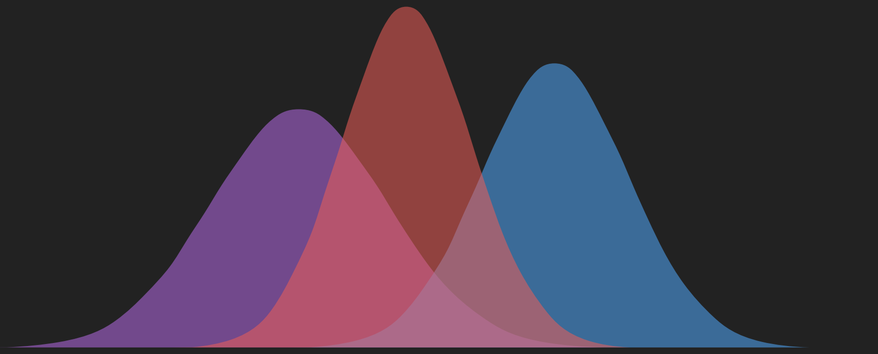


Bayesian A/B-Testing

RecSys Summer School - June 2023

Copenhagen

Morten Arngren
Lead Data Scientist

WUNDERMAN THOMPSON
MAP

Morten Arngren

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

Lead Data Scientist at
WundermanThompson

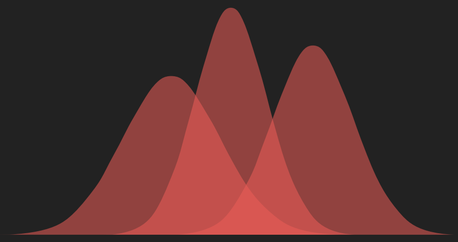
Code and keynote
available on GitHub

https://github.com/Arngren

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

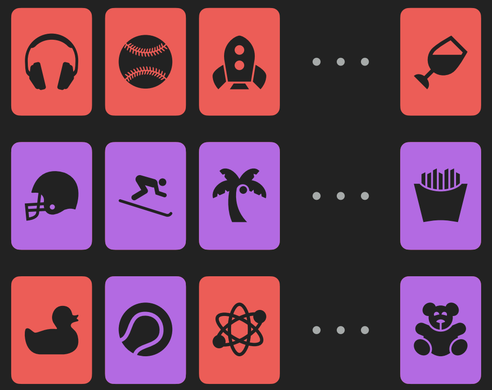
The stage

....

Two Rec. Engines

Movie Streaming

Recommendations



Slates

A

Running in production

New developed Rec. Engine

B

which has
the best
performance
?

News site

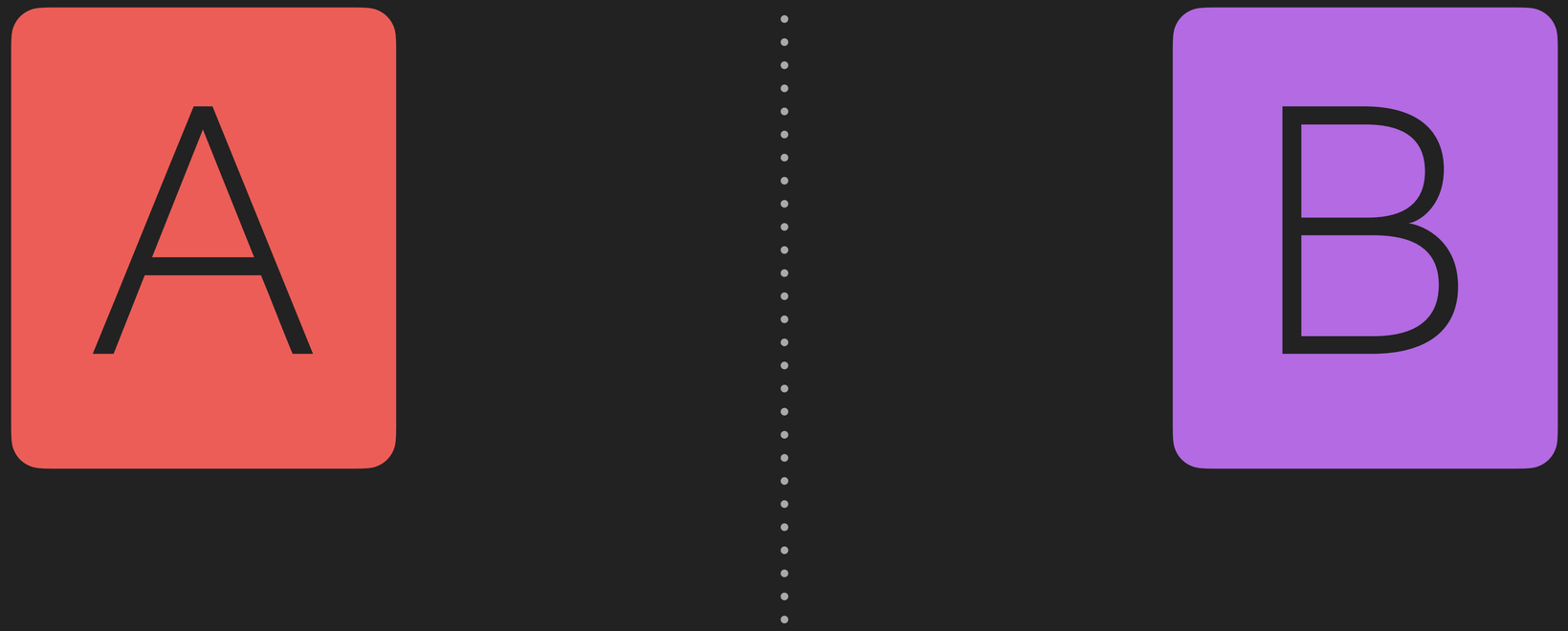
Advertisement

Advertisements

The stage

....

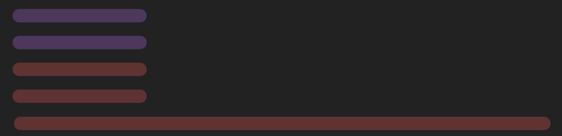
Two Rec. Engines



which has
the best
performance
?

A/B testing

Classic approach



state null & alternative hypothesis

H_0 H_1

define metric CTR, CVR, CpC?

define experiment

assume specific data distribution

which test to use?

calc. sample size to reduce epistemic uncertainty

....

activate

execute campaigns

statistical test

p-value < 5%?

reject H_0?

or

bayesian test

p(B>A)
Loss?

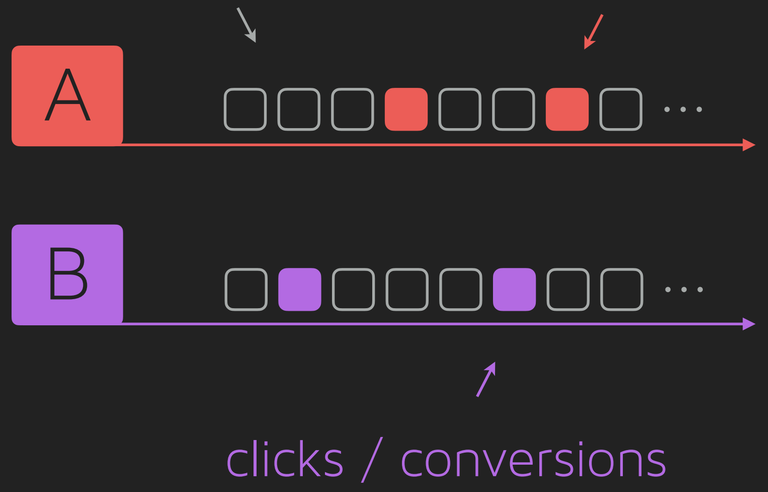
decide

...and go home

The stage

logs

impressions clicks / conversions



business metrics

CTR	Click-Through-Rate Fraction of clicks? [%]
CVR	Conversion Rate Product buys [%]
CpC	Cost-per-Click Price of a click [$]
???	Webshops margins Earnings [$]

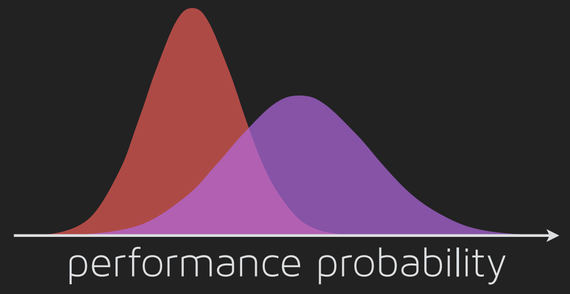
evaluation

Hypothesis testing

p-value < 5%?

or

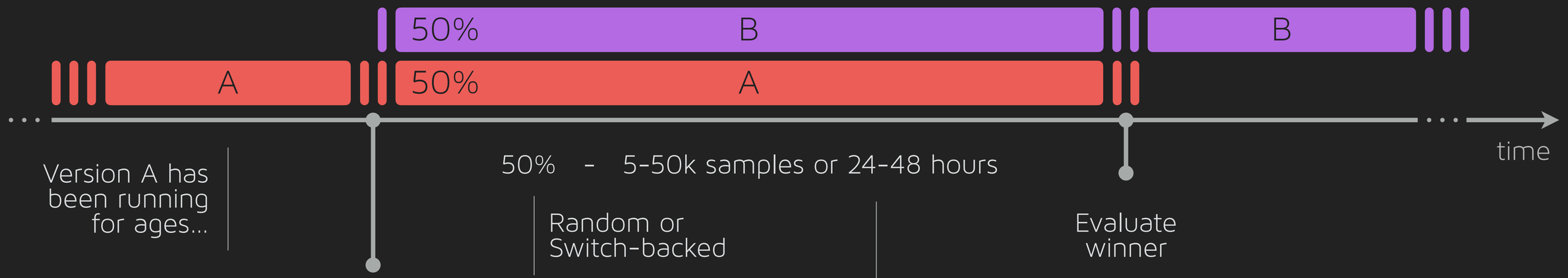
Bayesian testing



Introduction

Classic A/B-test

Timeline...



Version A has been running for ages...

A lot of planning / assumptions...

How many versions / samples?

What / how to measure?

Start test

50% - 5-50k samples or 24-48 hours

Random or Switch-backed

Both A and B are put in production and we wait some time...

Evaluate winner

only evaluate when time is up...

hypothesis testing p-values...

decide p-value < 5%

Complex A/B-test

...or the more right way...

~20%

Reserve

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

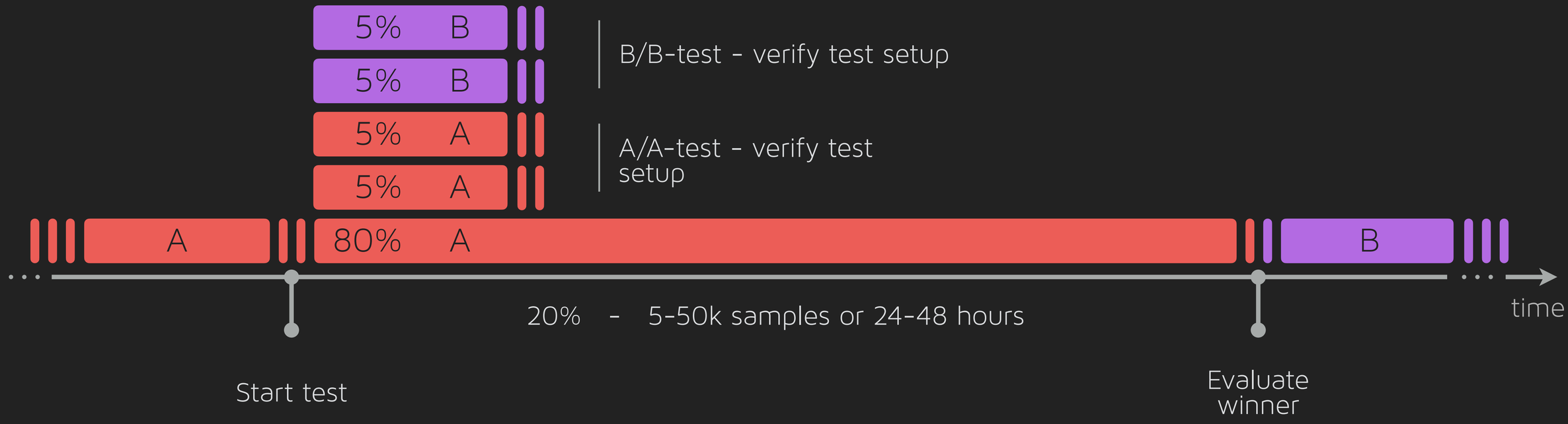
consider business requirements
and price...!

A/A test

Verify

split into several smaller tracks

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



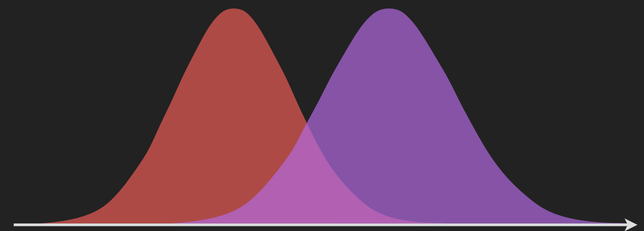
Bayesian A/B-test

Bayesian evaluating - probability of winner

models performance as probability distributions

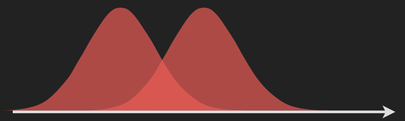
captures the epistemic uncertainty

allow us to quantify the confidence



A/A-tests

successful



failed

Bandit A/B-test

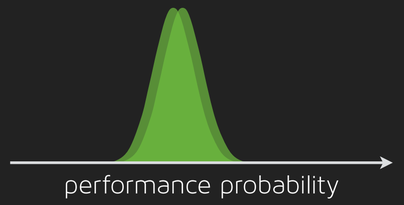
Dynamic architecture - runs self-contained - drop in new versions as cold start

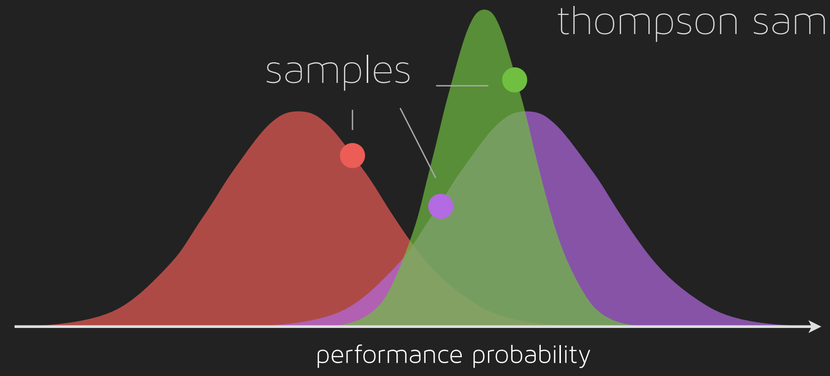
models performance as probability distributions

captures the epistemic uncertainty

allow us to quantify the confidence

A/A-tests





16% C

16% C

16% B

16% B

16% A

16% A

A

successful

failed

time

Start test

select dynamically from Thompson sampling

evaluate continuously

can run until sufficient confidence.....and A/A test has passed

What we really want to know...

