Foundations of large-scale "doubly-sequential" experimentation

(KDD tutorial in Anchorage, on 4 Aug 2019)

Aaditya Ramdas

Assistant Professor
Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

www.stat.cmu.edu/~aramdas/kdd19/

A/B-testing : tech   ::   clinical trials : pharma

A/B-testing : tech :: clinical trials : pharma

Large-scale A/B-testing or other related forms of randomized experimentation has revolutionized the tech industry in the last 15yrs.

A/B-testing : tech :: clinical trials : pharma

Large-scale A/B-testing or other related forms of randomized experimentation has revolutionized the tech industry in the last 15yrs.

In 2013, a team from Microsoft (Bing) claimed that they run tens of thousands of such experiments, leading to millions of dollars in increased revenue.

Kohavi et al. '13

A/B-testing : tech :: clinical trials : pharma

Large-scale A/B-testing or other related forms of randomized experimentation has revolutionized the tech industry in the last 15yrs.

In 2013, a team from Microsoft (Bing) claimed that they run tens of thousands of such experiments, leading to millions of dollars in increased revenue.

Much has been discussed about doing A/B testing the "right" way, both theoretically and practically in real-world systems.

Many companies contributing to this vast and growing literature.

Kohavi et al. '13

Audience poll (for the speaker)

Audience poll (for the speaker)

How many of you have written papers on A/B testing or online experimentation?
(or work in the area, or consider yourselves experts?)

Audience poll (for the speaker)

How many of you have written papers on A/B testing or online experimentation?
(or work in the area, or consider yourselves experts?)

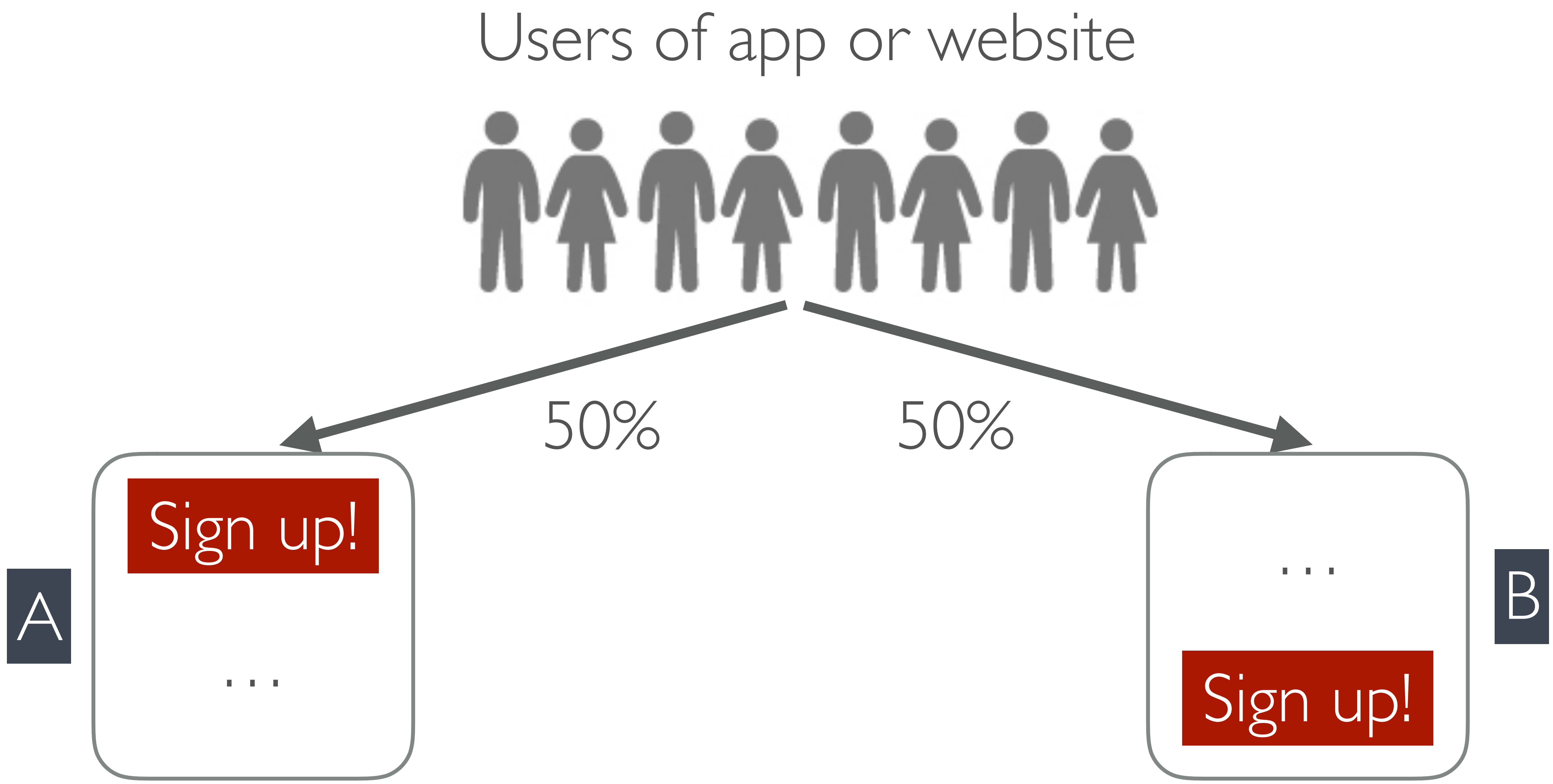
How many of you have read papers on A/B testing and know what it is, but want to know more?

