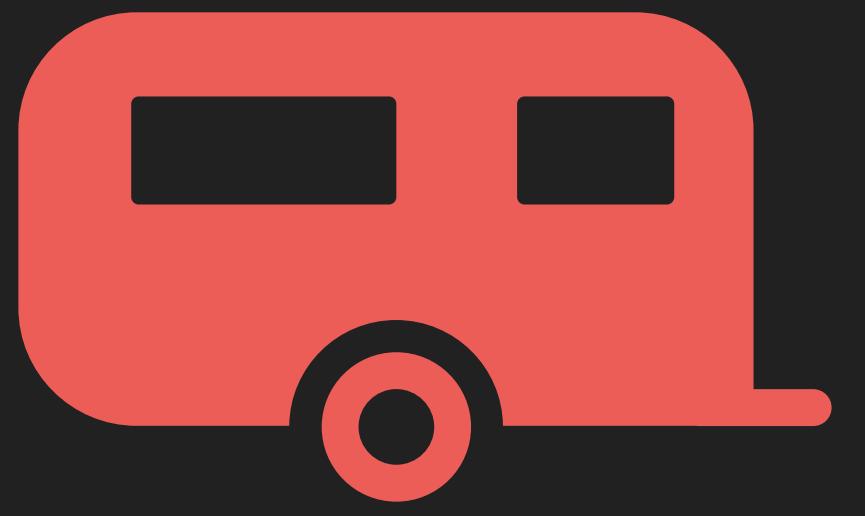
Click-
Through-
Rate

After 1000
impressions

Cost-per-
Click

Aleatoric
uncertainty
still lingers....

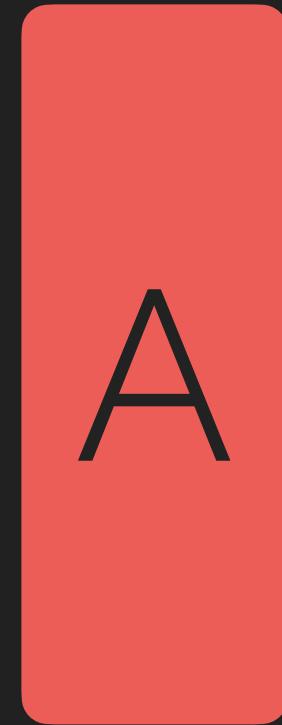




Campaign

A/B Testing - Campaign simulation

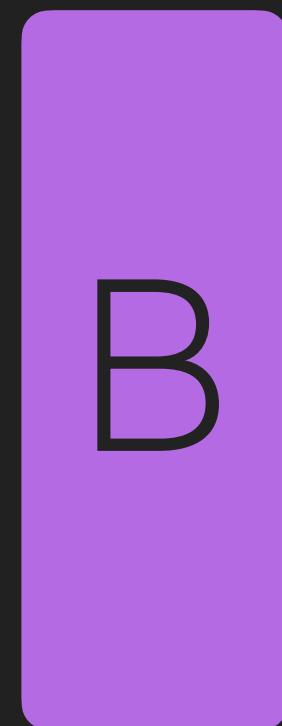
...