λ Engagement performance – eg. CTR / CVR / Open-Rate

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

Probability of B
being better than A

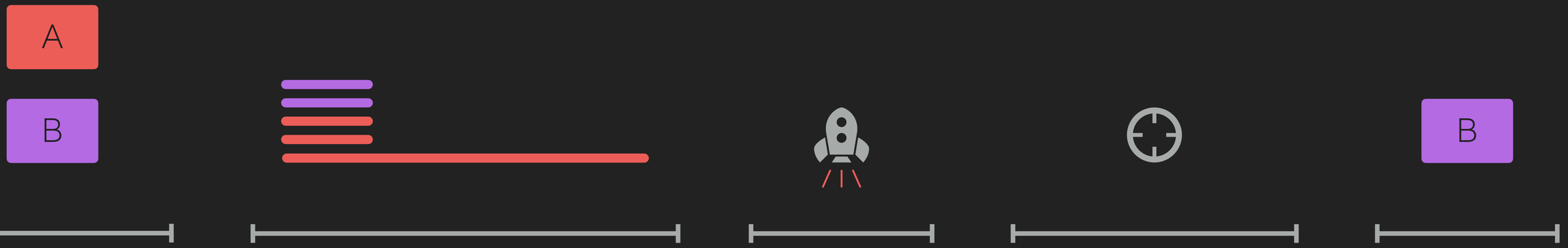
Hypothesis test answers
another question

Bayesian approach
answers this question

Hypothesis Testing

Null-hypothesis significance testing

What's specific to it...?



state null & alternative hypothesis

H_0 H_1

define metric
CTR, CVR, CpC?

define experiment

assume specific data distribution

which test to use?

calc. sample size to reduce epistemic uncertainty

....

activate

execute campaigns

statistical test

p-value < 5%?

reject H_0 ?

simple
accept / reject,
not size

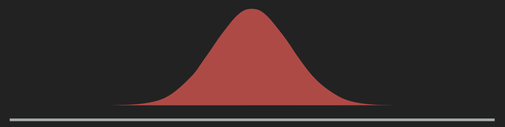
decide

...and go home

Null-hypothesis significance testing

Popular hypothesis tests...

Students T-Test



Assumptions

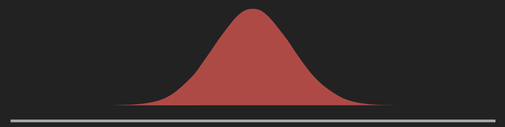
Student-t distributions

Very few samples

Mean of two identical distributions

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

Z-test



Assumptions

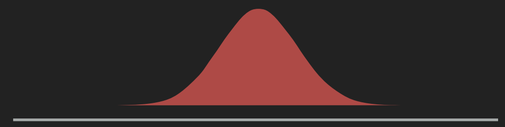
Gaussian distributions

Homogeneity of variance

Mean of two identical Gaussian distributions

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

Fisher's exact test



Assumptions

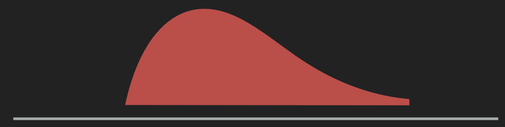
Contingency tables

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

Null-hypothesis significance testing

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%

R. Fischer - “Lady tasting tea” - 1935

Milk or tea first?

Dr. Murial Bristol claimed she could...

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

Dr. Murial Bristol nailed all of them...

4 successes has 1 chance out of 70 (≈ 1.4% < 5%)

= p-value

https://en.wikipedia.org/wiki/Lady_tasting_tea

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

https://splitmetrics.com/blog/mobile-a-b-testing-sample-size/
https://vwo.com/blog/how-to-calculate-ab-test-sample-size/

Single-sided or double sided test?

What are we testing for?

H_1 A < B

H_1 A <> 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!

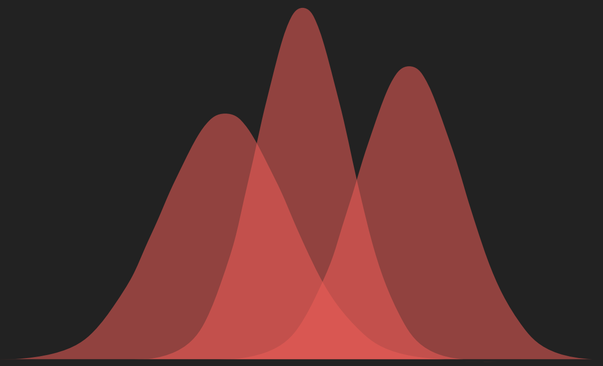
relies on estimating the significance when samples is reached

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

two main types

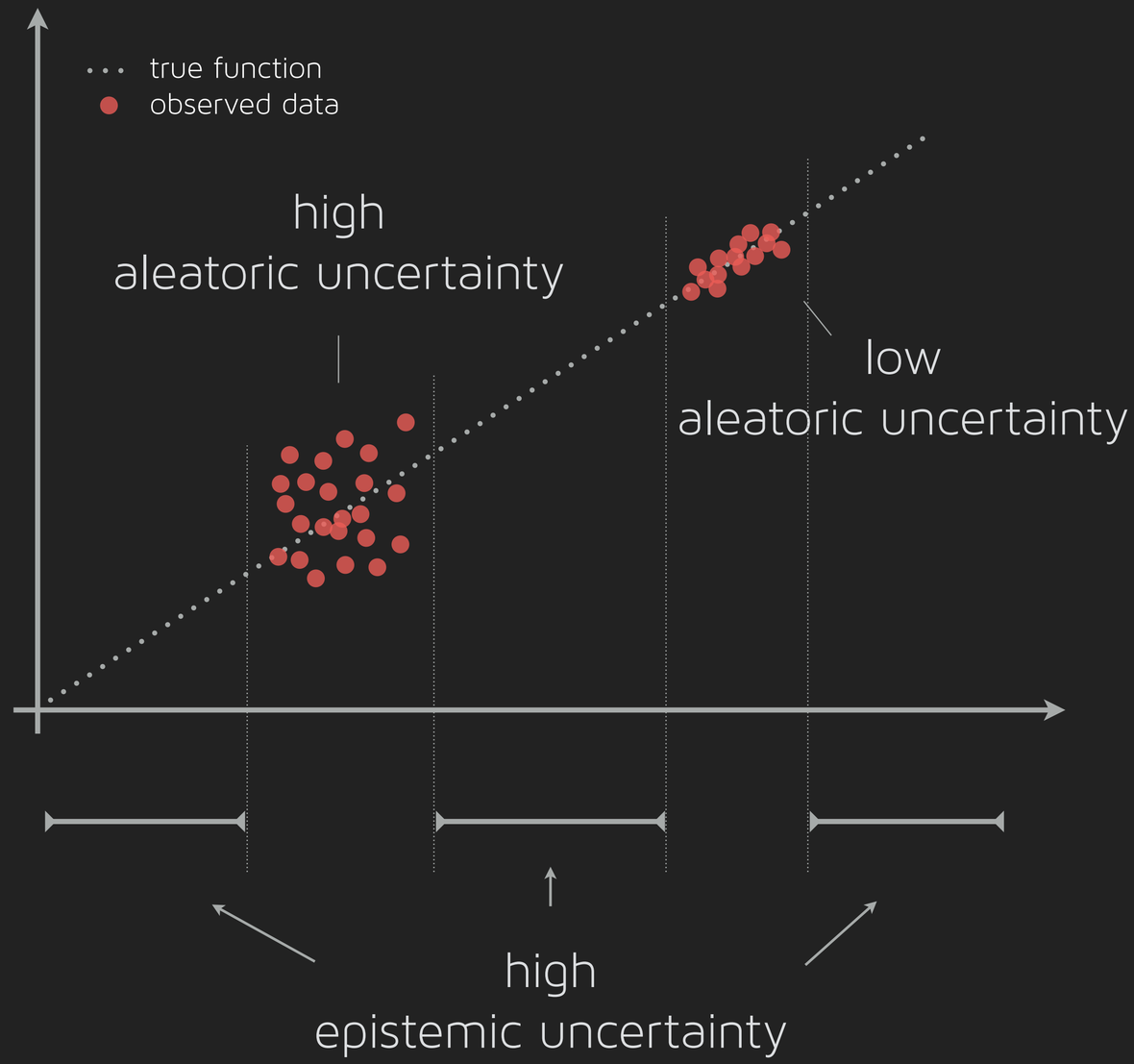
aleatoric uncertainty

- natural stochasticity in observations (noise)
- can't be reduced with more data

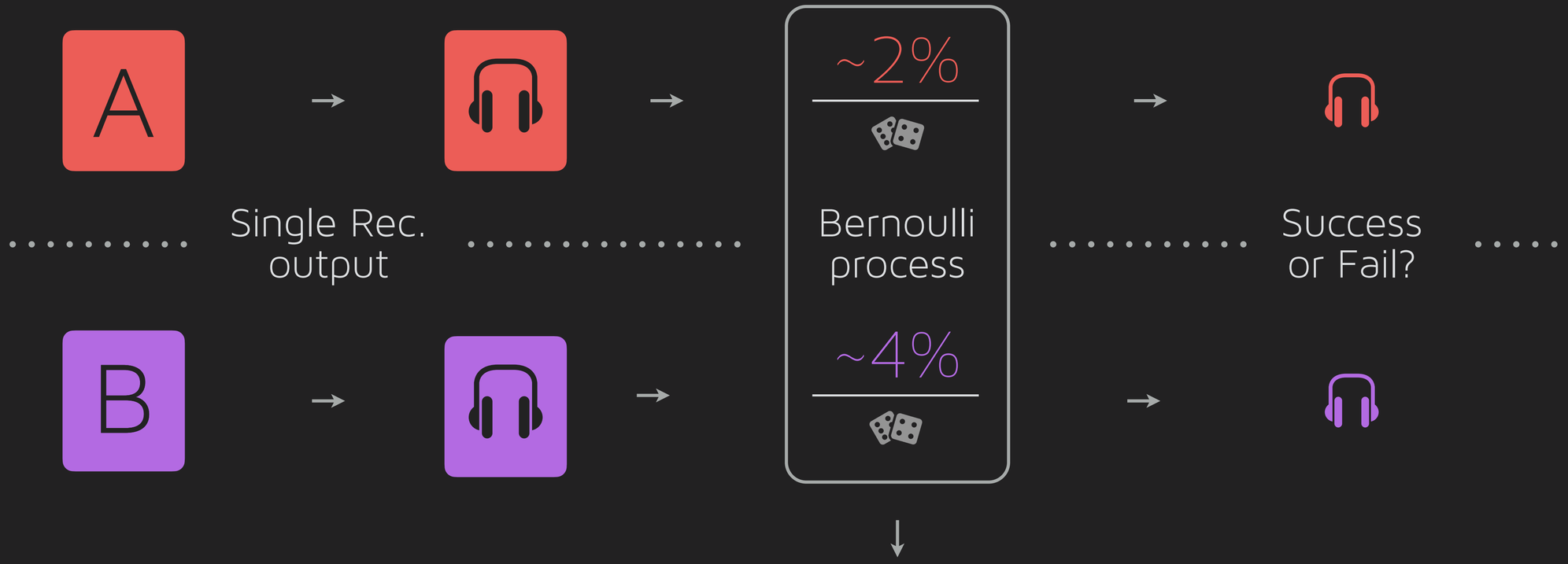
epistemic uncertainty

- uncertainty around too few data points
- out-of-distribution outliers

example



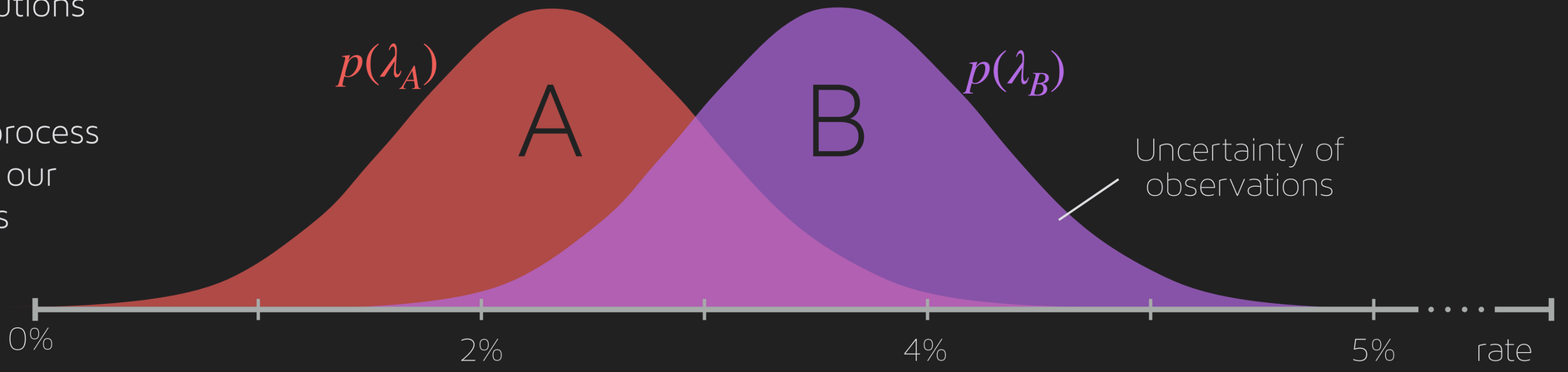
Stochastic process behind the scenes...