Audience poll (for the speaker)

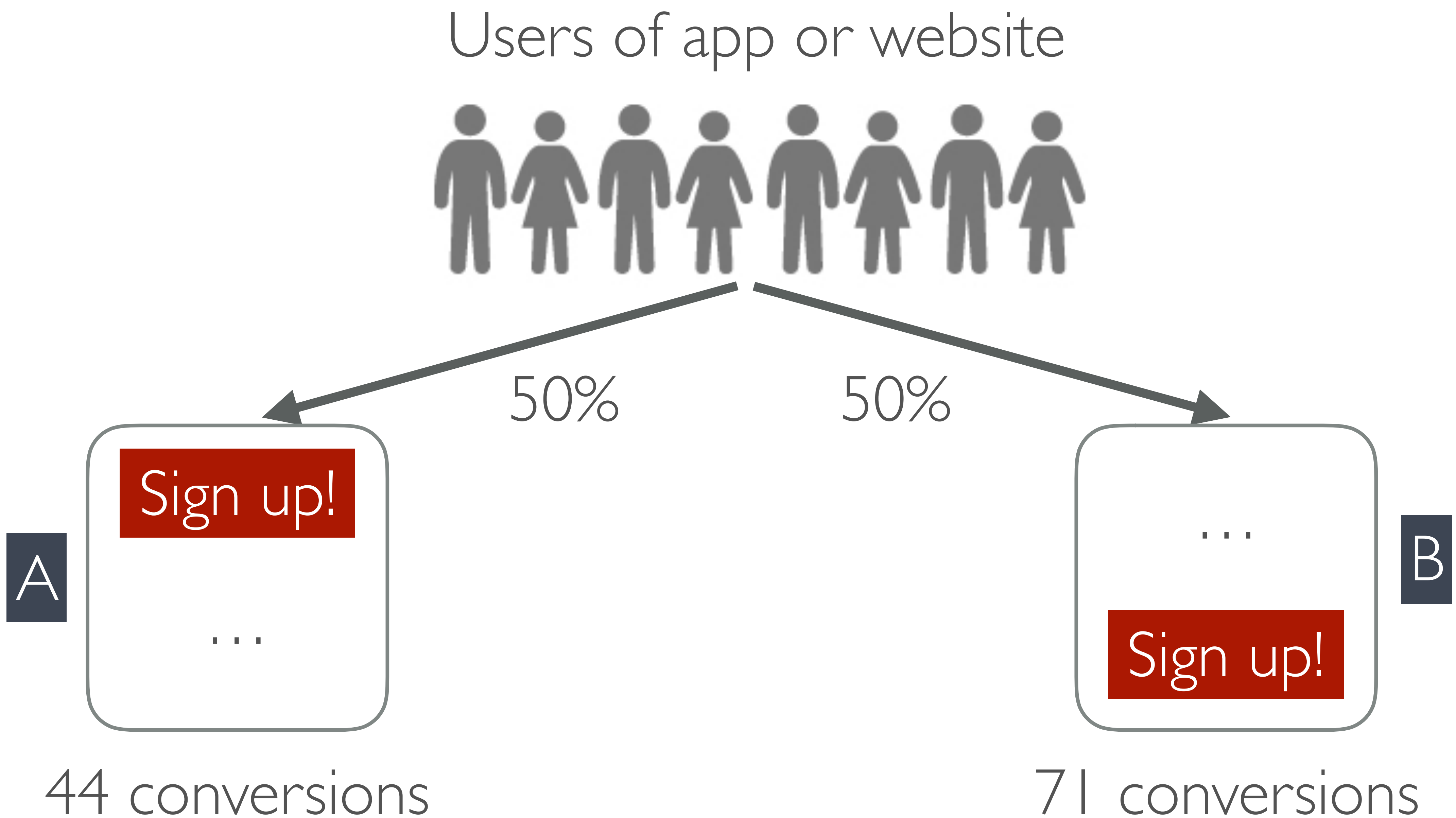
How many of you have written papers on A/B testing or online experimentation?
(or work in the area, or consider yourselves experts?)