$\frac{10\%}{\text{CTR A}}$

\$11

$\frac{\$11}{\text{CPM A}}$

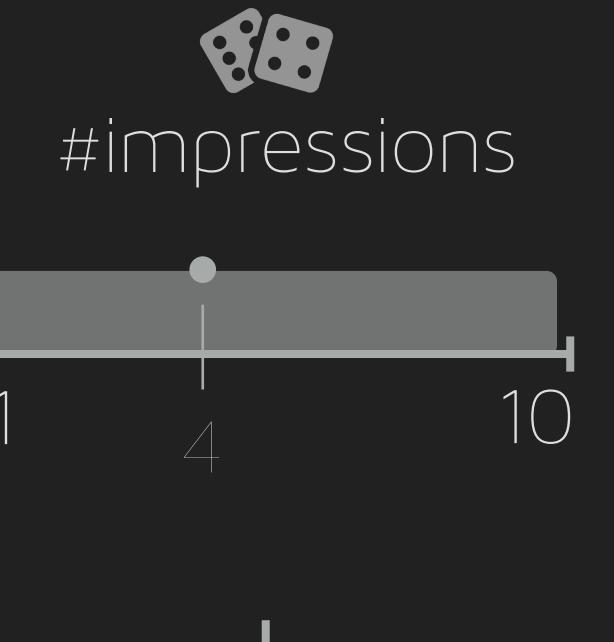


$\frac{12\%}{\text{CTR B}}$

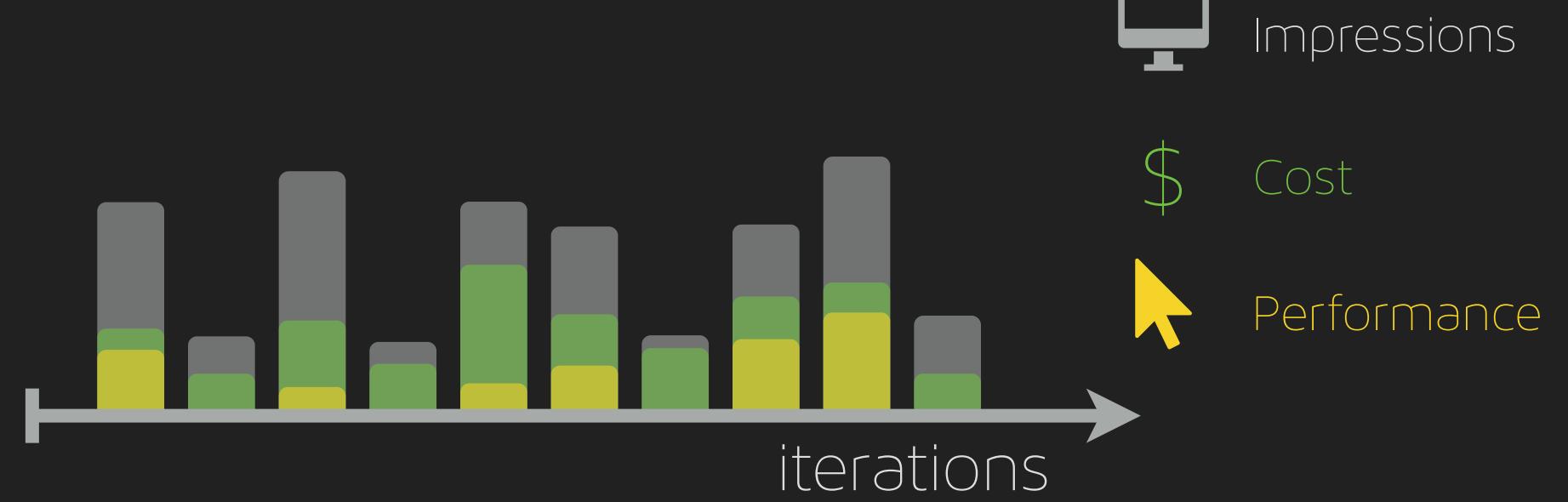
\$19

$\frac{\$19}{\text{CPM B}}$

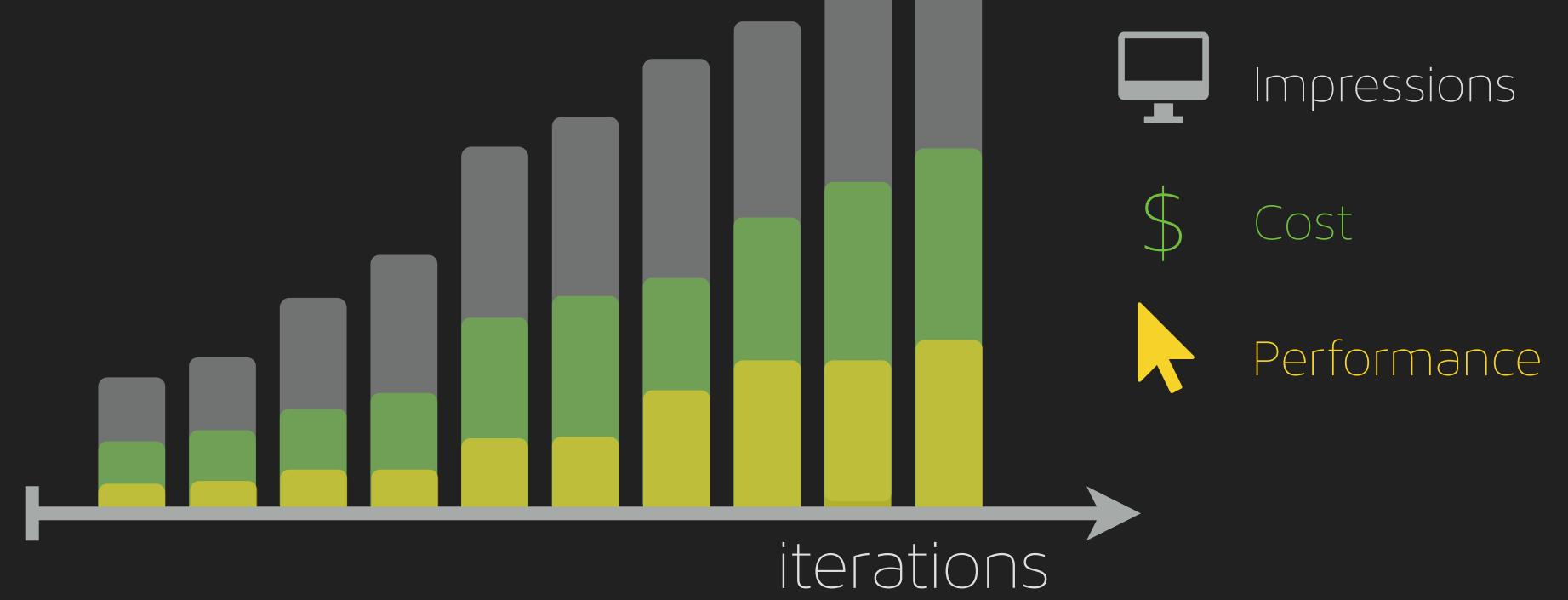
Simulate impression,
clicks and cost



logs



accumulated



Impressions

Cost

Performance

Impressions

Cost

Performance

Campaign simulation

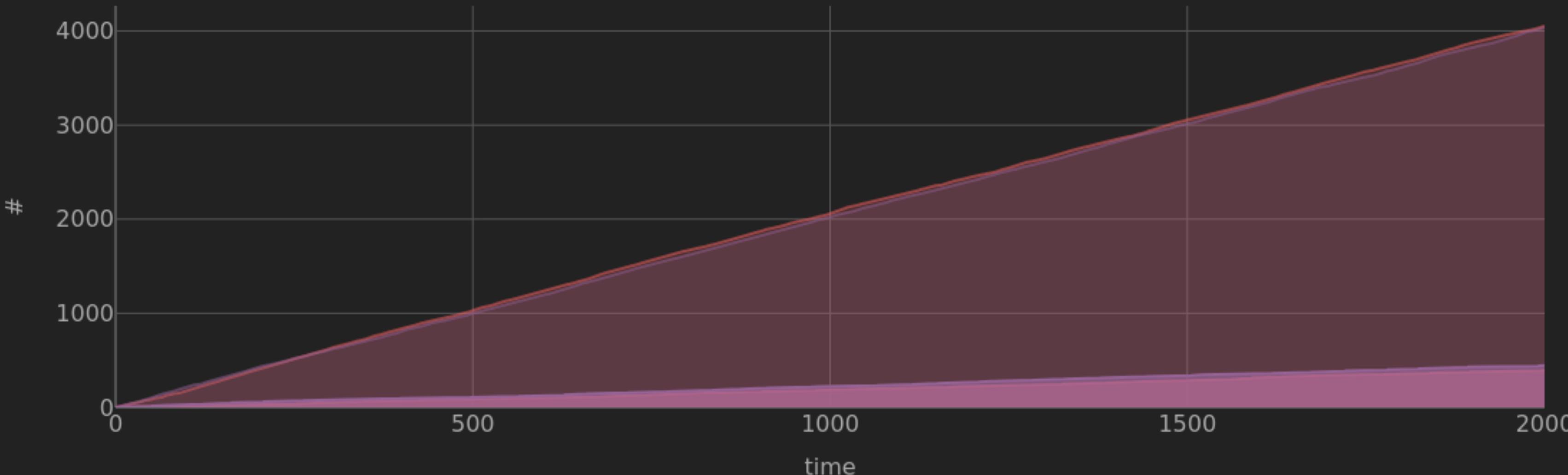
Observations and performance

A

10%
CTR A

\$11
CPM A

Observations - impr. & clicks