Prior Beta distributions

$p(\lambda)$

for the Bernoulli process that generated our observations



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.

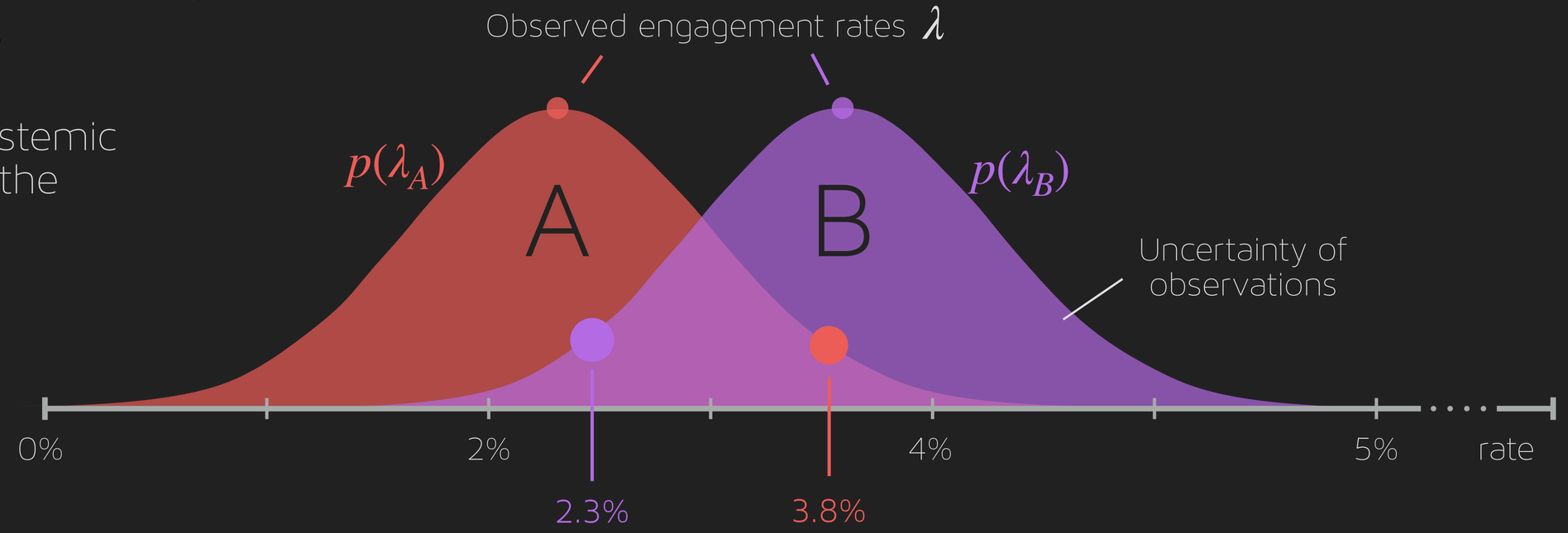
Example of two candidates and their engagement rates with uncertainty.

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

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

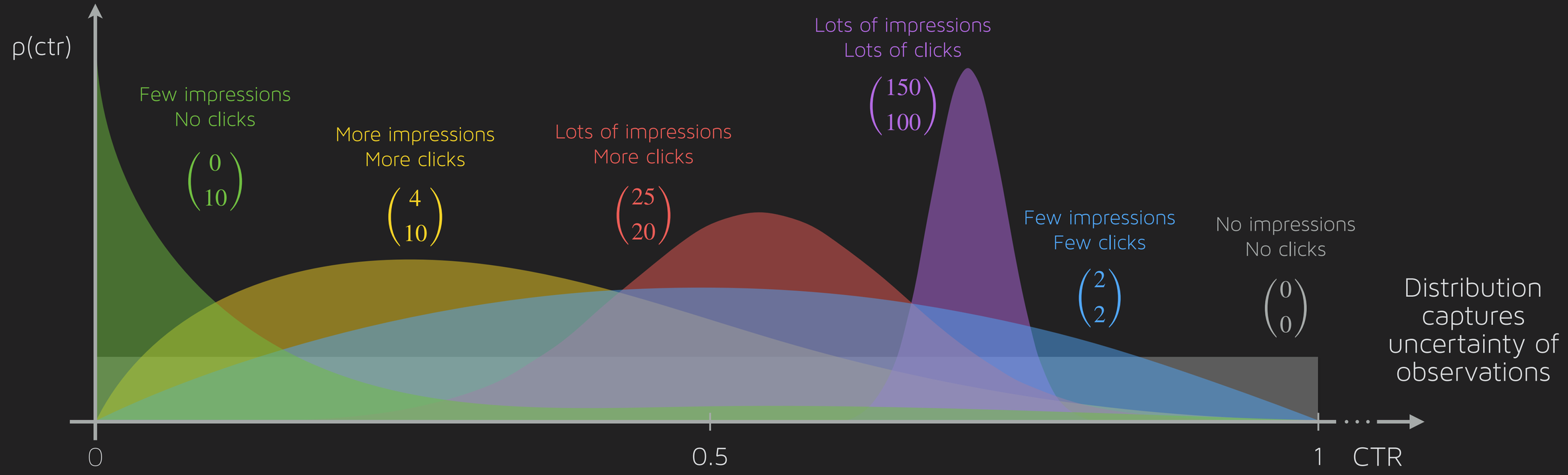
$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} clicks \\ impression - clicks \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

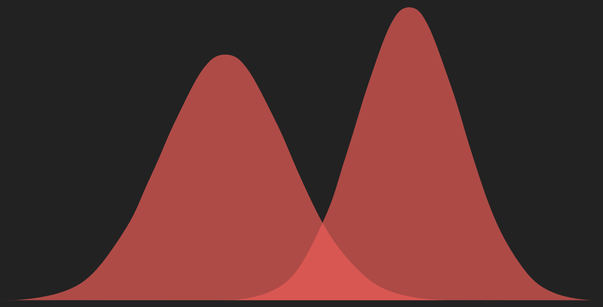
cost / click

gamma distribution

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

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





Bayesian Metrics
A/A testing

Bayesian same/same test

Are A/A the same distribution?

Selected metrics for measuring differences between two probability distributions

Kolmogorov-Smirnov

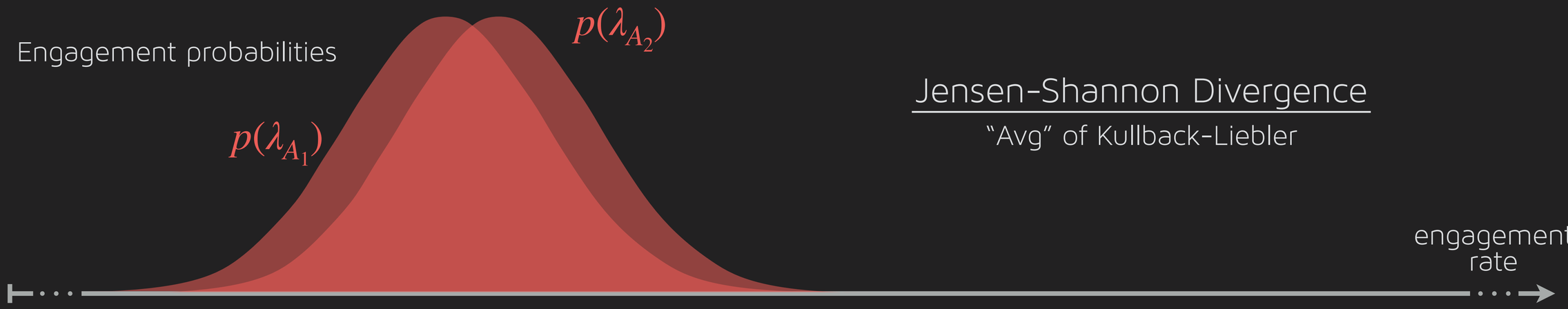
How much are the distribution overlapping?

Wasserstein

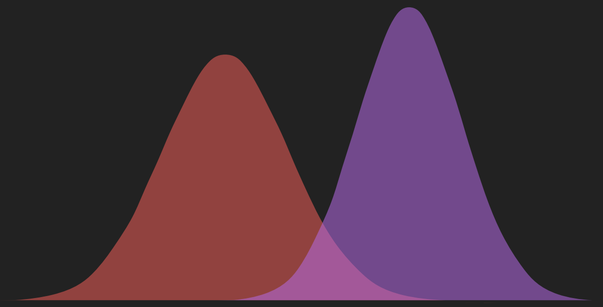
How much mass is moved?
(1 - wasserstein)

Jensen-Shannon Divergence

"Avg" of Kullback-Liebler



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 $p_{\lambda_B} > p_{\lambda_A}$

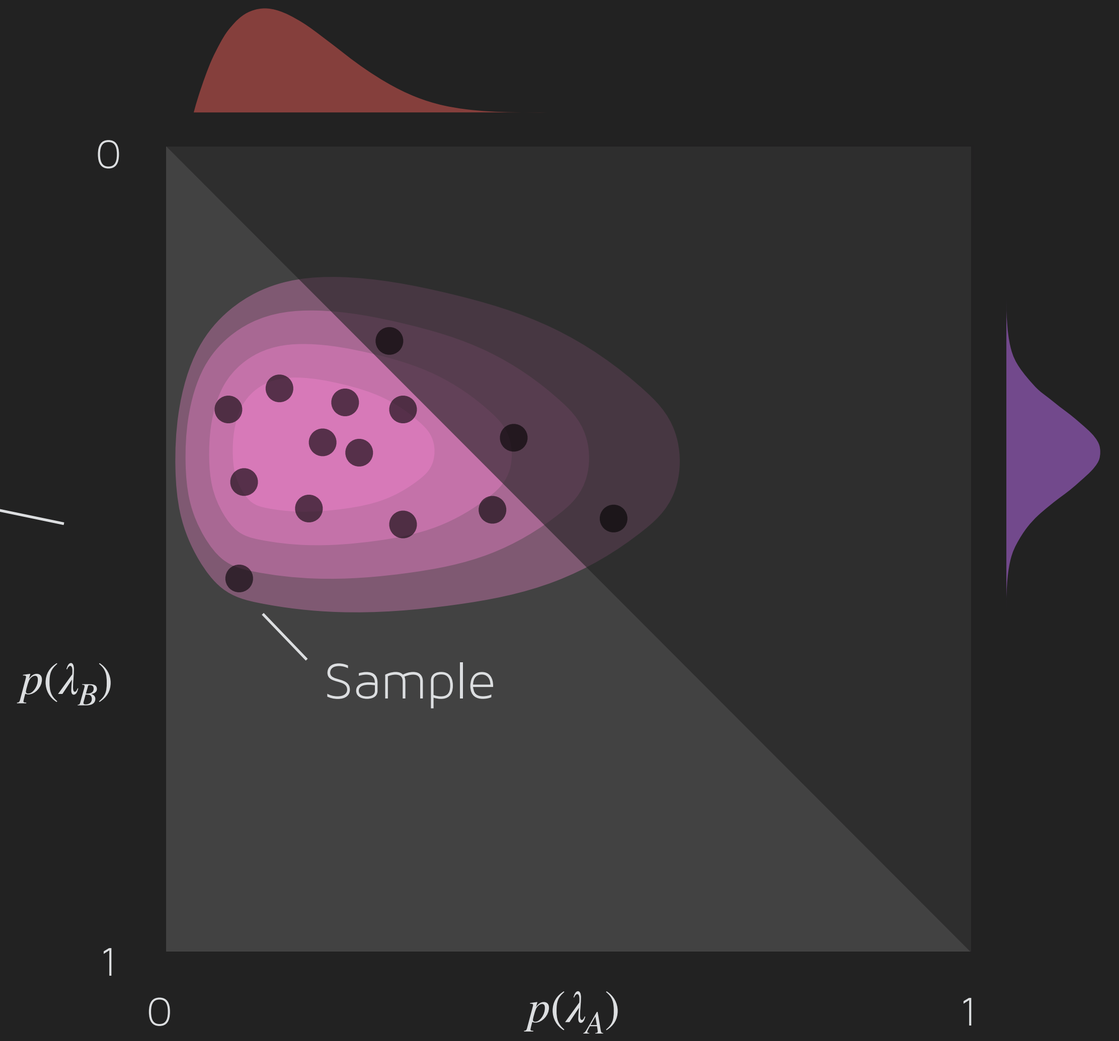
This marginalisation can be estimated via Monte Carlo sampling

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

$$P(p_B > p_A) = \int_0^1 \int_0^{\lambda_A} P(\lambda_A, \lambda_B)\ \mathrm{d}\lambda_B \mathrm{d}\lambda_A$$

$$= \sum_{i=0}^{\alpha_B - 1} \frac{B(\alpha_A + i, \beta_A + \beta_B)}{(\beta_B + i) B(1 + i, \beta_B) B(\alpha_A, \beta_A)}$$

Joint probability of engagement-rates



https://www.evanmiller.org/bayesian-ab-testing.html
https://www.chrisstucchio.com/blog/2014/bayesian_ab_decision_rule.html