How many of you have read papers on A/B testing and know what it is, but want to know more?

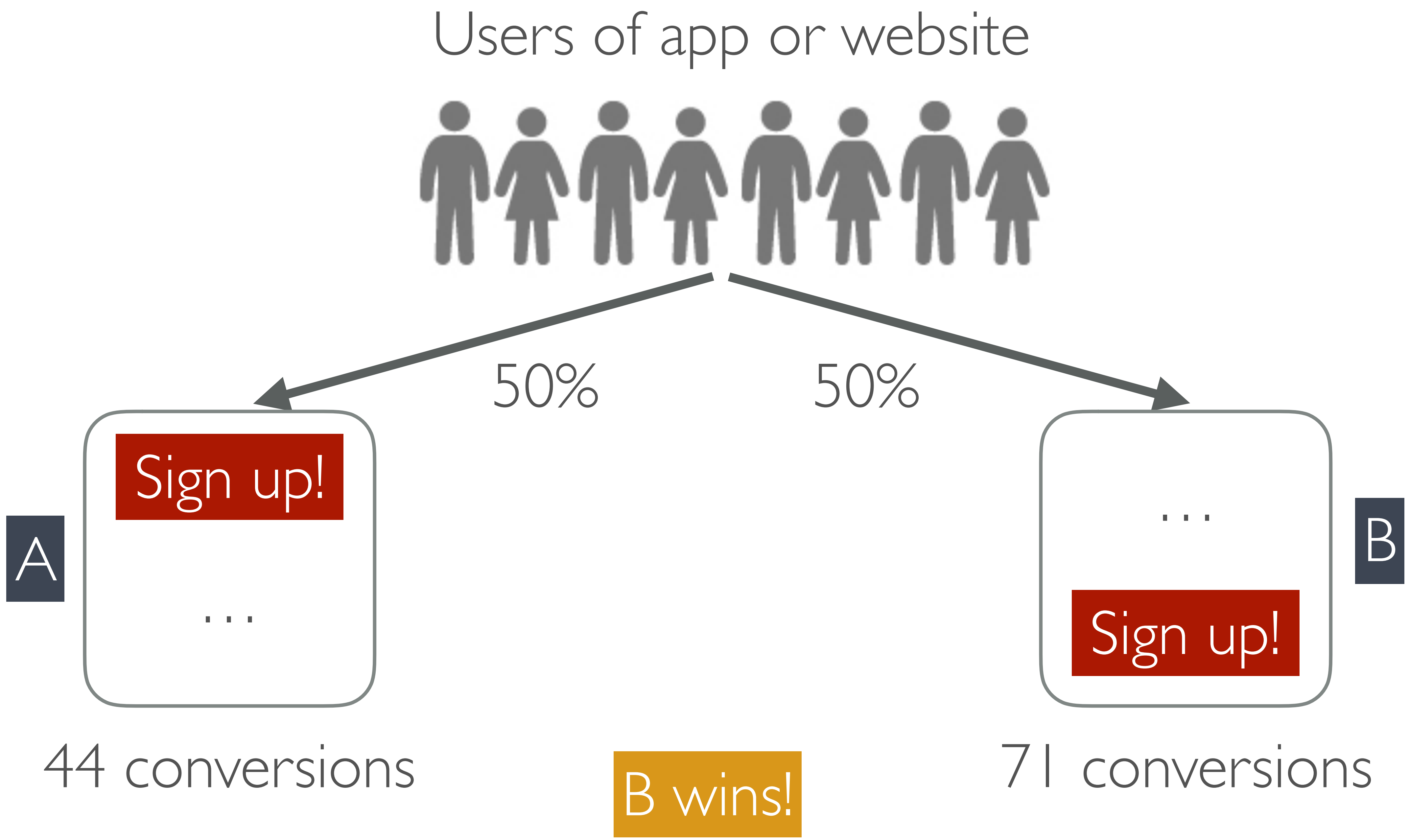
How many have no idea what I’m talking about?