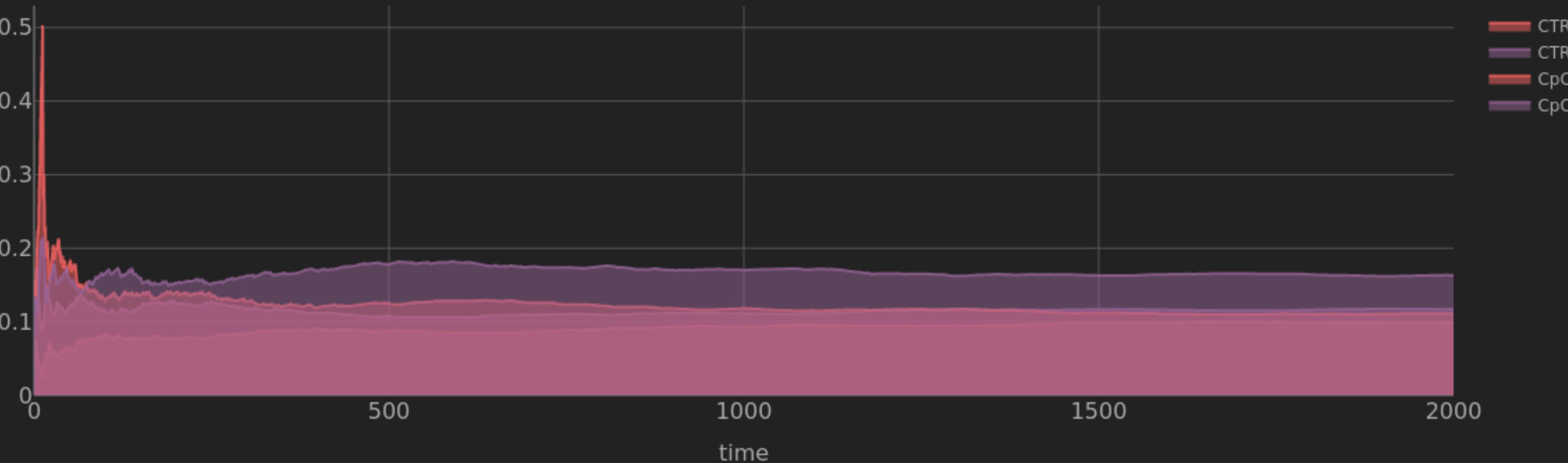


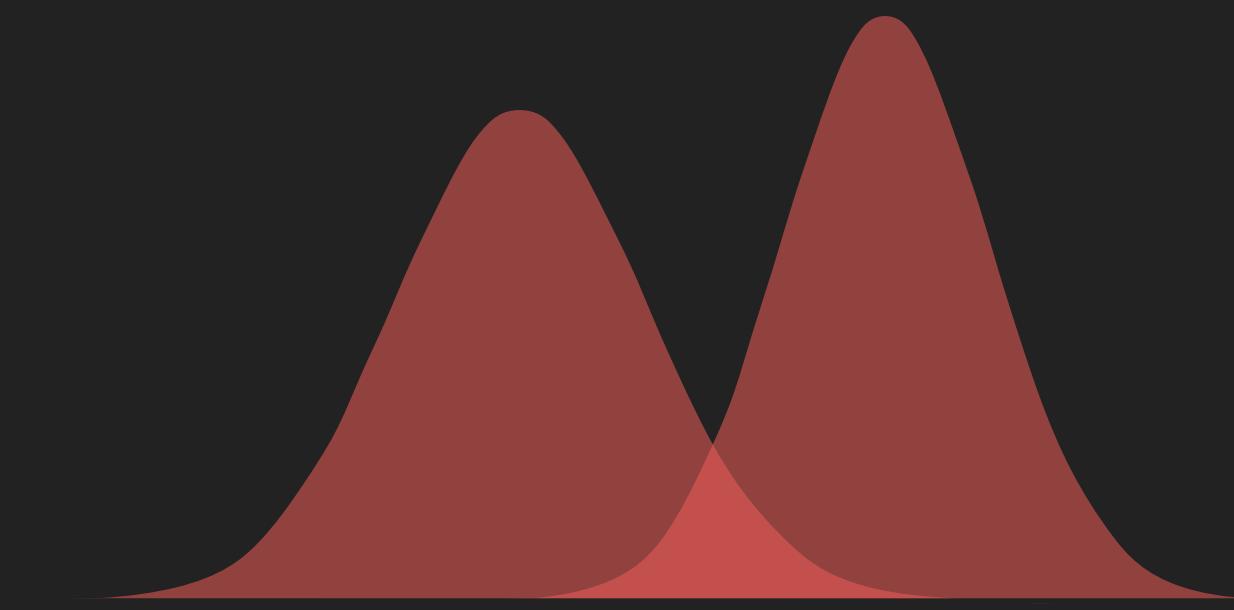
B

12%
CTR B

\$19
CPM B

Performance - CTR & CpC

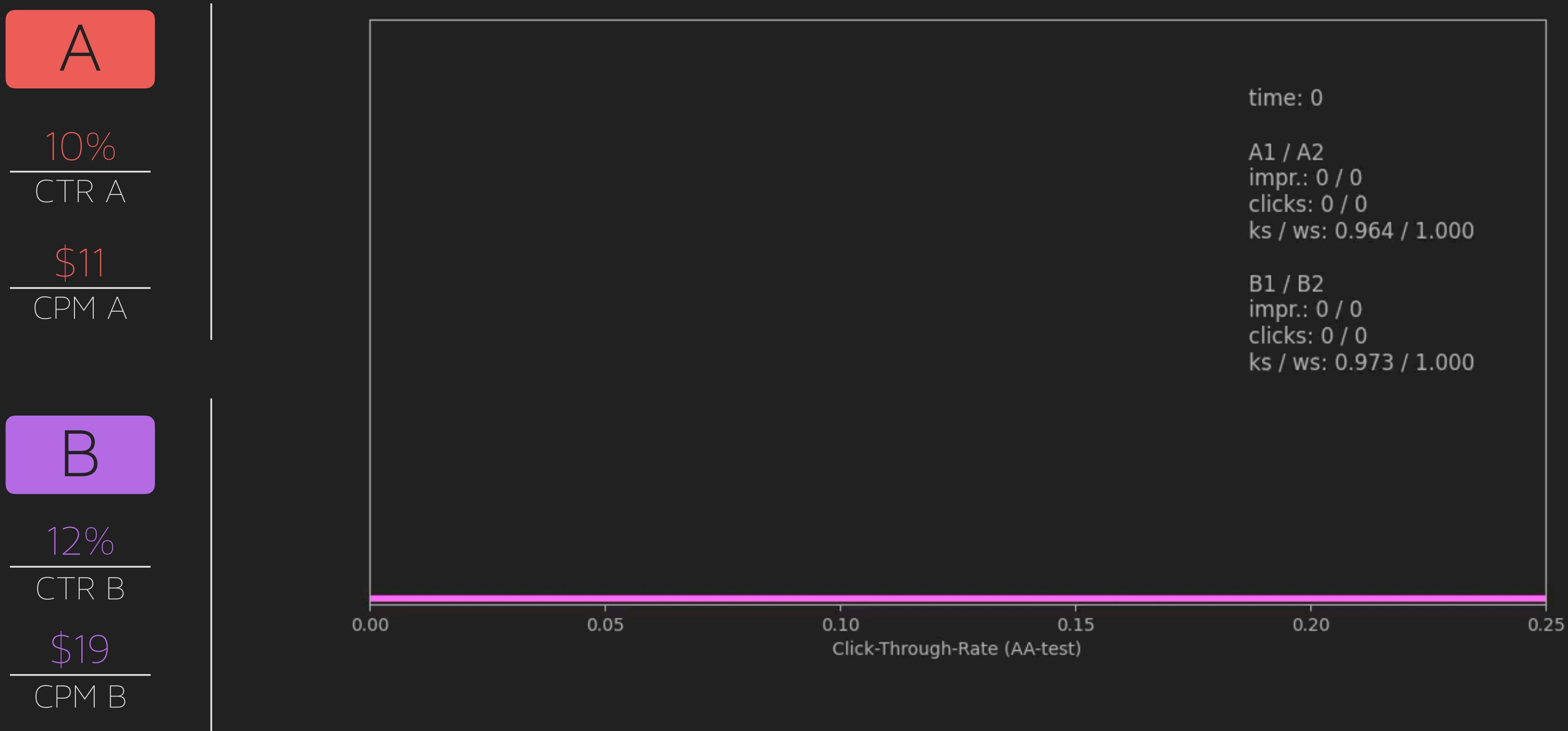




Bayesian A/A testing

Hypothesis A/A Testing - Campaign simulation

Same / Same test



Hypothesis A/A Testing - Campaign simulation

Same / Same test

A

10%
CTR A

\$11
CPM A

4054 / 4078

impressions

393 / 410

clicks

45 / 45

cost

B

12%
CTR B

\$19
CPM B

4040 / 4088

impressions

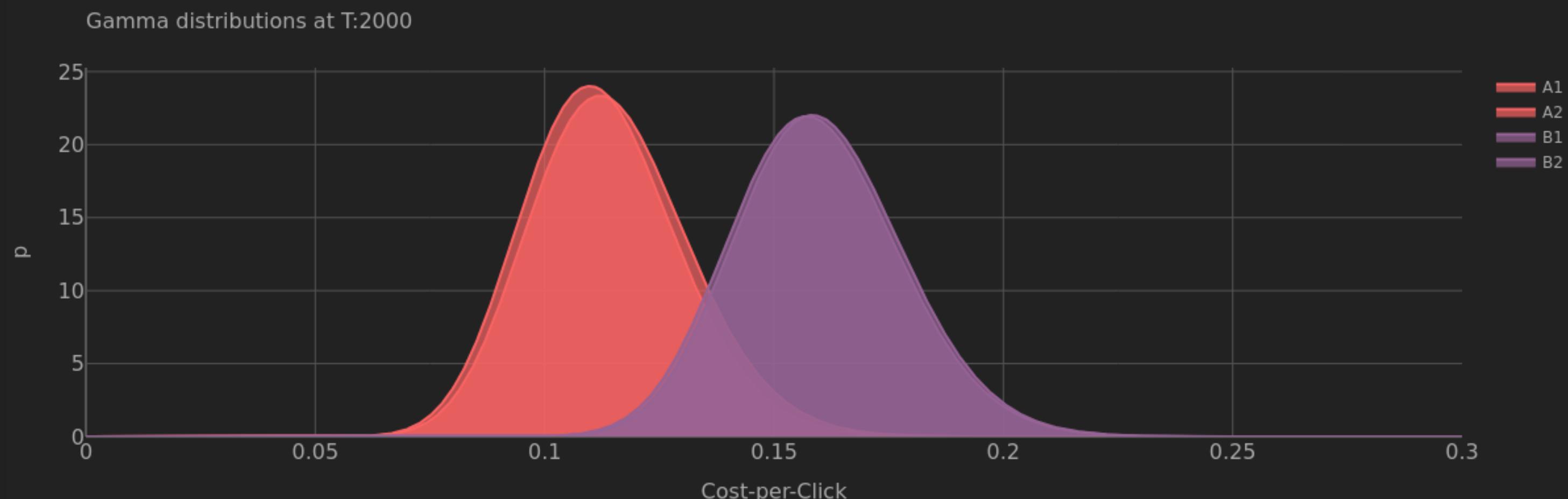
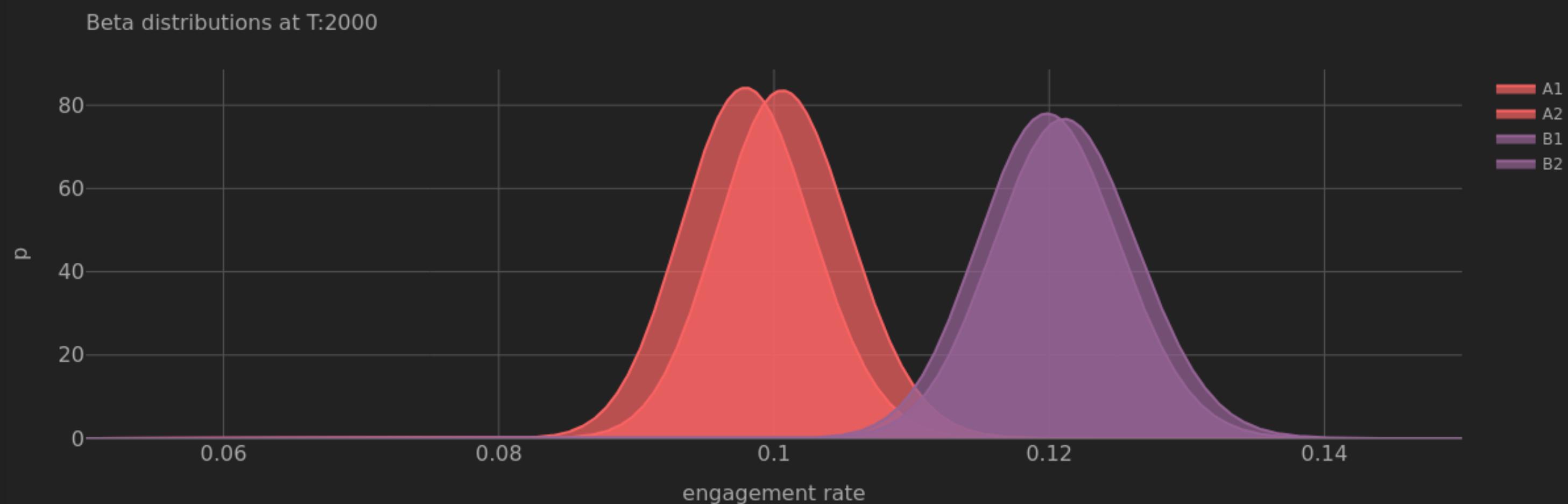
447 / 503