Bayesian Metrics - Expected Loss

Average performance loss for wrong choice

Evaluation of loss if choosing wrong

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

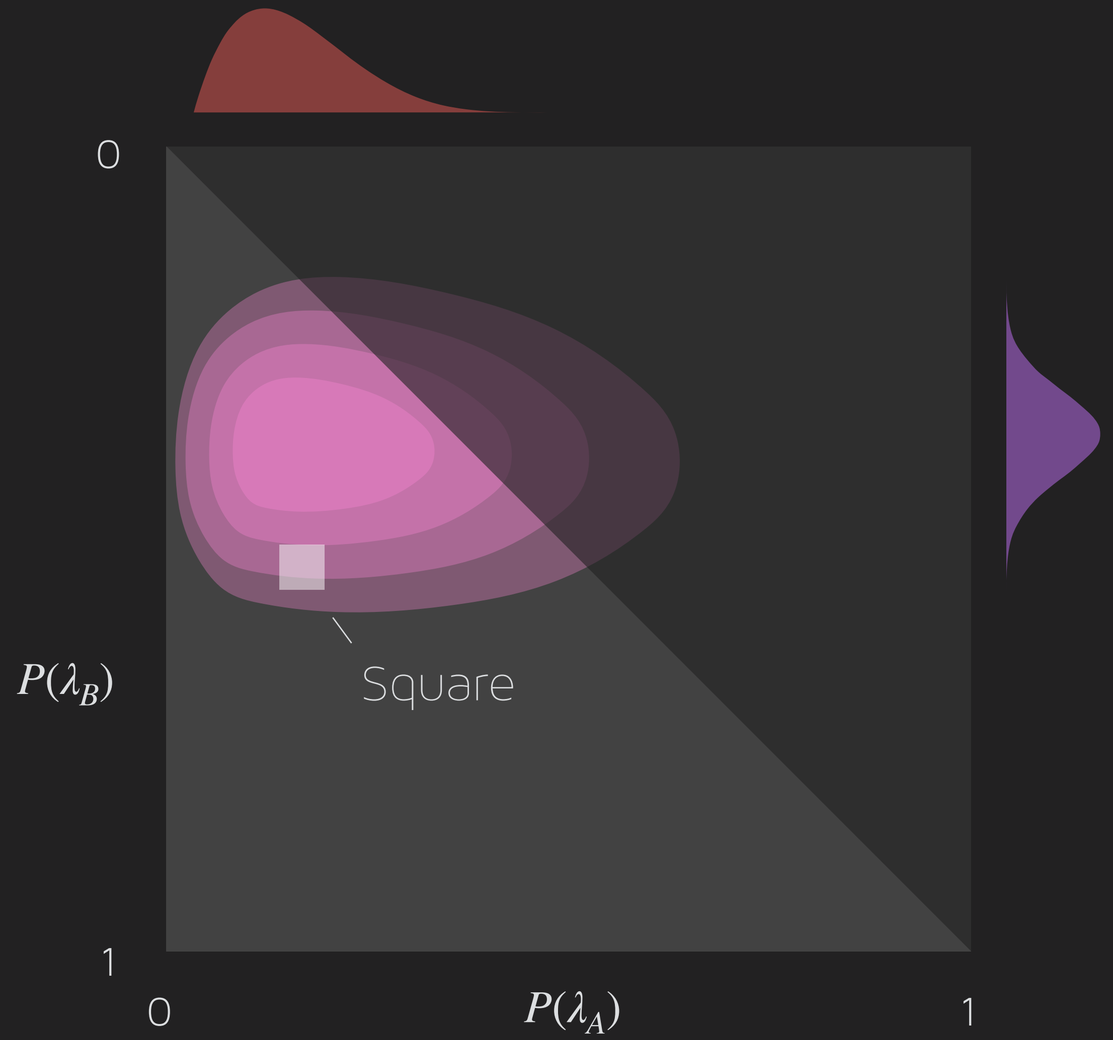
$$\mathscr{L}(\lambda_A, \lambda_B, B) = max(\lambda_A - \lambda_B, 0)$$

Expectation of loss by integrating across joint distribution

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

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

Joint probability of engagement-rates



https://www.chrisstucchio.com/blog/2014/bayesian_ab_decision_rule.html
https://medium.com/vptech/introduction-to-bayesian-a-b-testing-in-python-df81a9b3f5fd

Single Example

Example....

Observations

A	1000 impressions	100 clicks	$11 Cost
B	1000 impressions	120 clicks	$19 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

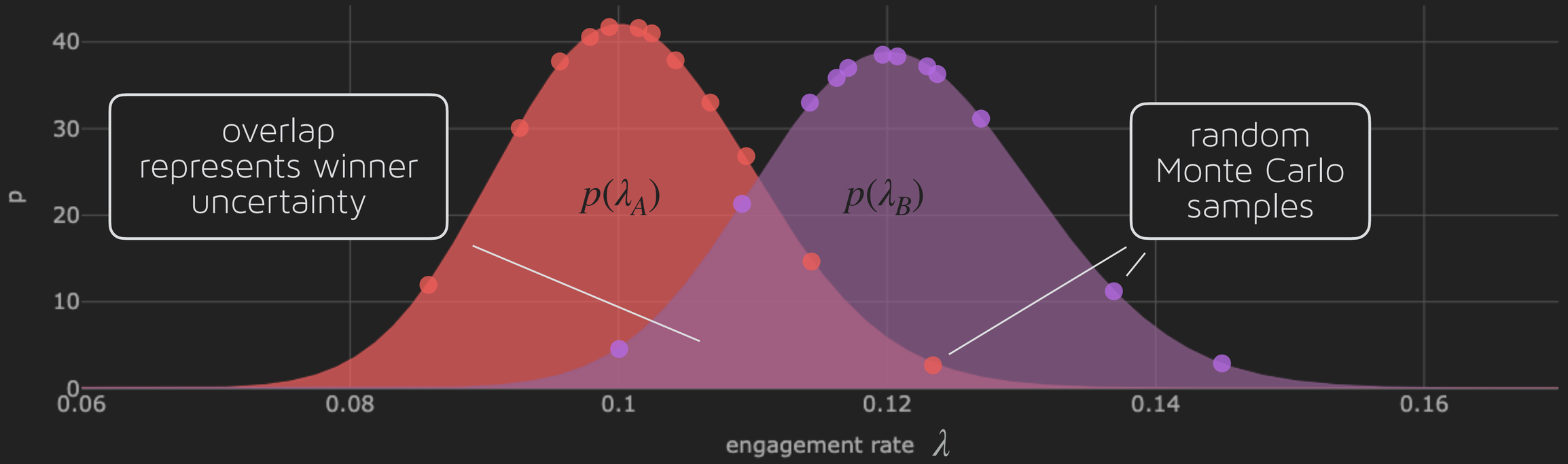
Impressions	1000
Clicks A	100
Clicks B	120

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

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

→ $\frac{\text{B}}{\text{winner}}$?

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	1000
Clicks A	100
Clicks B	120

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

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

$$\rightarrow \quad \frac{\text{B}}{\text{winner}} \quad ?$$

Probability distribution of lift ratio $\frac{\lambda_B}{\lambda_A}$

Bayesian Metrics - Probability of winner & lift analysis

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

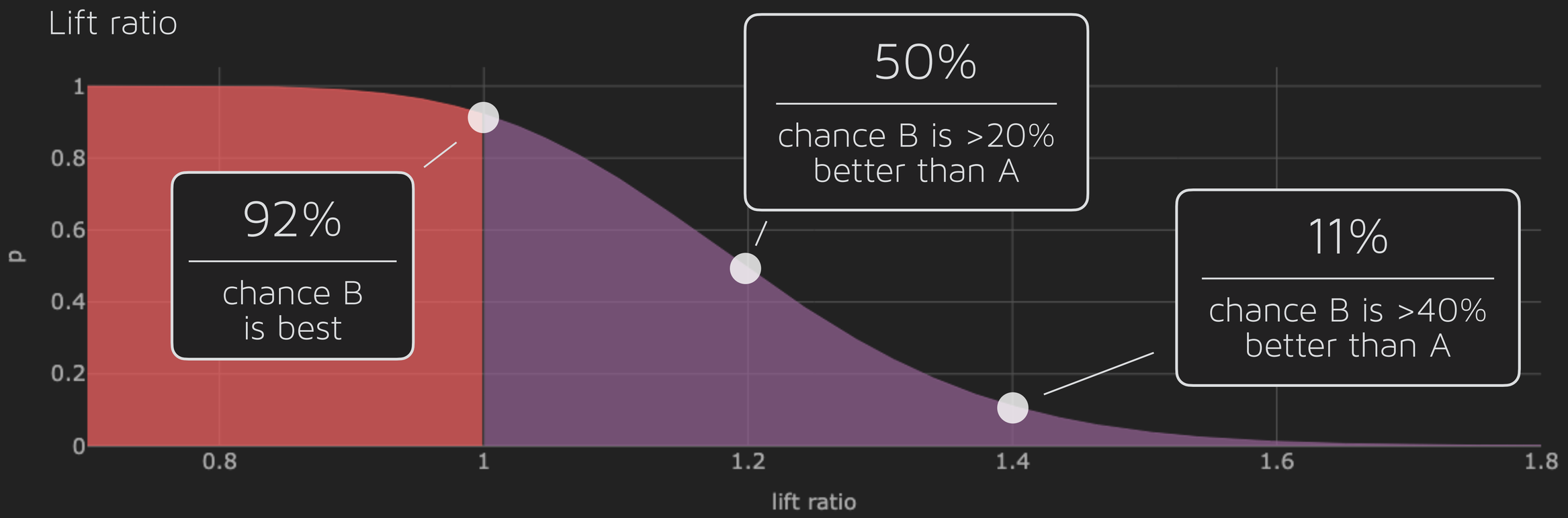
Example

Impressions	1000
Clicks A	100
Clicks B	120

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

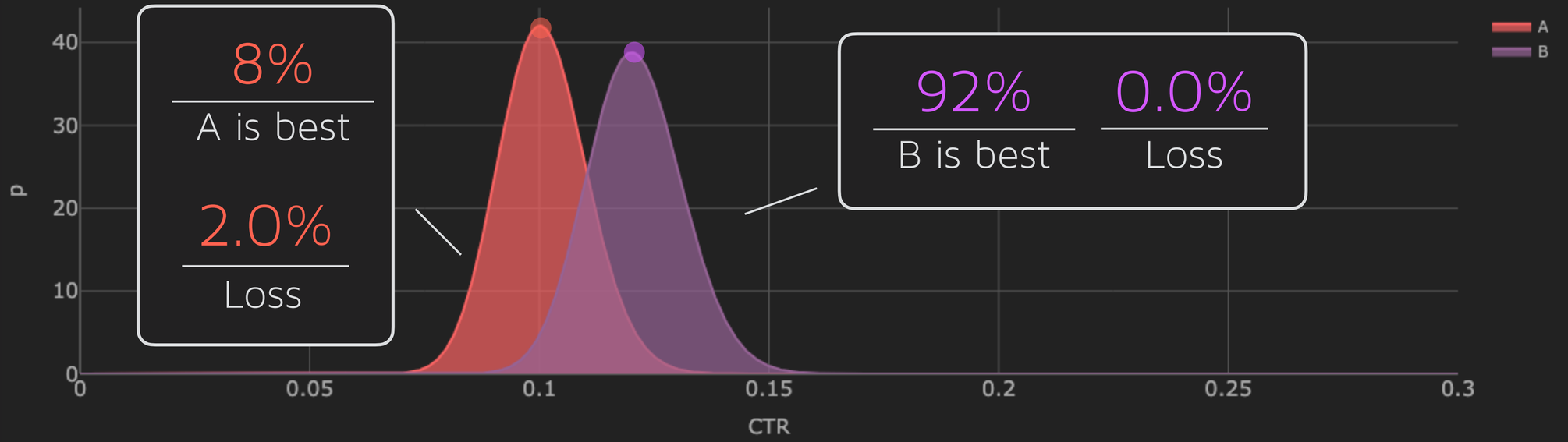
$$\lambda_B = \frac{120}{1000} = 12\,\%$$

→ B / winner ?