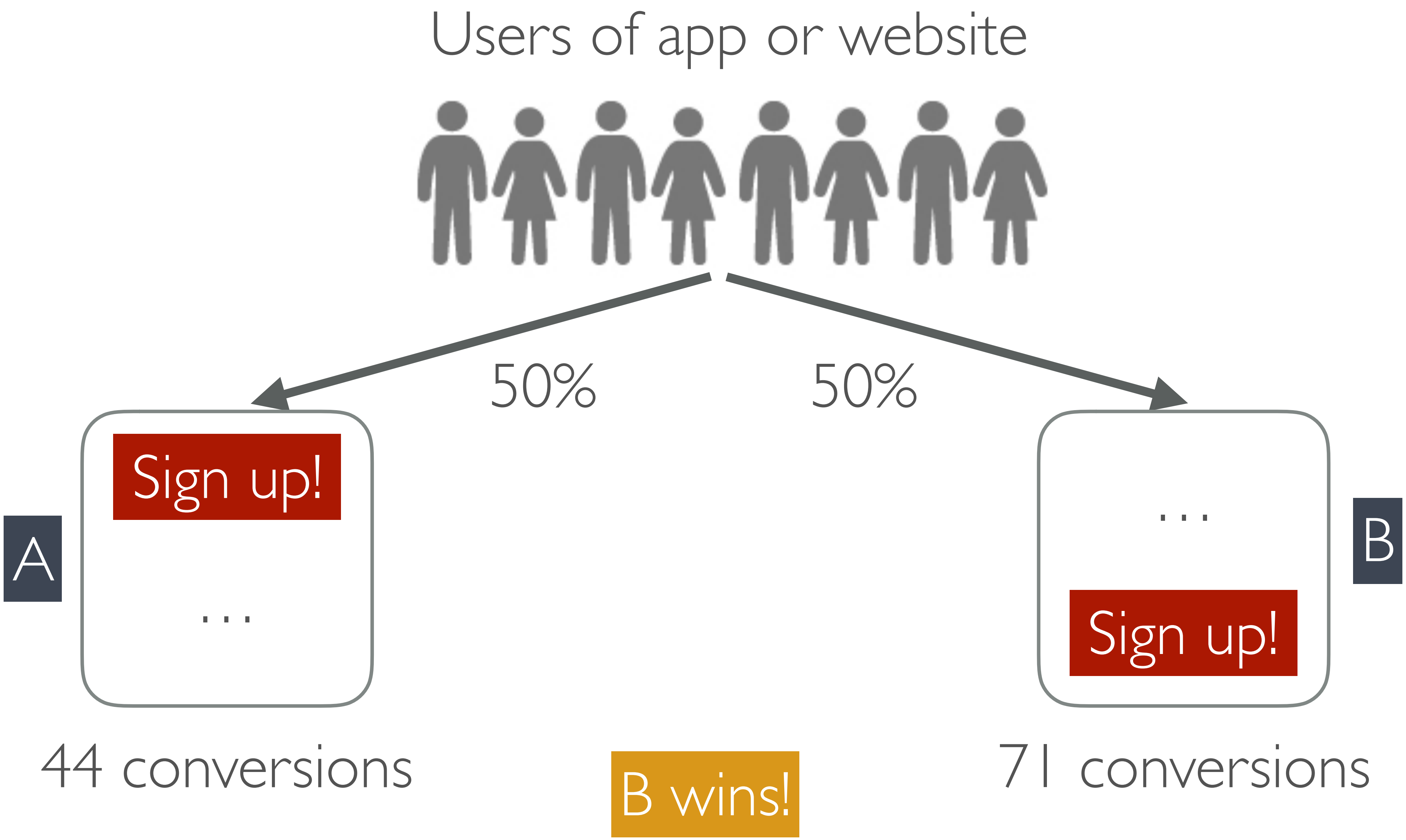


Users of app or website
50%
50%
A
Sign up!
...
B
...
Sign up!



Users of app or website
50%
50%
A
Sign up!
...
B
...
Sign up!
44 conversions
71 conversions


Users of app or website
50%
50%
A
Sign up!
...
B
...
Sign up!
44 conversions
B wins!
71 conversions


Users of app or website
50%
50%
A
Sign up!
...
B
...
Sign up!
44 conversions
B wins!
71 conversions


BUWHAHAHA!
I'M USING A–B TESTING TO MANIPULATE IRRATIONAL HUMANS!
Dilbert.com DilbertCartoonist@gmail.com
BEND TO MY WILL AND CHOOSE THE ORANGE BUTTON, YOU MINDLESS CLICK–PUPPETS!
AND THIS IS LEGAL?
I OWN YOU NOW!

What we will NOT cover today

What we will NOT cover today

Pre-experiment analysis and difference-in-differences

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments
Deployment with controlled, phased rollouts

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments
Deployment with controlled, phased rollouts

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments
Deployment with controlled, phased rollouts

Pitfalls of long experiments (survivorship bias, perceived trends)

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments
Deployment with controlled, phased rollouts

Pitfalls of long experiments (survivorship bias, perceived trends)
ML meets causal inference meets online experiments

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments
Deployment with controlled, phased rollouts

Pitfalls of long experiments (survivorship bias, perceived trends)
ML meets causal inference meets online experiments
Experimentation in marketplaces or with network effects

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments
Deployment with controlled, phased rollouts

Pitfalls of long experiments (survivorship bias, perceived trends)
ML meets causal inference meets online experiments
Experimentation in marketplaces or with network effects
Ethical aspects of running experiments

What we will NOT cover today

Pre-experiment analysis and difference-in-differences
What makes a good metric? (directionality and sensitivity)
Combining metrics into an overall evaluation criterion
Time-series aspects: dealing with periodicity and trends
Causal inference in observational studies

Parametric methods for A/B testing (like SPRT and variants)
Bayesian A/B testing
Side effects and risks associated with running experiments
Deployment with controlled, phased rollouts

Pitfalls of long experiments (survivorship bias, perceived trends)
ML meets causal inference meets online experiments
Experimentation in marketplaces or with network effects
Ethical aspects of running experiments