clicks

77 / 78

cost

End-of-simulation



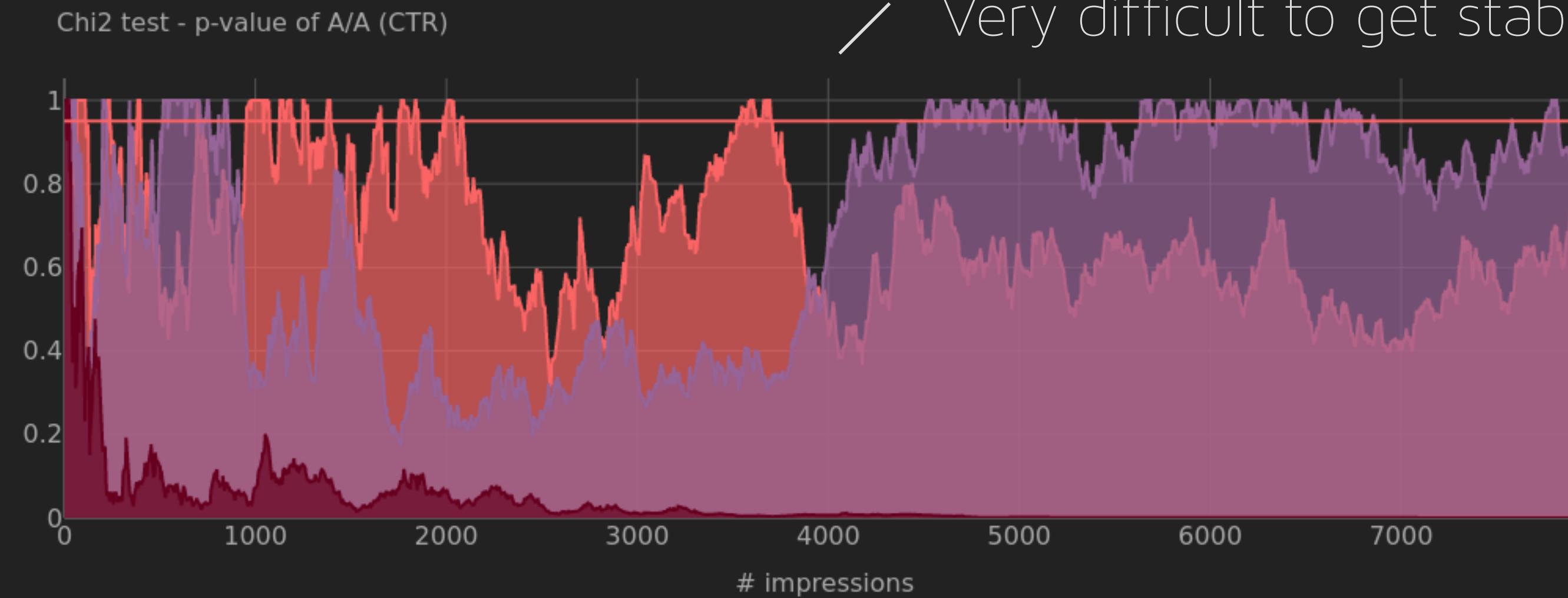
Hypothesis A/A Testing - Campaign simulation

Same / Same test

A

$\frac{10\%}{\text{CTR A}}$

$\frac{\$11}{\text{CPM A}}$



Aleatoric uncertainty
Very difficult to get stable...

A didn't make it...

$\frac{65\%}{\text{p-value}}$

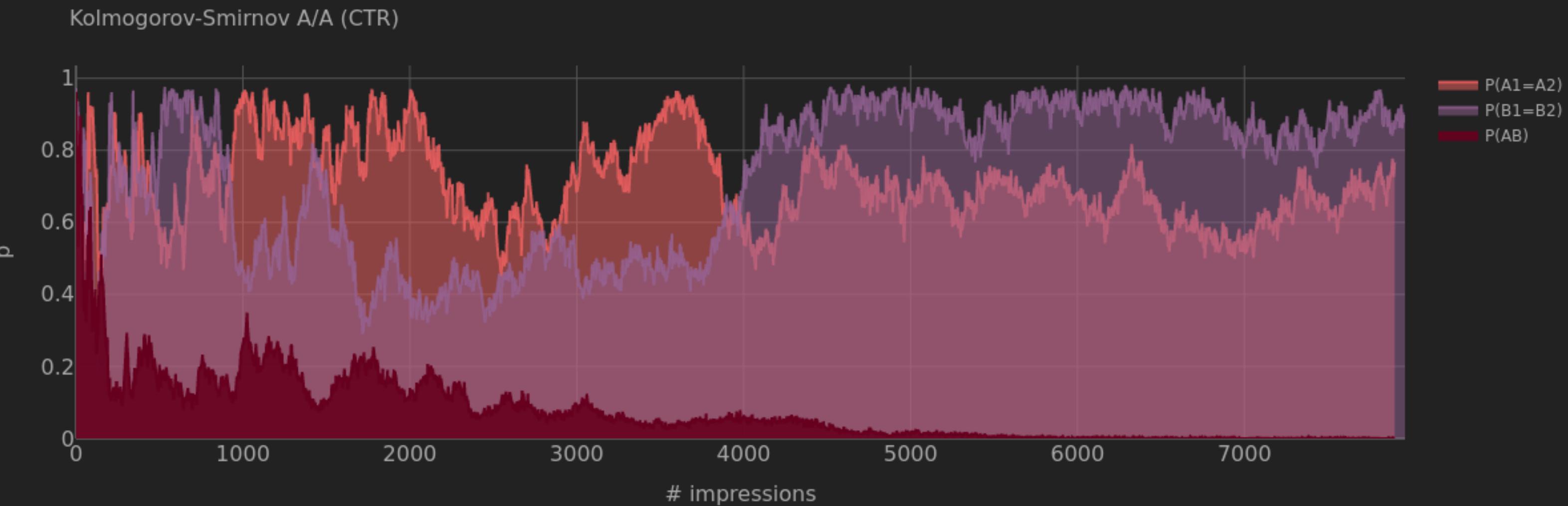
$\frac{97\%}{\text{p-value}}$

B

$\frac{12\%}{\text{CTR B}}$

$\frac{\$19}{\text{CPM B}}$

AA & BB are more similar than AB



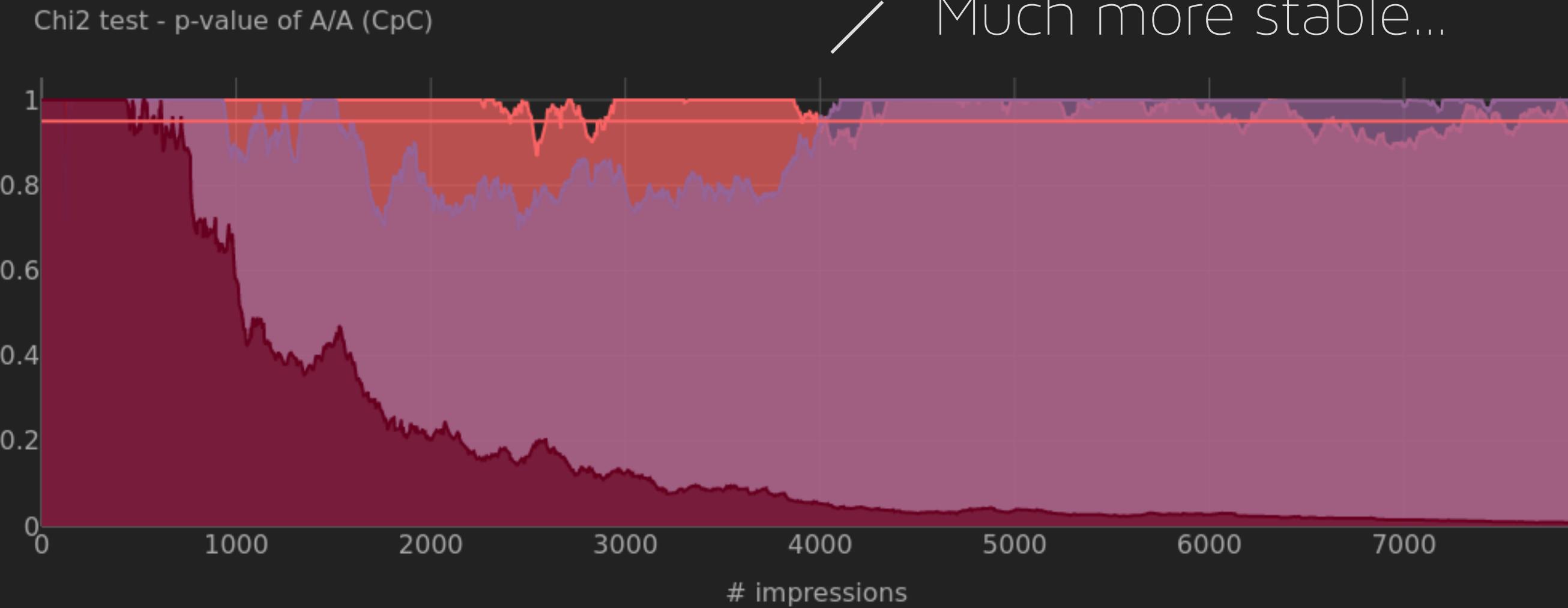
Hypothesis A/A Testing - Campaign simulation