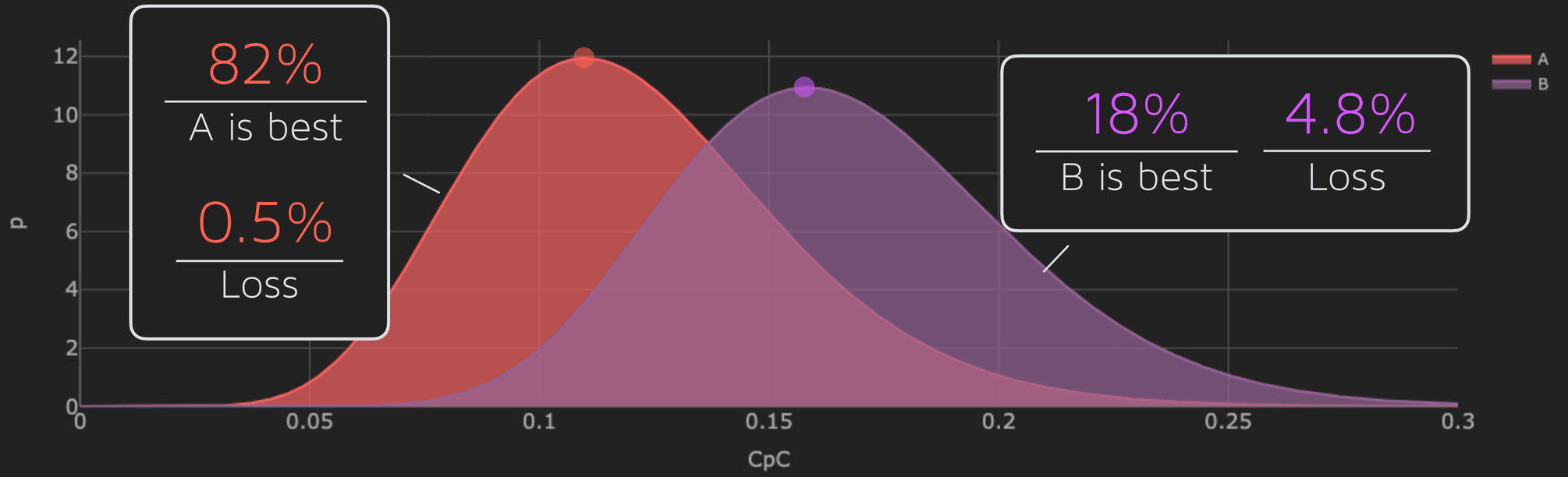
Bayesian A/B Testing - Decision

Click-Through-Rate



After 1000 impressions

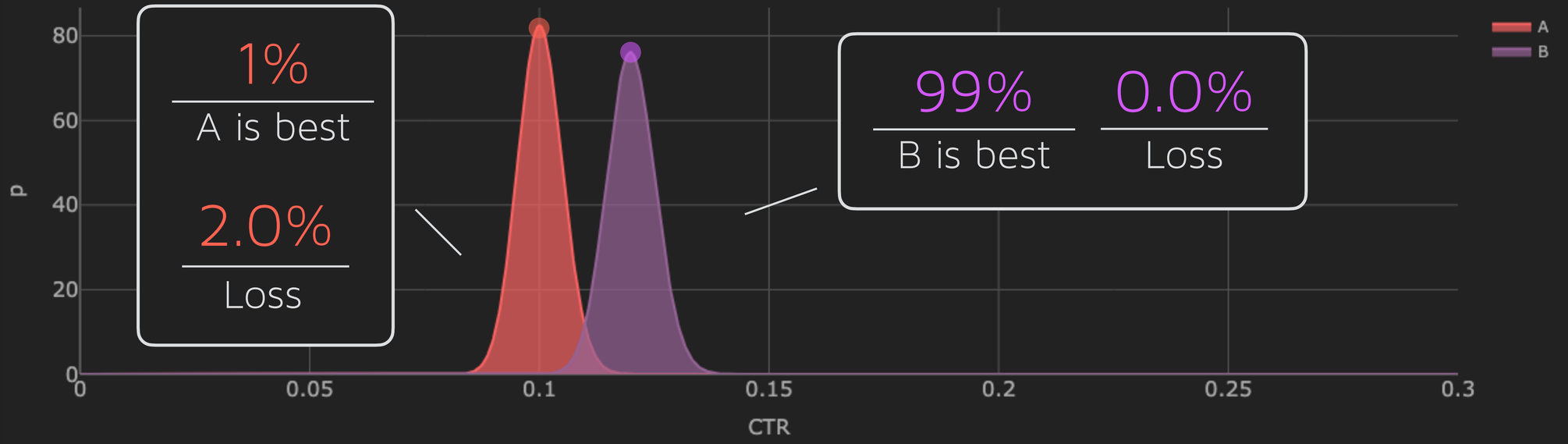
Cost-per-Click



Aleatoric uncertainty still lingers....

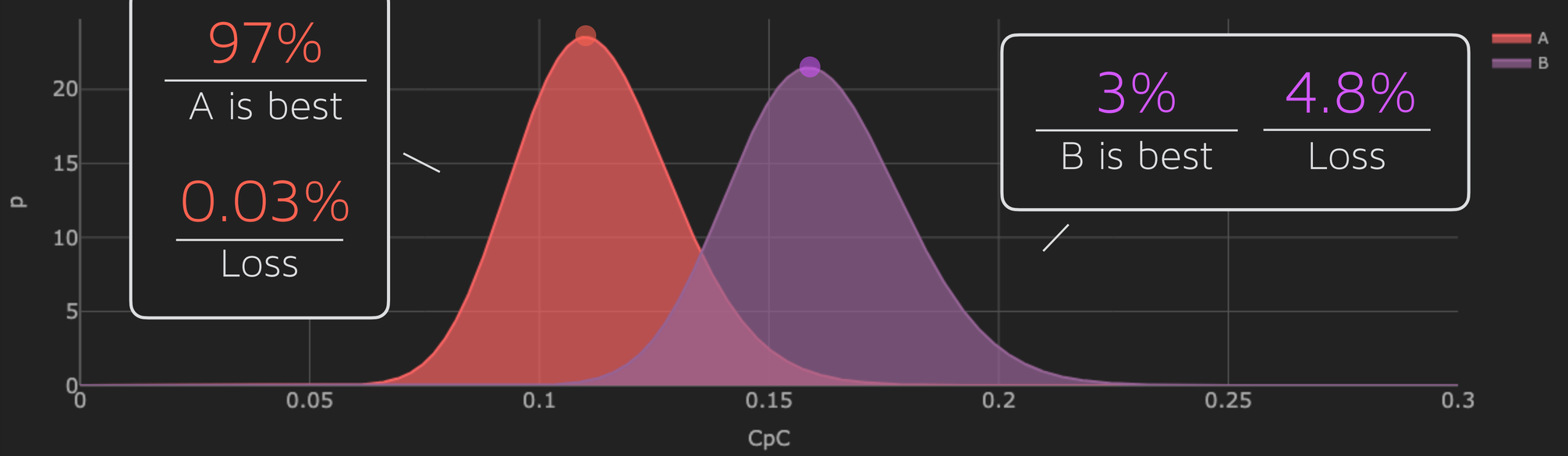
Bayesian A/B Testing - Decision

Click-Through-Rate



After 3840 impressions

Cost-per-Click



Aleatoric uncertainty still lingers....

Campaign

A/B Testing - Campaign simulation

...