Same / Same test

A

$\frac{10\%}{\text{CTR A}}$

$\frac{\$11}{\text{CPM A}}$



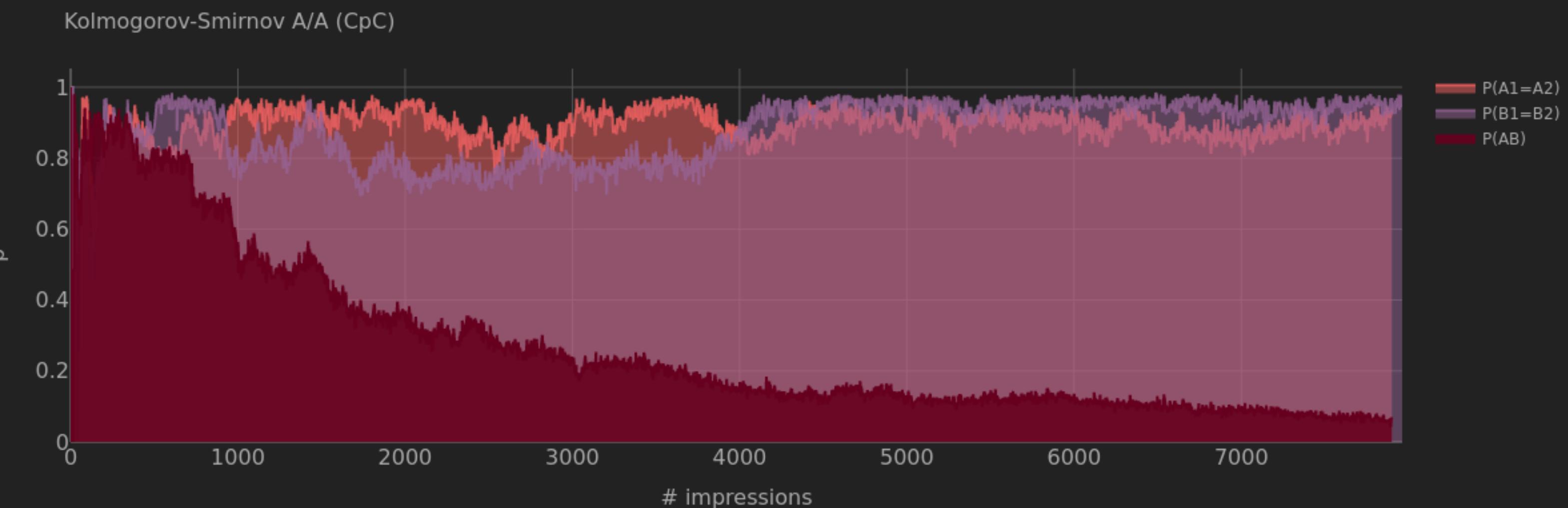
$\frac{98\%}{\text{p-value}}$ $\frac{100\%}{\text{p-value}}$

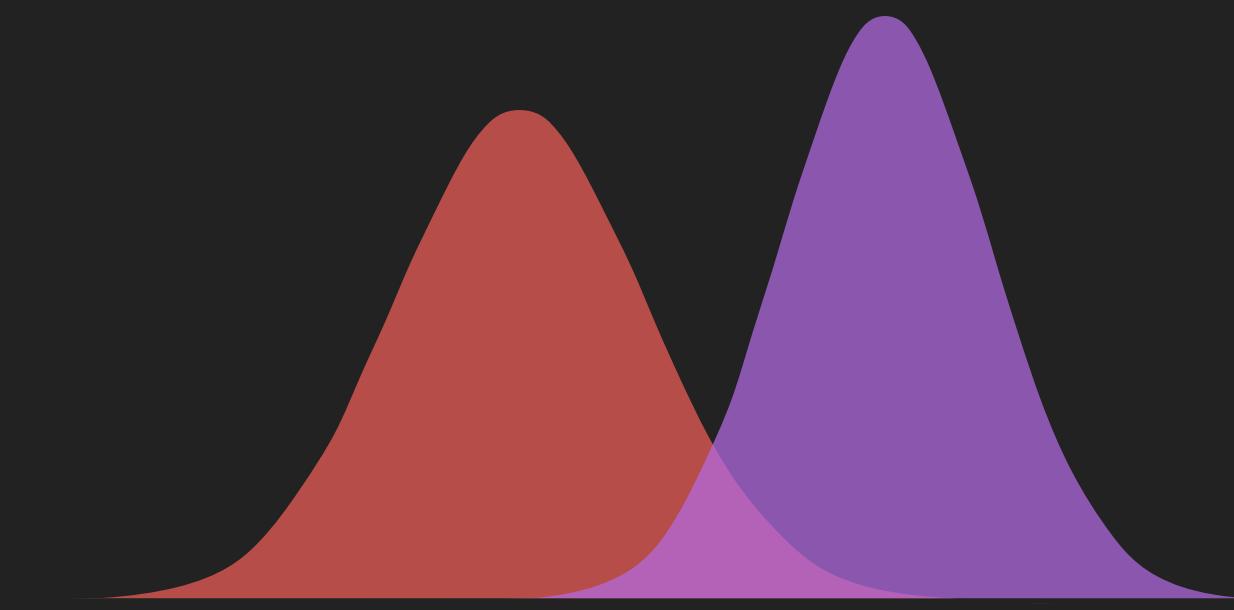
B

$\frac{12\%}{\text{CTR B}}$

$\frac{\$19}{\text{CPM B}}$

AA & BB are
more similar
than AB





Bayesian A/B testing

TODO: Bayesian A/B Testing - Campaign simulation

....