A

10%

CTR A

11

CPM A

B

12%

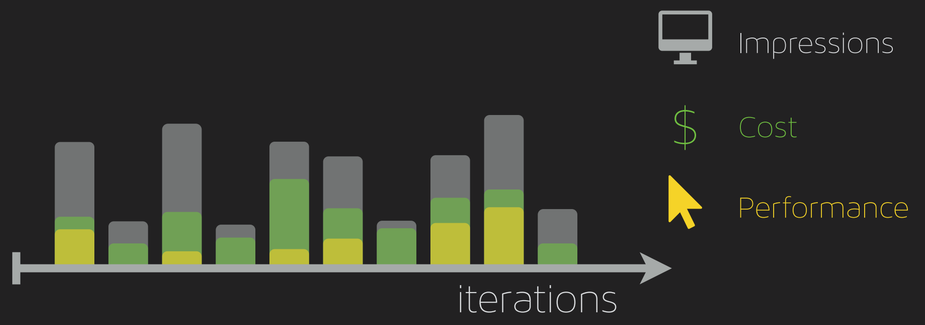
CTR B

19

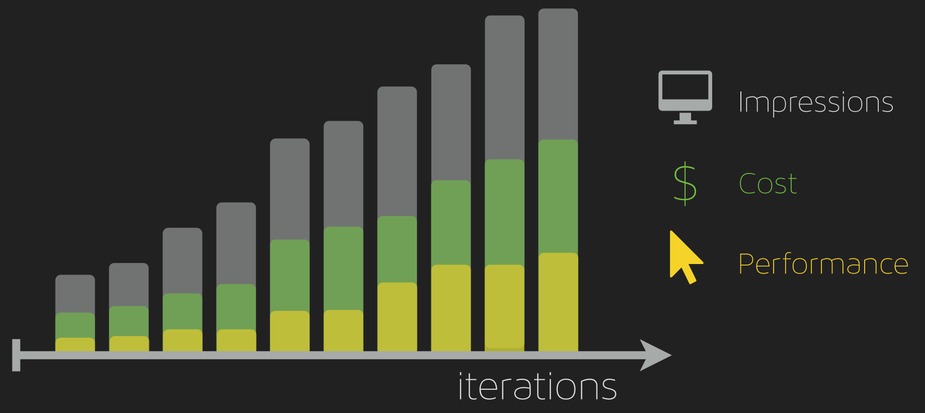
CPM B

Simulate impression, clicks and cost

logs



accumulated



Campaign simulation

Observations and performance

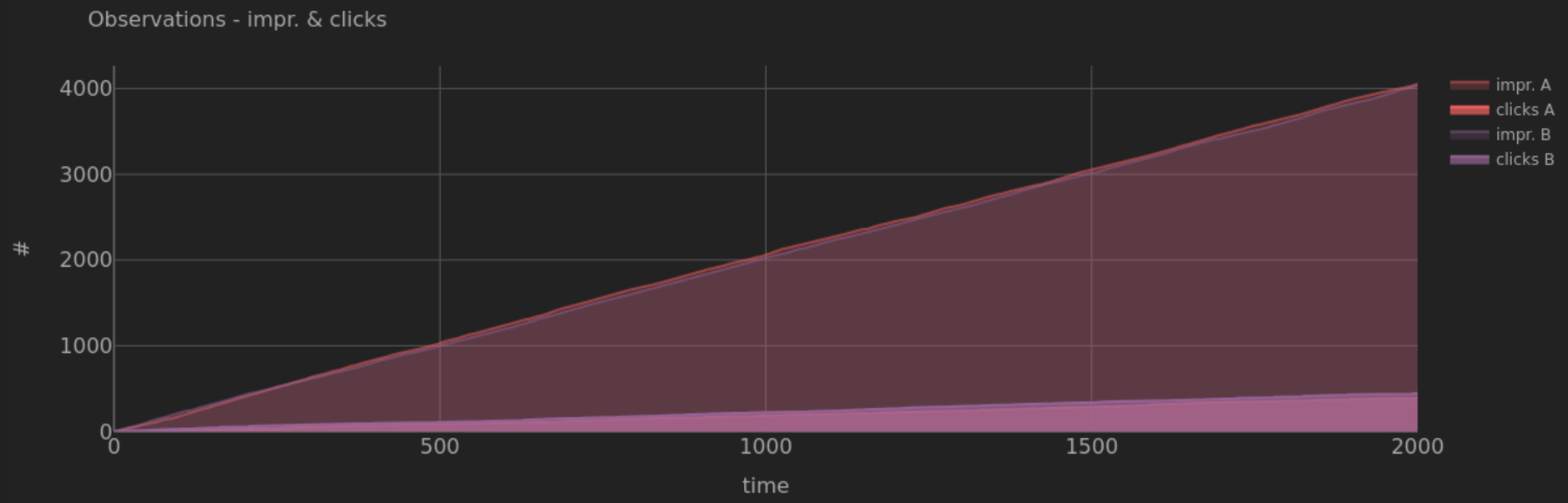
A

10%

CTR A

11

CPM A



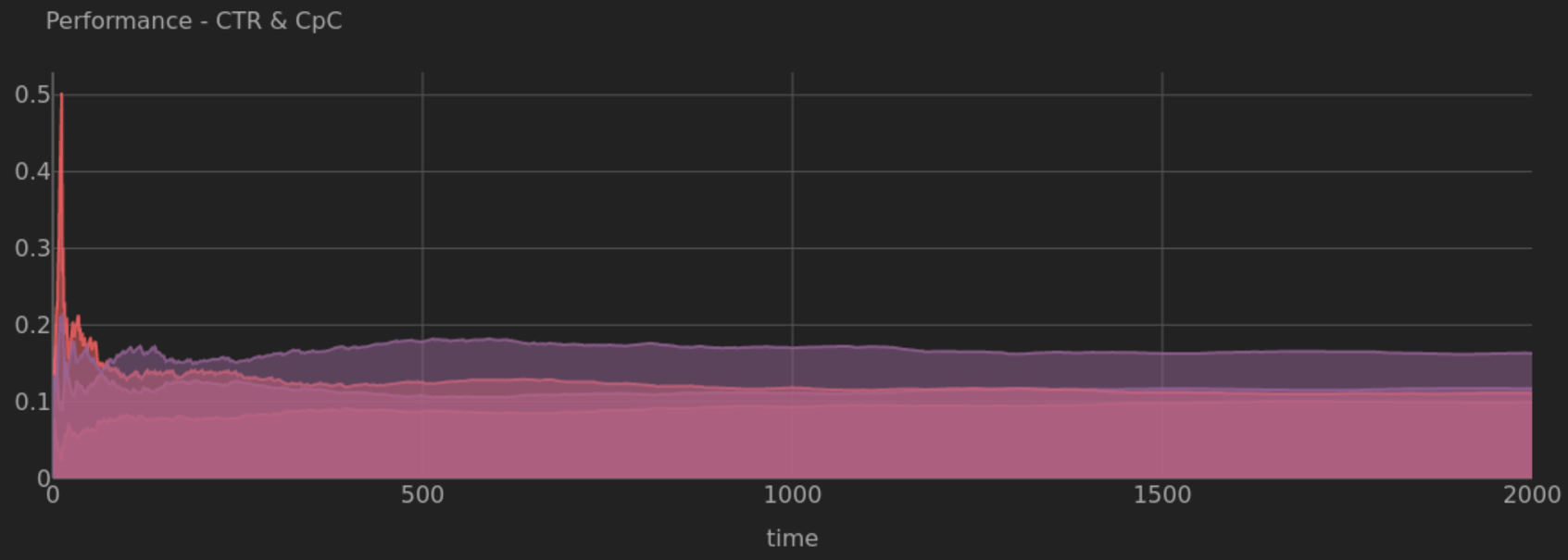
B

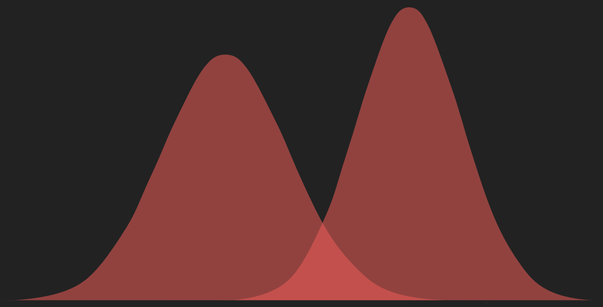
12%

CTR B

19

CPM B





Bayesian A/A testing

Hypothesis A/A Testing - Campaign simulation

Same / Same test

A

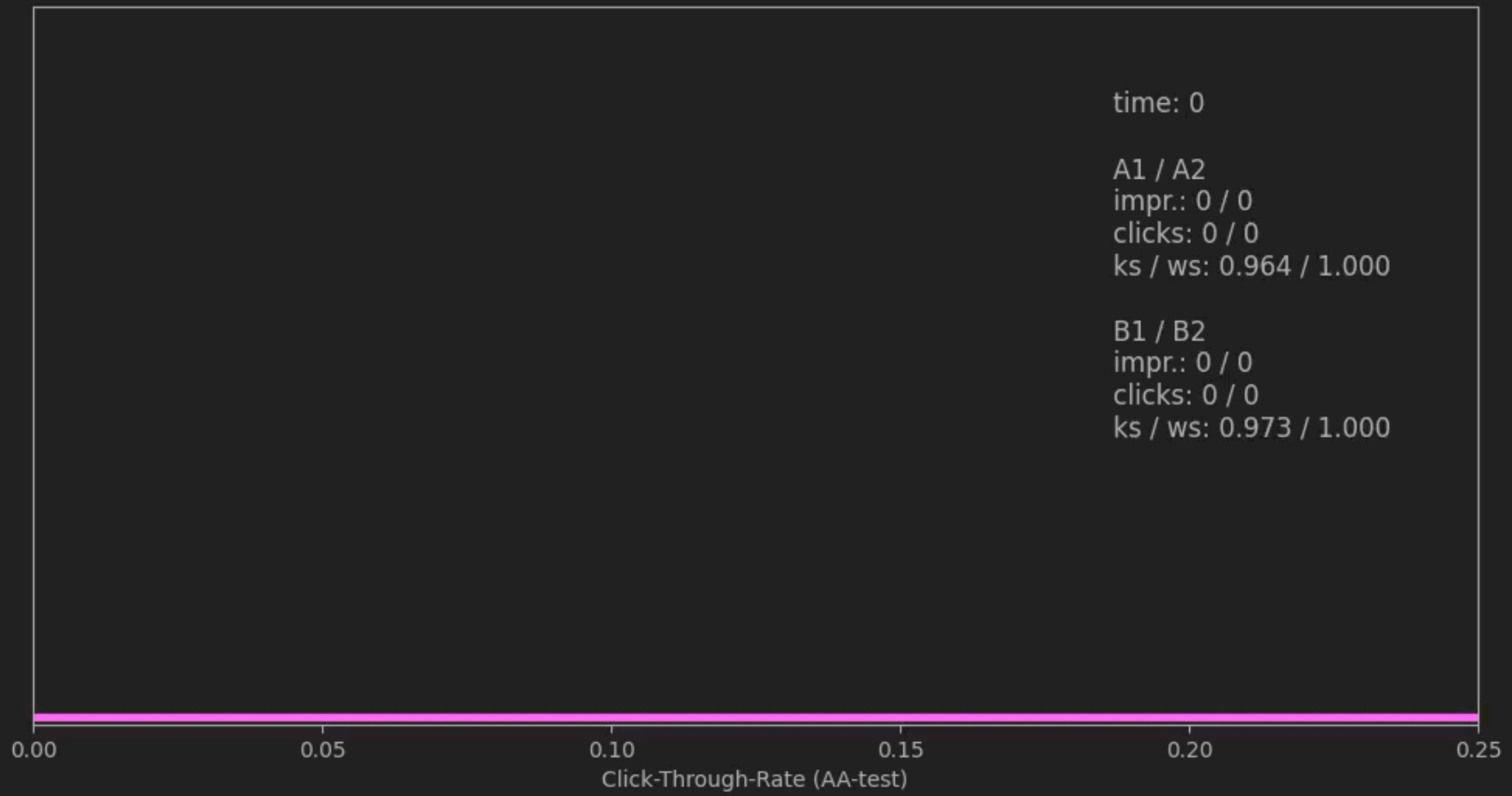
10%
CTR A

11
CPM A

B

12%
CTR B

19
CPM B



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

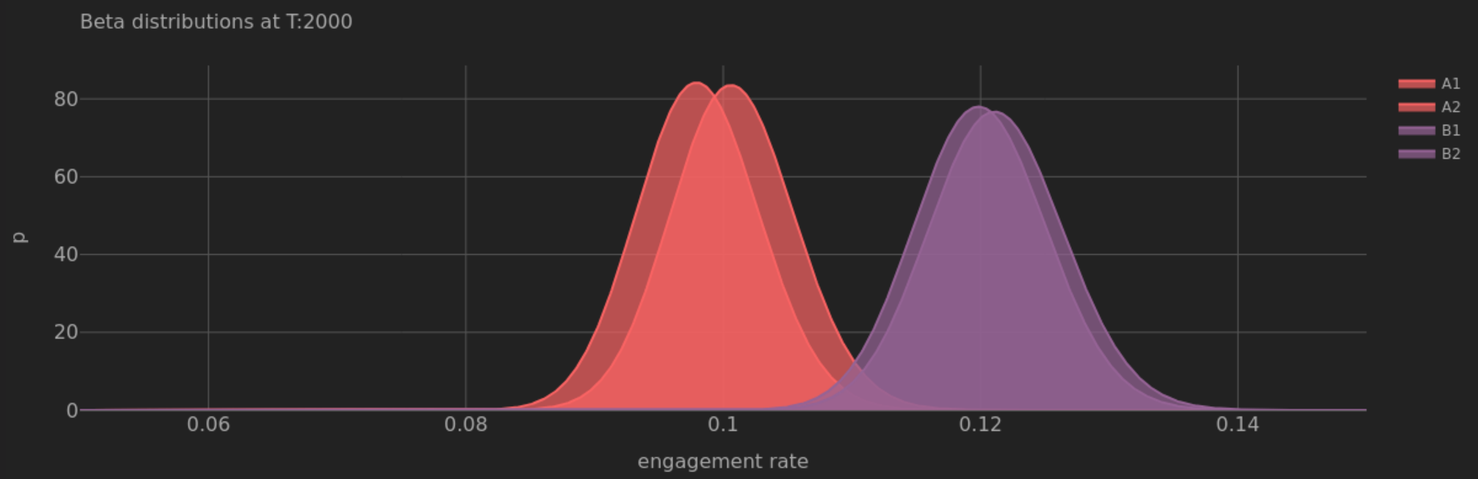
19
CPM B

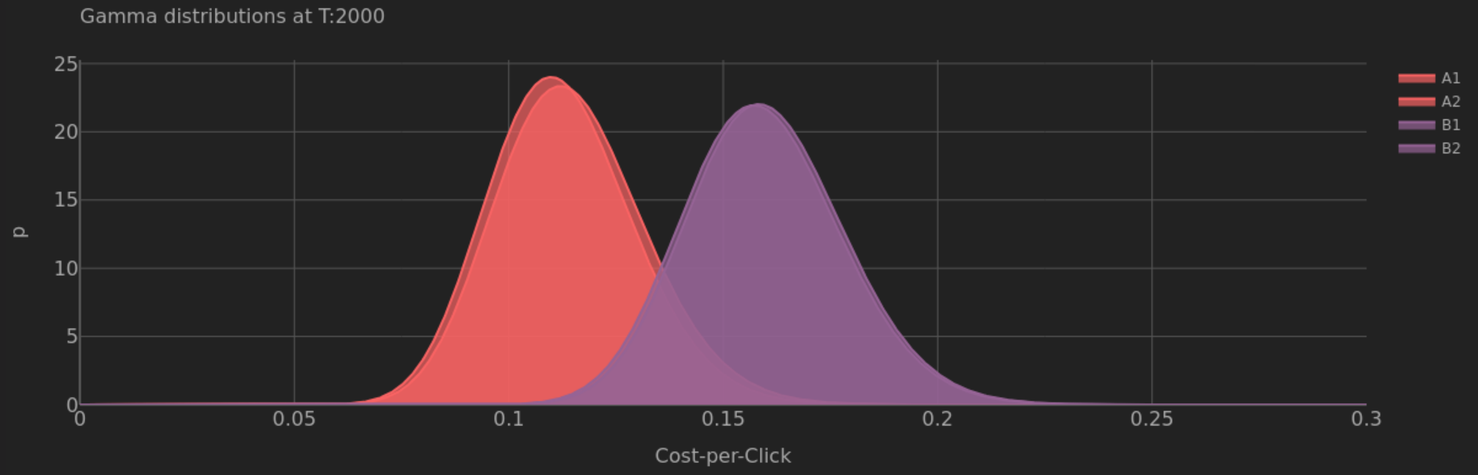
4040 / 4088
impressions

447 / 503
clicks

77 / 78
cost

End-of-simulation





Hypothesis A/A Testing - Campaign simulation

Same / Same test

A

10%
CTR A

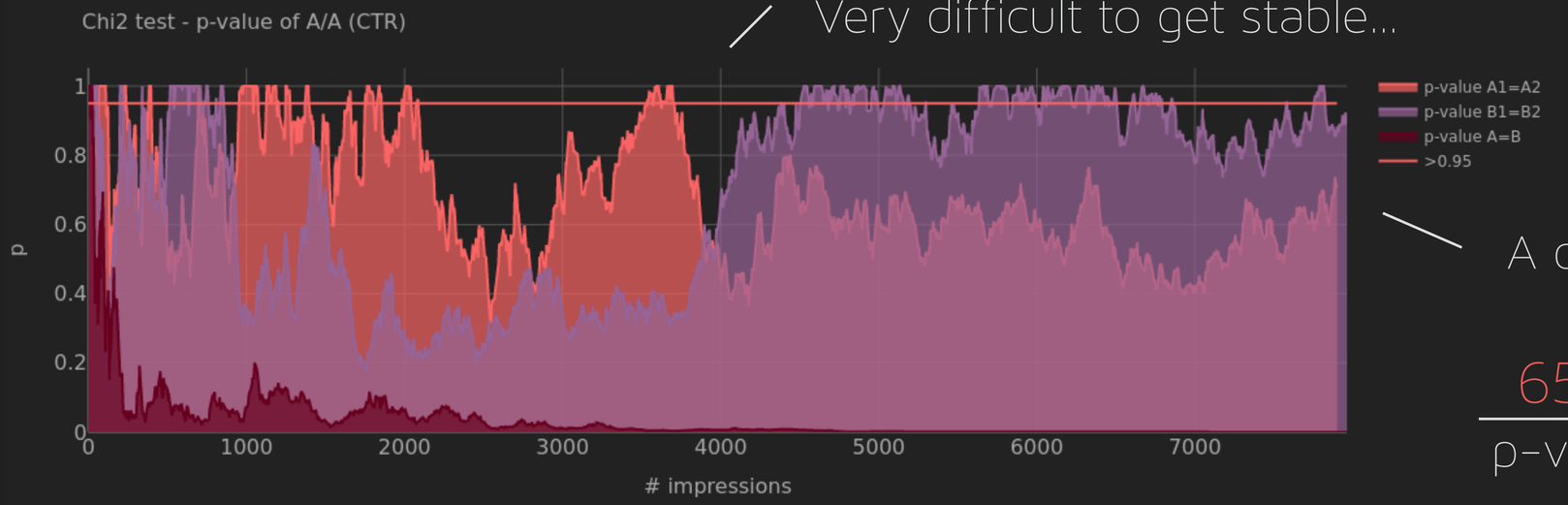
11
CPM A

B

12%
CTR B

19
CPM B

Aleatoric uncertainty
Very difficult to get stable...

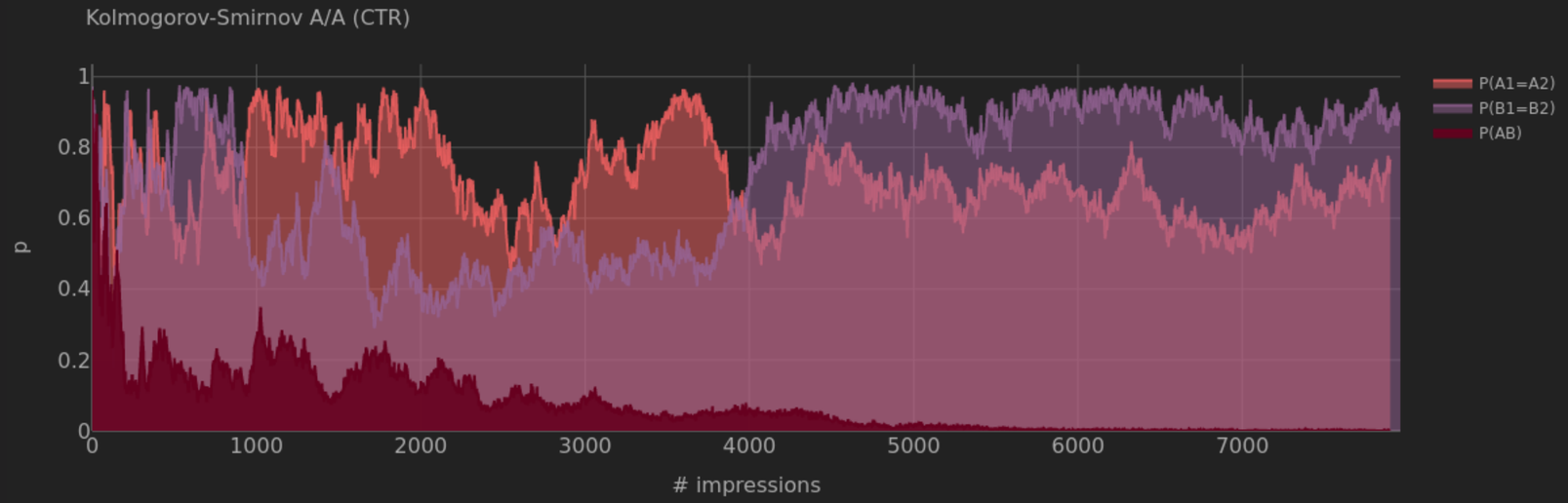


A didn't make it...