A

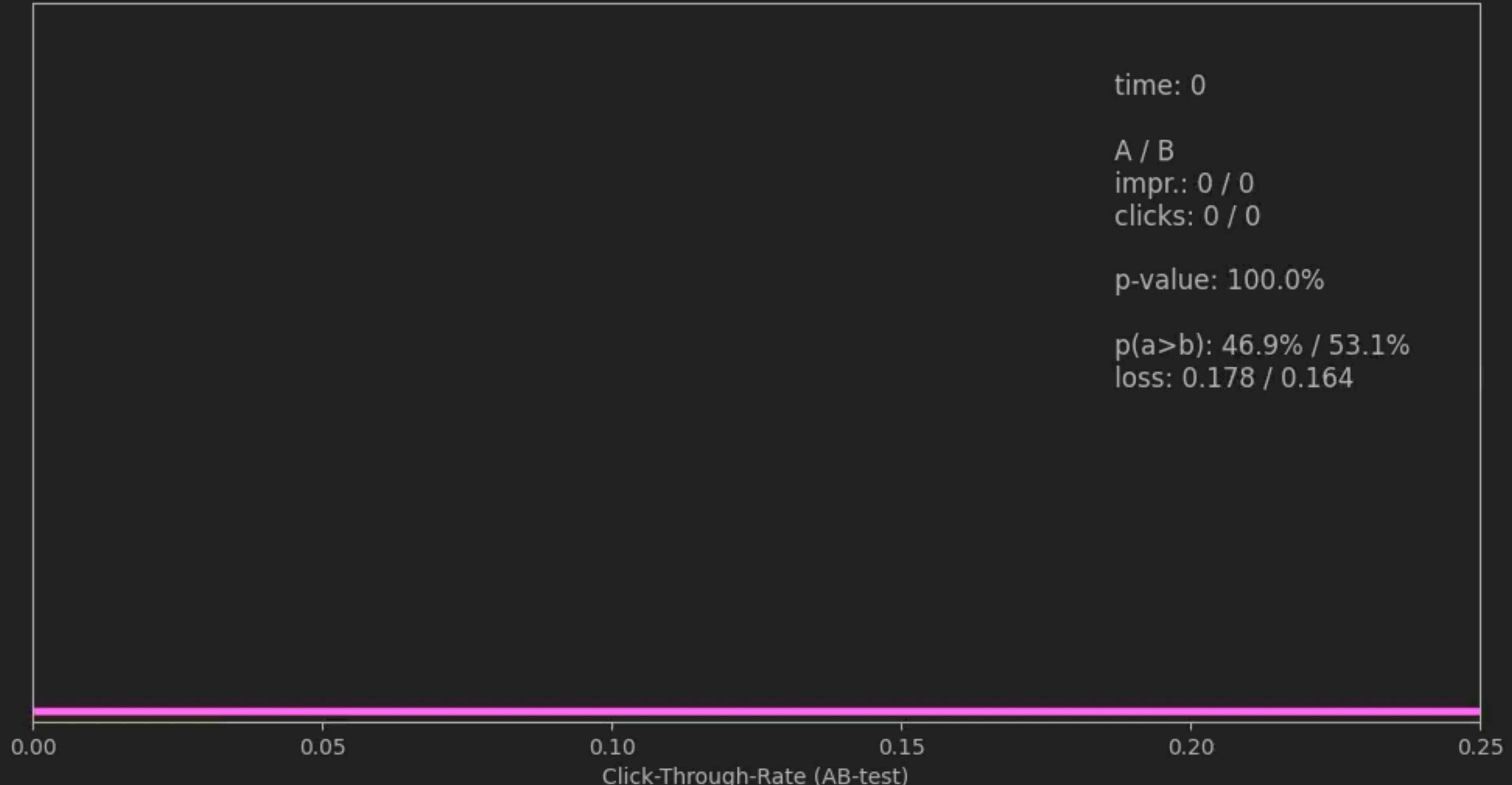
10%
—
CTR A

\$11
—
CPM A

B

12%
—
CTR B

\$19
—
CPM B



Hypothesis A/B Testing - Campaign simulation

A/B-test

A

10%
CTR A

\$11
CPM A

7898

impressions

784

clicks

87

cost

B

12%
CTR B

\$19
CPM B

7961

impressions

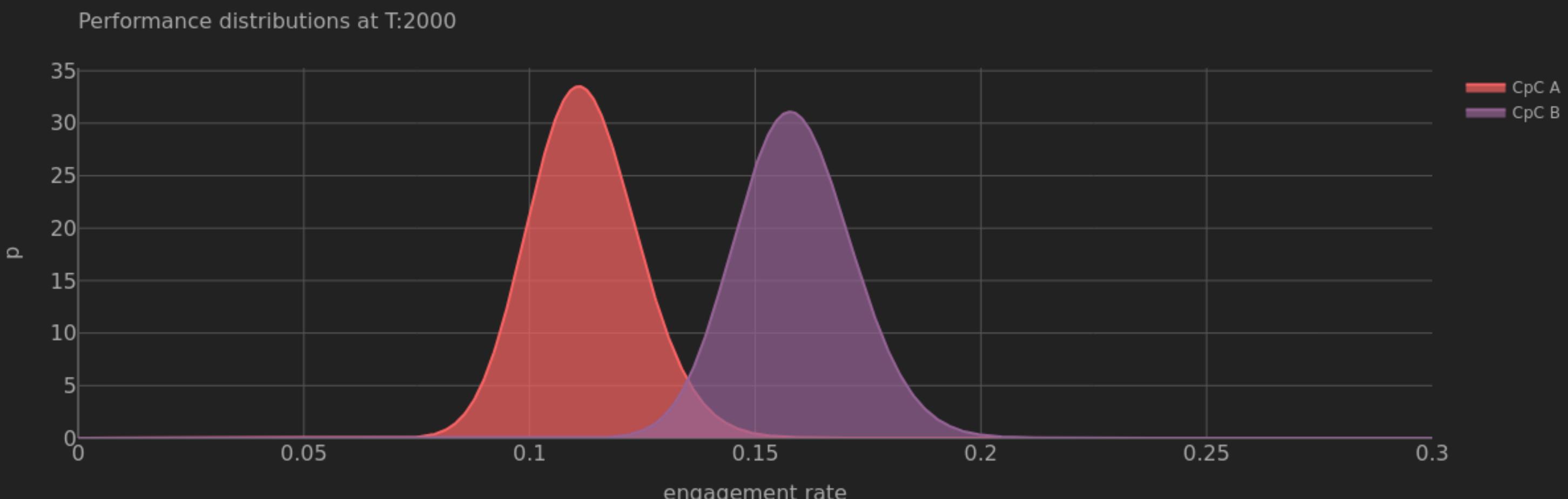
959

clicks

151

cost

End-of-simulation



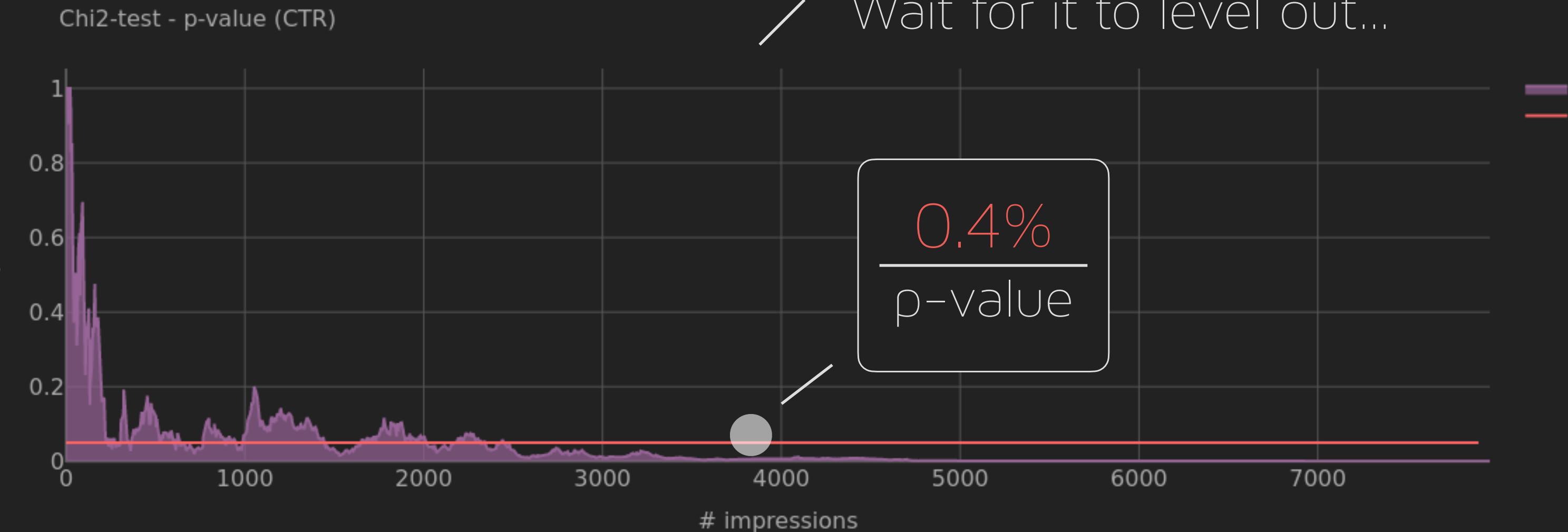
TODO - Hypothesis A/B Testing - Campaign simulation

A/B test

A

$\frac{10\%}{\text{CTR A}}$

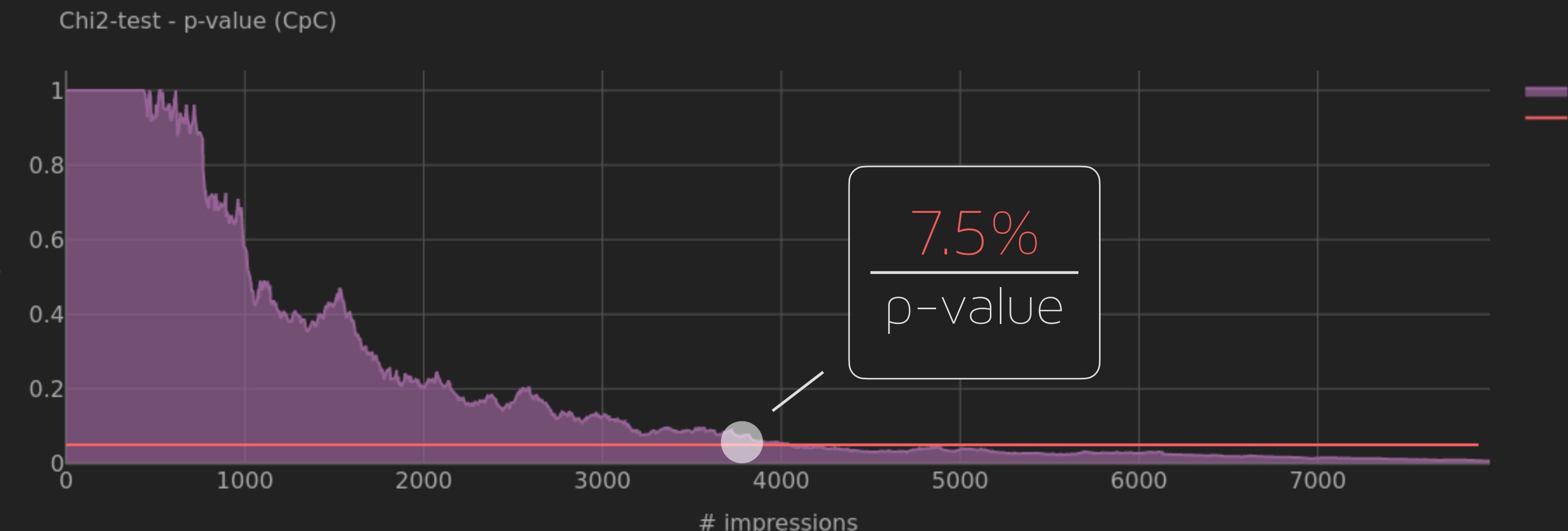
$\frac{\$11}{\text{CPM A}}$



B

$\frac{12\%}{\text{CTR B}}$

$\frac{\$19}{\text{CPM B}}$



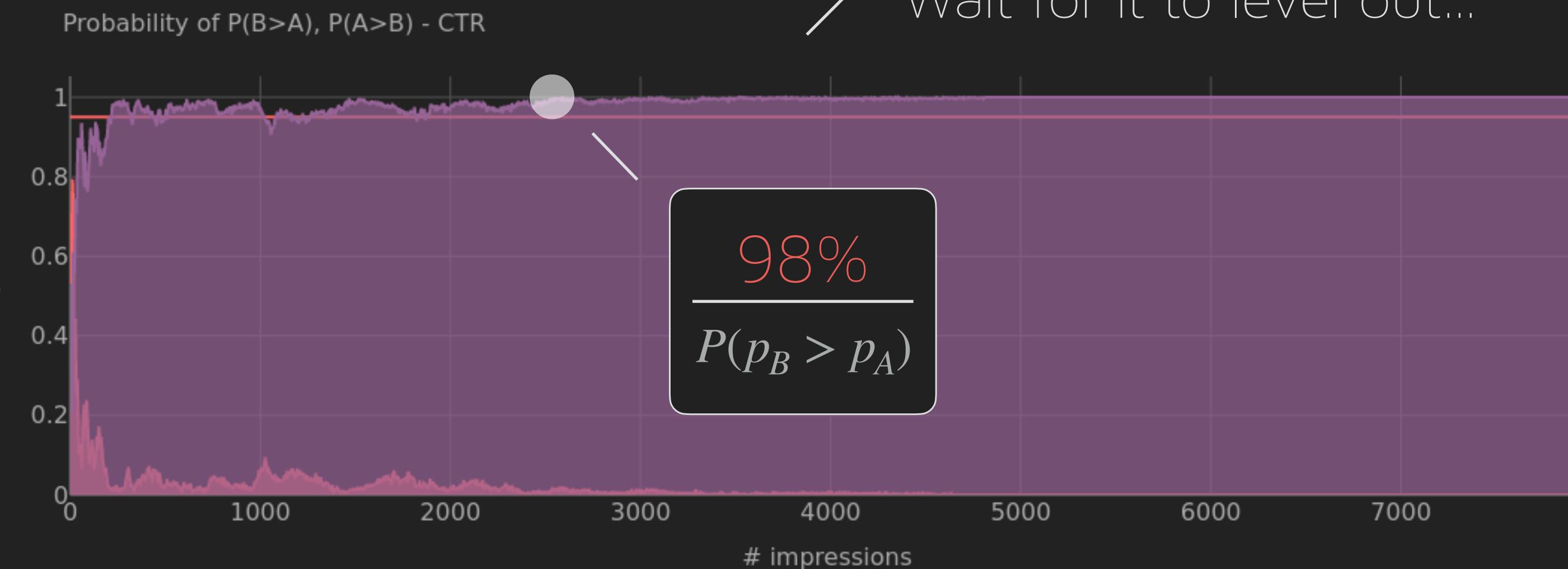
Bayesian A/B Testing - Campaign simulation

A/B test

A

10%
CTR A

\$11
CPM A



Aleatoric uncertainty
Wait for it to level out...

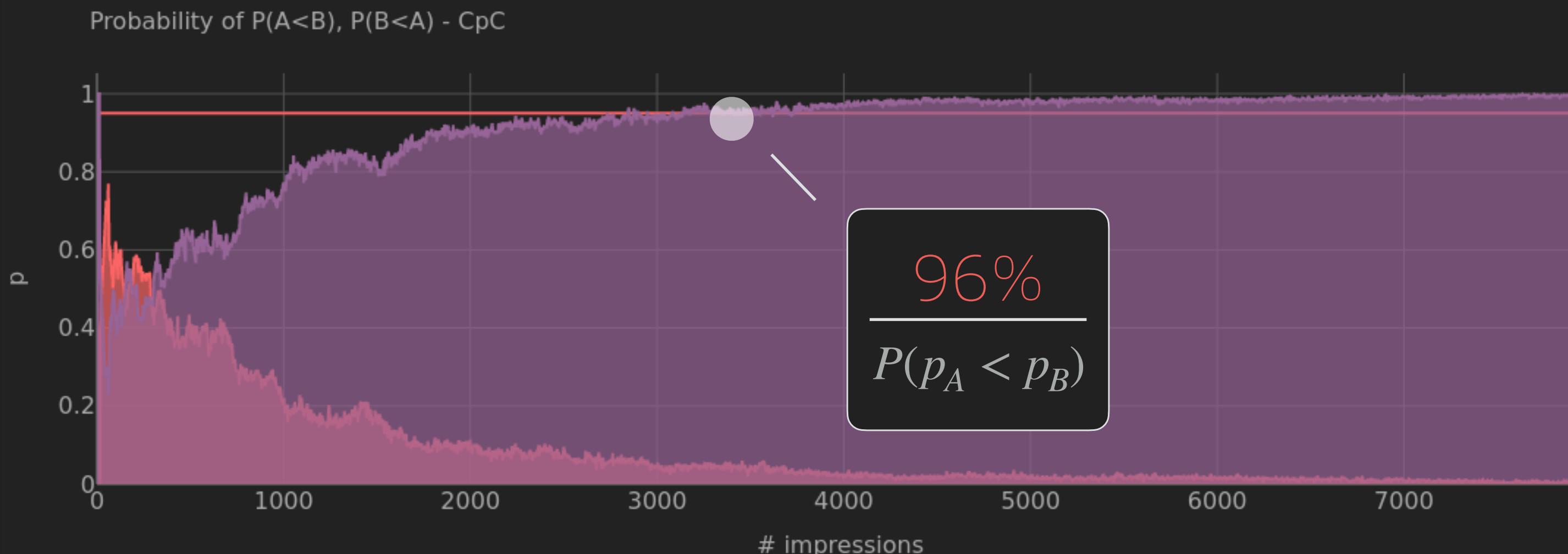
Could we get
results earlier?

$\frac{\sim 500-2500}{\# \text{samples}}$

B

12%
CTR B

\$19
CPM B



Aleatoric
uncertainty
messes with
us...