65%
p-value

97%
p-value

AA & BB are more similar than AB



Hypothesis A/A Testing - Campaign simulation

Same / Same test

A

10%
CTR A

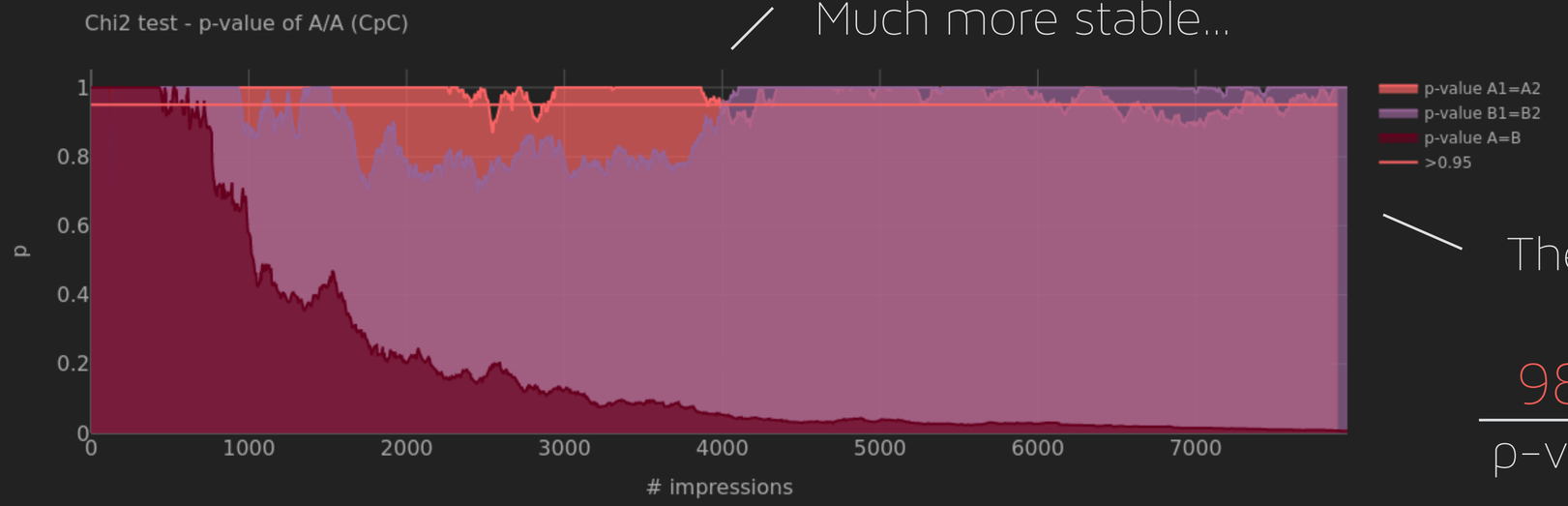
11
CPM A

B

12%
CTR B

19
CPM B

Aleatoric uncertainty
Much more stable...

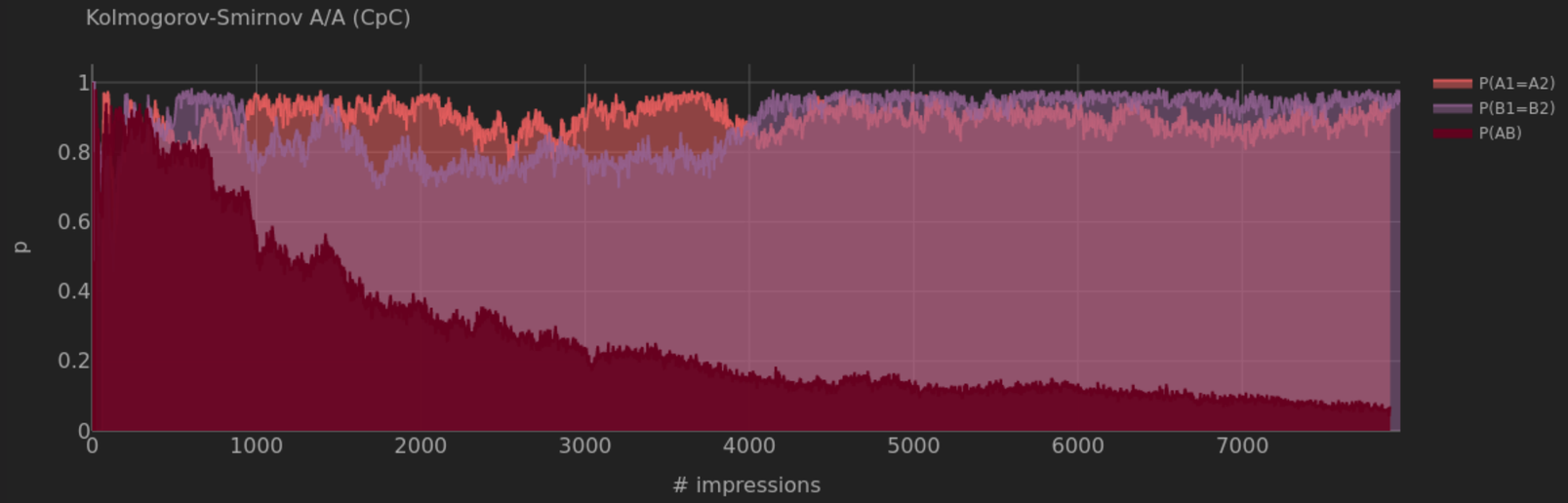


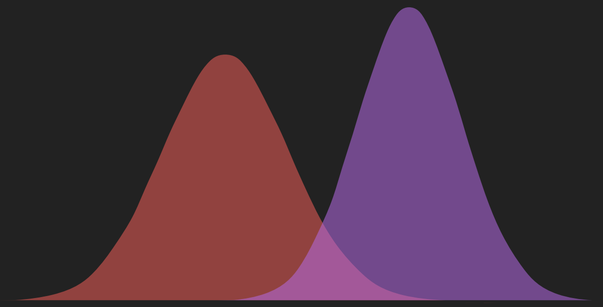
They made it...