Hypothesis A/B Testing - Campaign simulation

A/B test

A

$\frac{10\%}{\text{CTR A}}$

$\frac{\$11}{\text{CPM A}}$



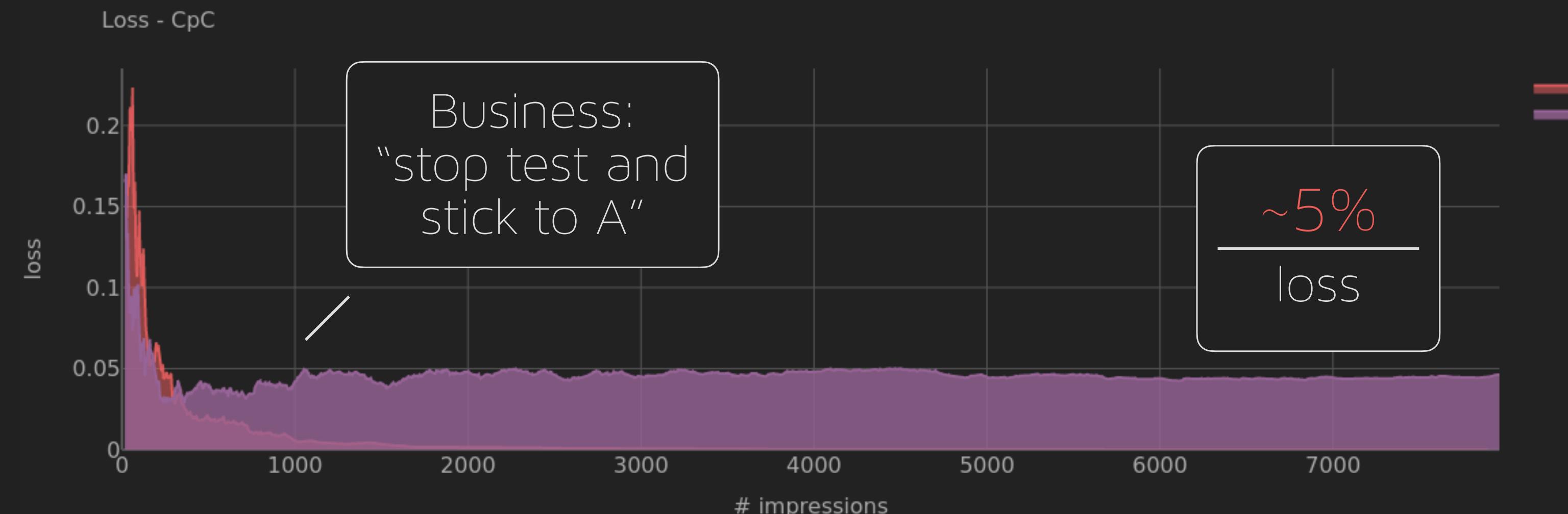
Aleatoric uncertainty not stable,
but results are clear

Choosing A
is pretty
expensive...

B

$\frac{12\%}{\text{CTR B}}$

$\frac{\$19}{\text{CPM B}}$



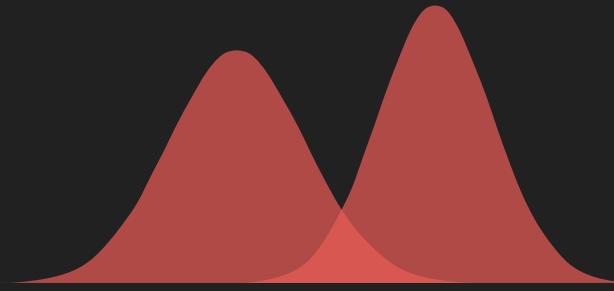
Choosing B is
pretty
expensive...

This time in
real \$\$\$

Bayesian End-of-Level summary



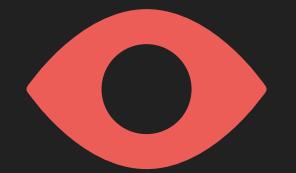
Uncertainty is important for practical evaluations



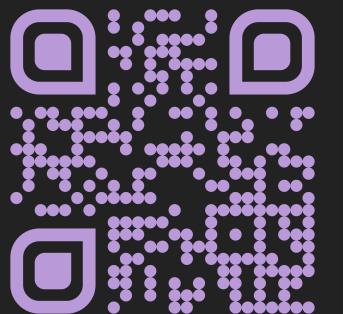
A/A-test - don't forget
Simple tools to evaluate - but not so simple to decide...?



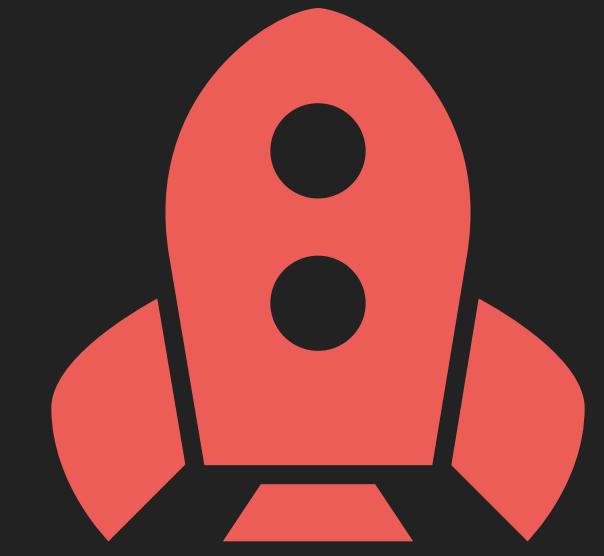
Bayesian A/B-testing - white box - easier to understand
Interpretable metrics - probabilities & losses



Peeking is “allowed”, but be carefull with the aleatoric uncertainty - it will mess with you....



... code available...:-)



References

References

Calculating Sample Size for A/B Testing: Formulas, Examples & Errors

<https://splitmetrics.com/blog/mobile-a-b-testing-sample-size/> (paywall)

Kees Schippers

How to Calculate A/B Testing Sample Sizes?

<https://vwo.com/blog/how-to-calculate-ab-test-sample-size/>

Cameron Davidson-Pilon

Probabilistic Programming and Bayesian Methods for Hackers

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

David J. C. MacKay

Information Theory, Inference, and Learning Algorithms

<https://www.inference.org.uk/itprnn/book.pdf>

Alicia A. Johnson, Miles Q. Ott, Mine Dogucu

Bayes Rules! An Introduction to Applied Bayesian Modeling

<https://www.bayesrulesbook.com>

Evan Miller

Bayesian A/B-Testing

<https://www.evanmiller.org/bayesian-ab-testing.html>

Michael Frasco

The Power of Bayesian A/B Testing

<https://medium.com/convoy-tech/the-power-of-bayesian-a-b-testing-f859d2219d5>

Urteaga et al.

Bayesian bandits: balancing the exploration-exploitation tradeoff via double sampling

<https://arxiv.org/pdf/1709.03162.pdf>

John Cook

Exact Calculation of Beta Inequalities

https://www.johndcook.com/exact_beta_inequalities.pdf

Chris Stucchio

Bayesian A/B Testing at VWO

https://vwo.com/downloads/vwo_SmartStats_technical_whitepaper.pdf

kruschke

Bayesian Estimation Supersedes the t Test

<https://jkkweb.sitehost.iu.edu/articles/Kruschke2013JEPG.pdf>

Hass et al.

Exact Bayesian Inference for A/B testing

<https://gist.github.com/lucidyan/89eee0db8ce353d91e6bfbc51b5dcf19>

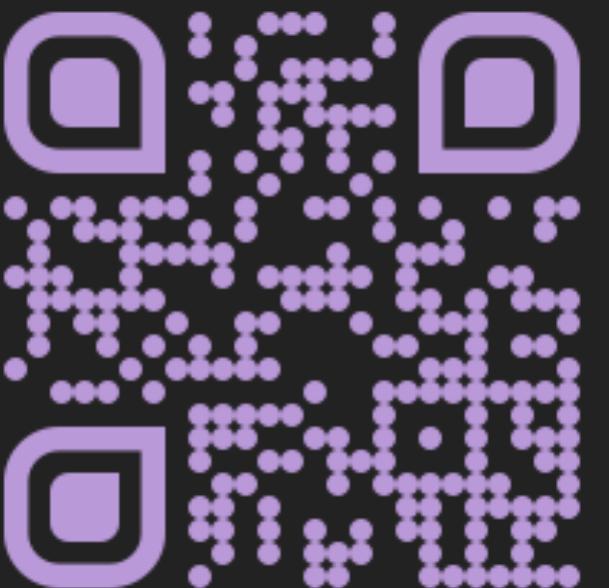
Sureshkumar

Bayesian experimentation methods for products

<https://towardsdatascience.com/bayesian-experimentation-methods-for-products-636514951e43>

“Uncertainty is important in
practical evaluations”

Thank you :-)



<https://github.com/Arngren>

Bayesian Conditioning...

Condition on certain scenarios - eg. reel position / category...

How? Redo calc. on filtered subset of data. Uncertainty will tell you, if you can decide - just as before...

Example	
"Vacation"	"Cars"
12:00 London ...	14:00 Paris ...
3	2

$P(\lambda_B > \lambda_A | \theta), \quad \theta =$

Item category	Time Geography ...	Slate location
Time Geography ...	12:00 London ...	3
Slate location	14:00 Paris ...	2