98%
p-value

100%
p-value

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

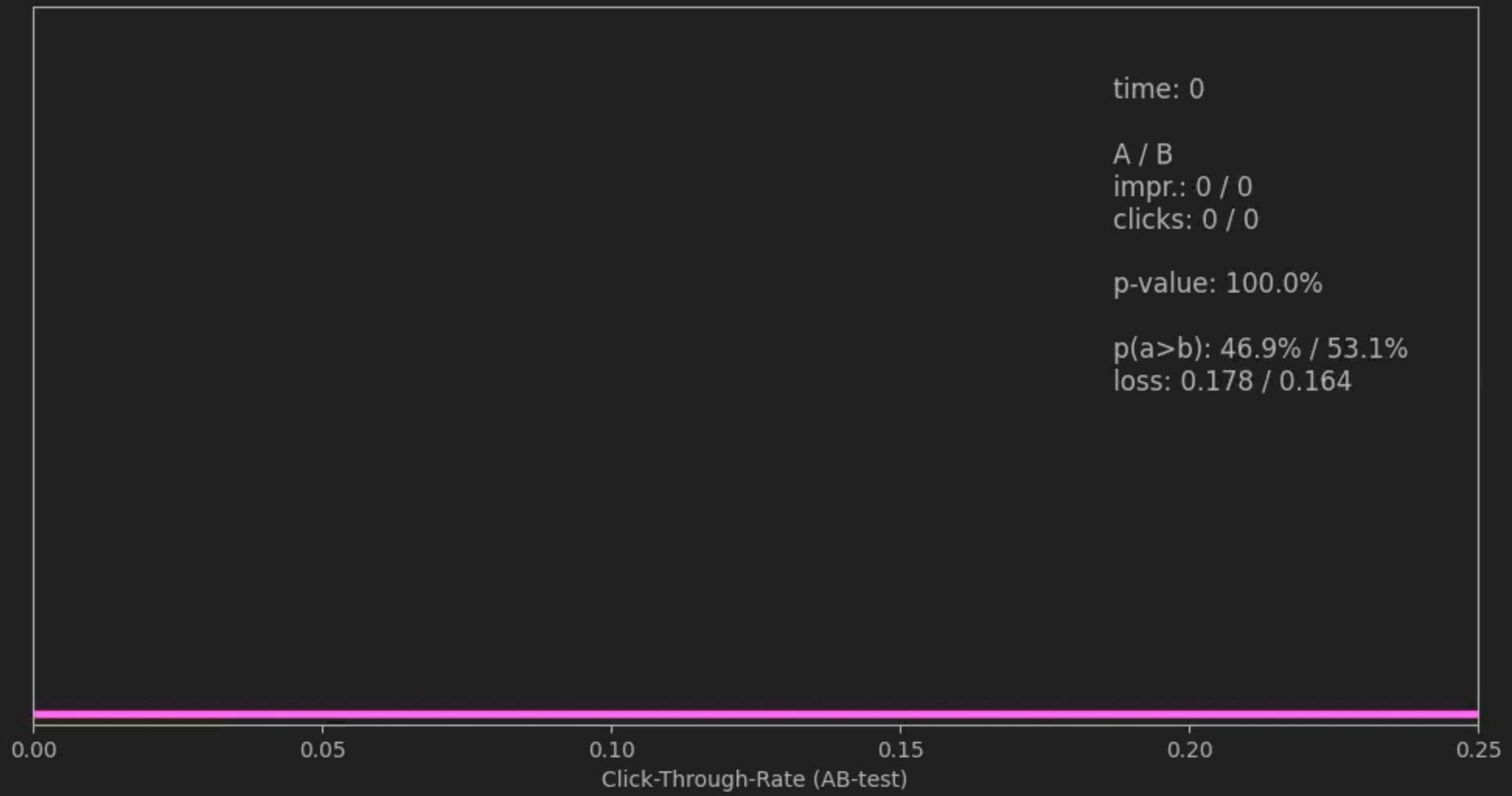
B

12%

CTR B

19

CPM B



Skip? Todo - 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

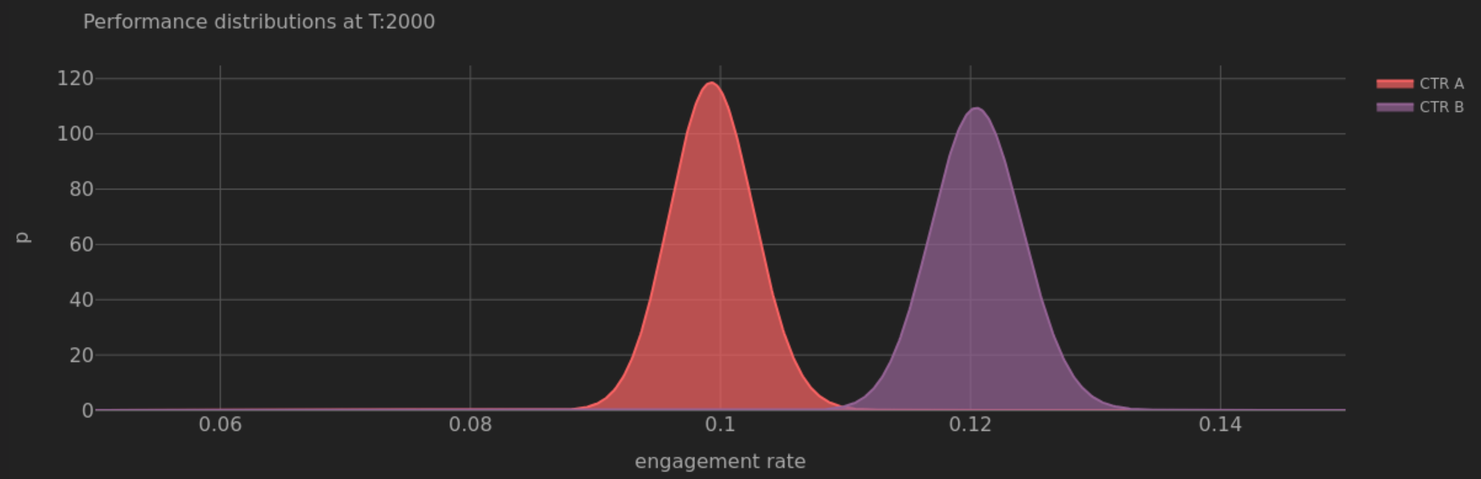
19
CPM B

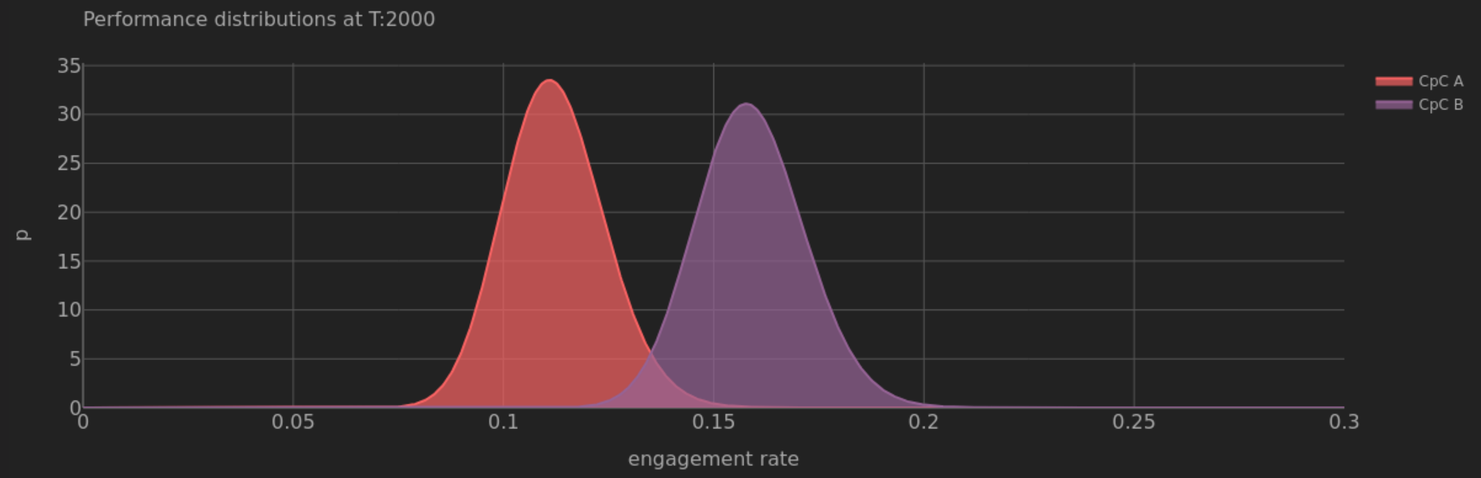
7961
impressions

959
clicks

151
cost

End-of-simulation





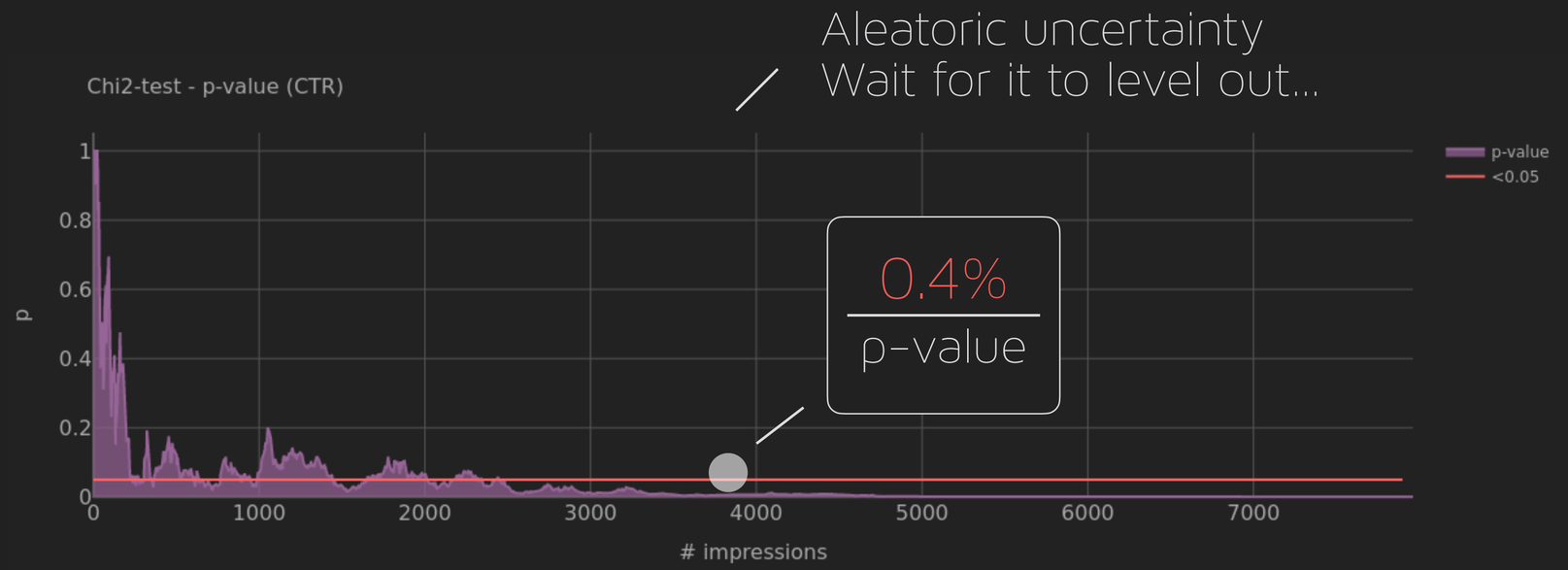
TODO - Hypothesis A/B Testing - Campaign simulation

A/B test

A

10%
CTR A

11
CPM A



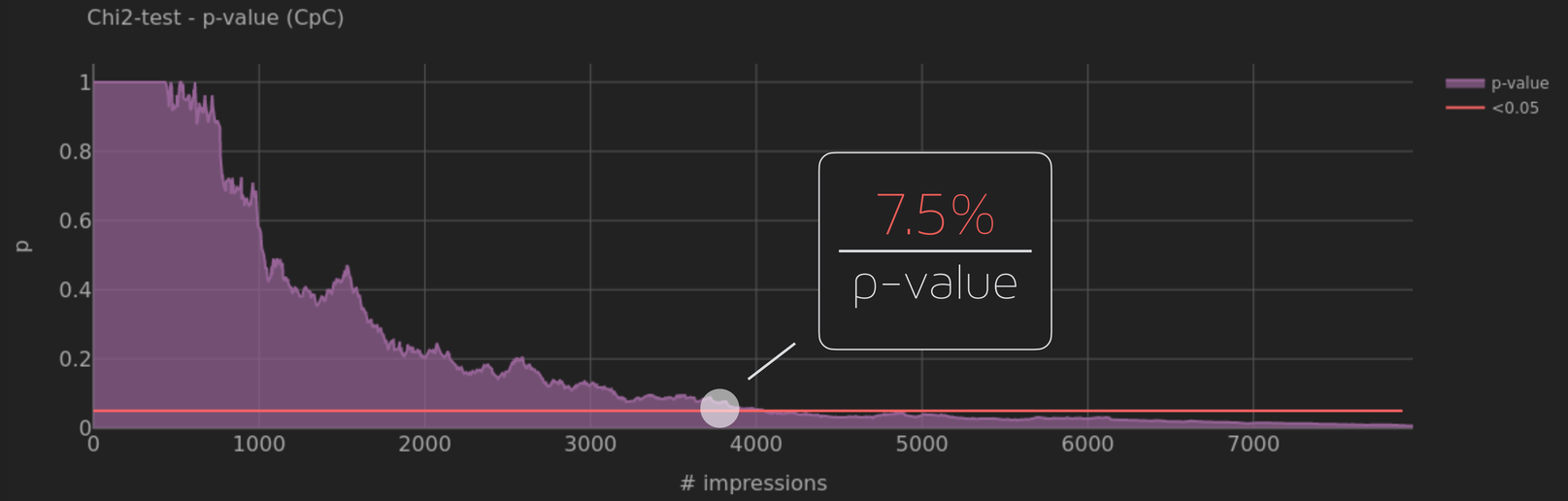
Required number of samples

3840
#samples

B

12%
CTR B

19
CPM B



Hmmm, B didn't pass...

Higher cost did not outweigh better click rate...

Bayesian A/B Testing - Campaign simulation

A/B test

A

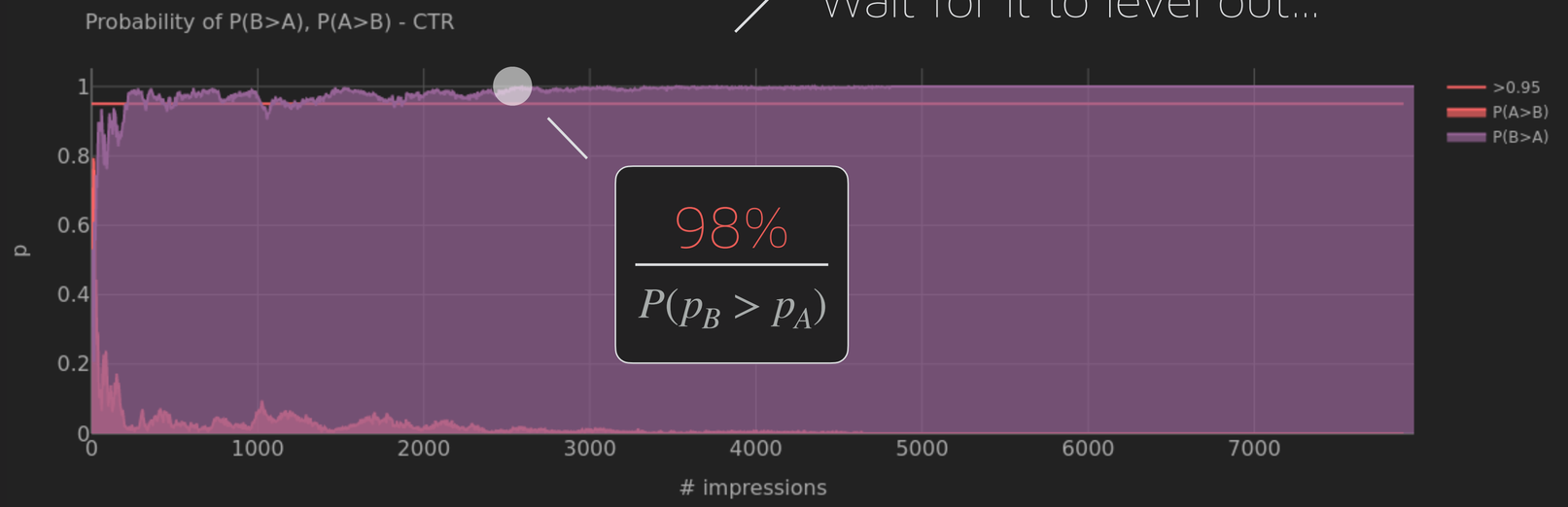
10%

CTR A

11

CPM A

Aleatoric uncertainty
Wait for it to level out...



Could we get results earlier?

~500-2500

#samples

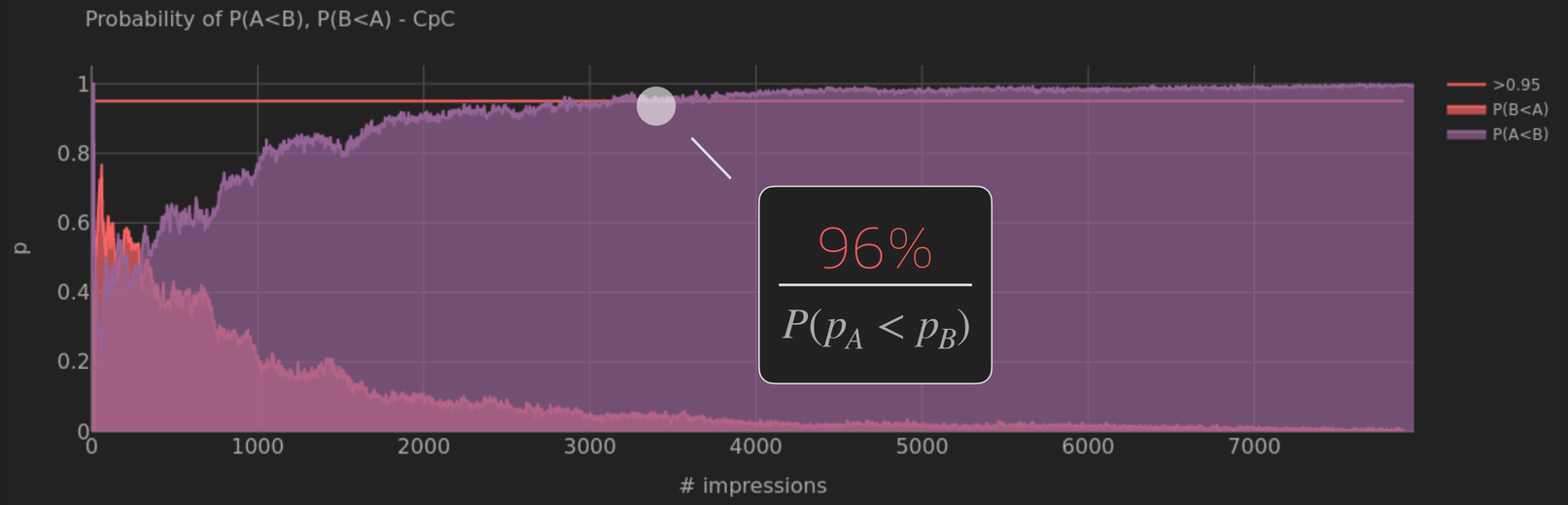
B

12%

CTR B

19

CPM B



Aleatoric uncertainty messes with us...

Hypothesis A/B Testing - Campaign simulation

A/B test

A

10%

CTR A

11

CPM A

Aleatoric uncertainty not stable, but results are clear

Choosing A is pretty expensive....

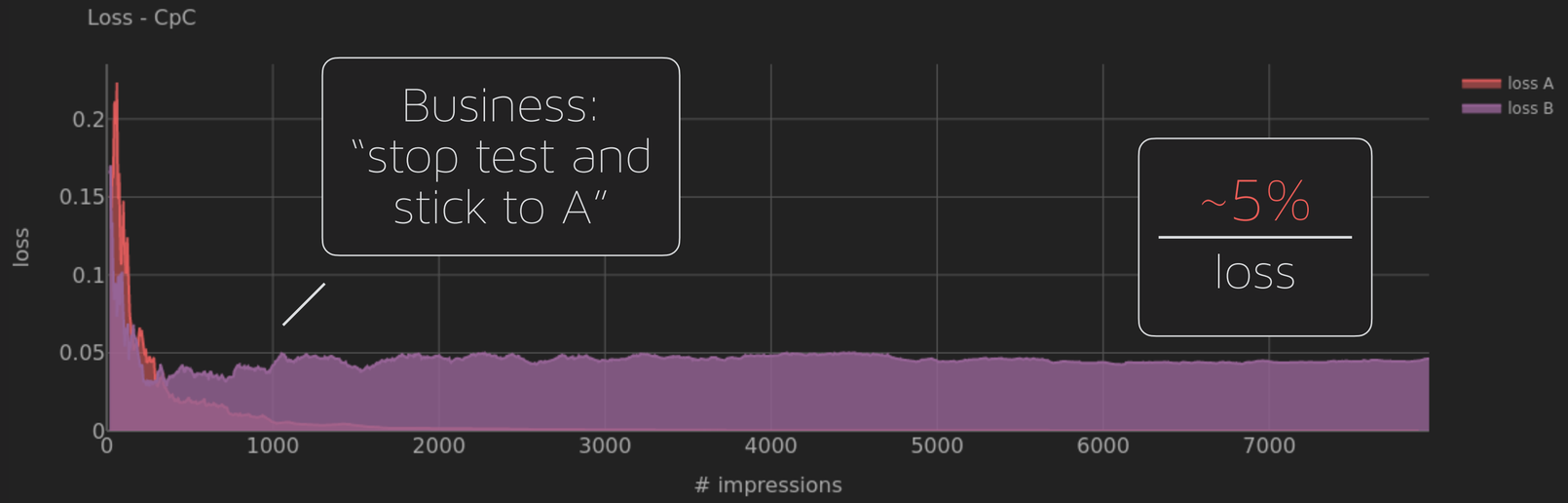
B

12%

CTR B

19

CPM B

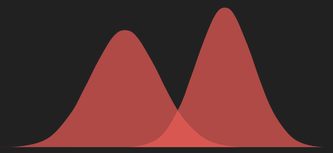


Choosing B is pretty expensive....

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”

Code