There are many resources for these topics

Yandex tutorial at The Web Conference '18

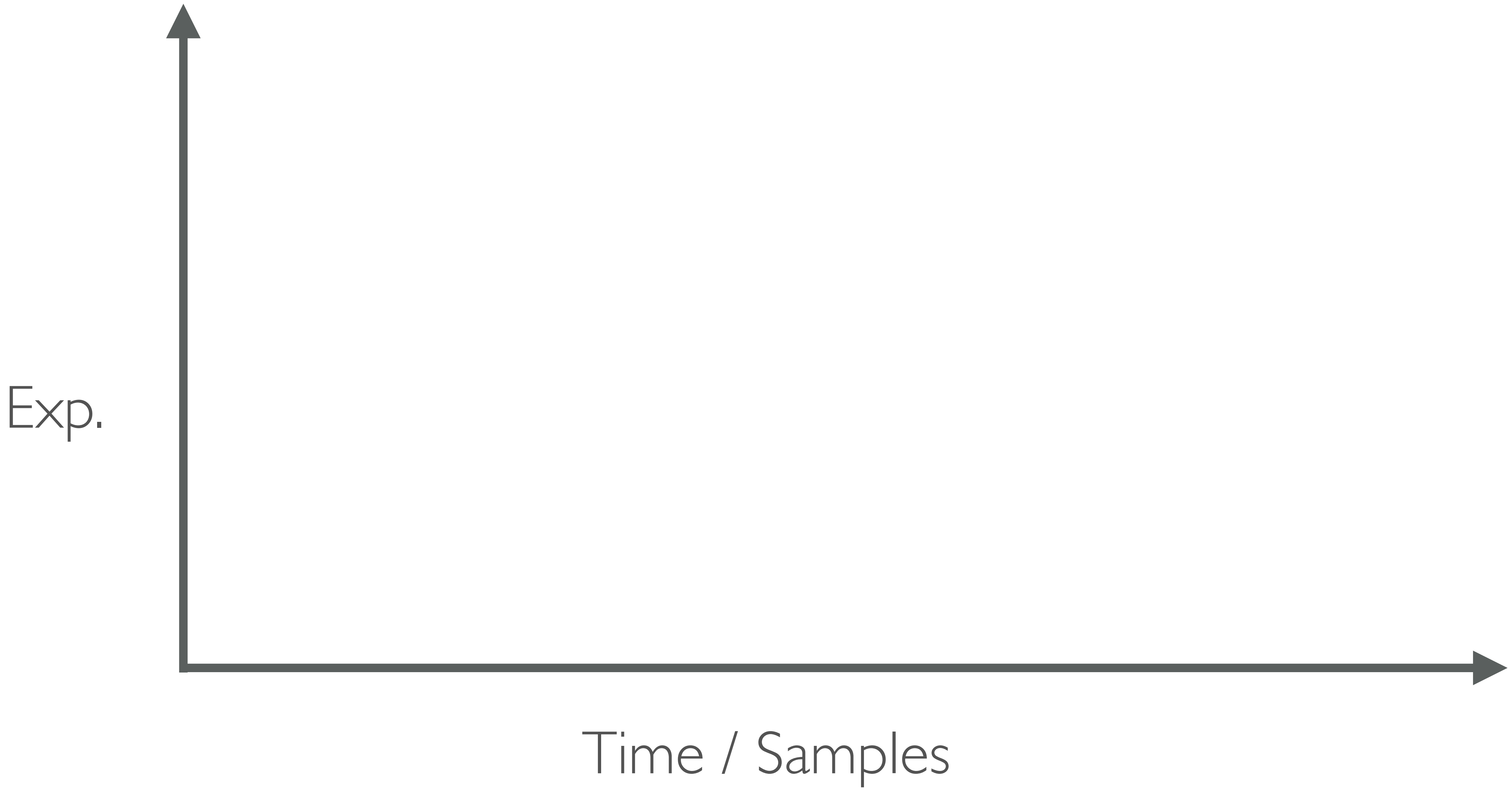
Microsoft tutorial at The Web Conference '19
(+ ExP Platform webpage)

Blog posts by Evan Miller, Etsy, Optimizely, etc.

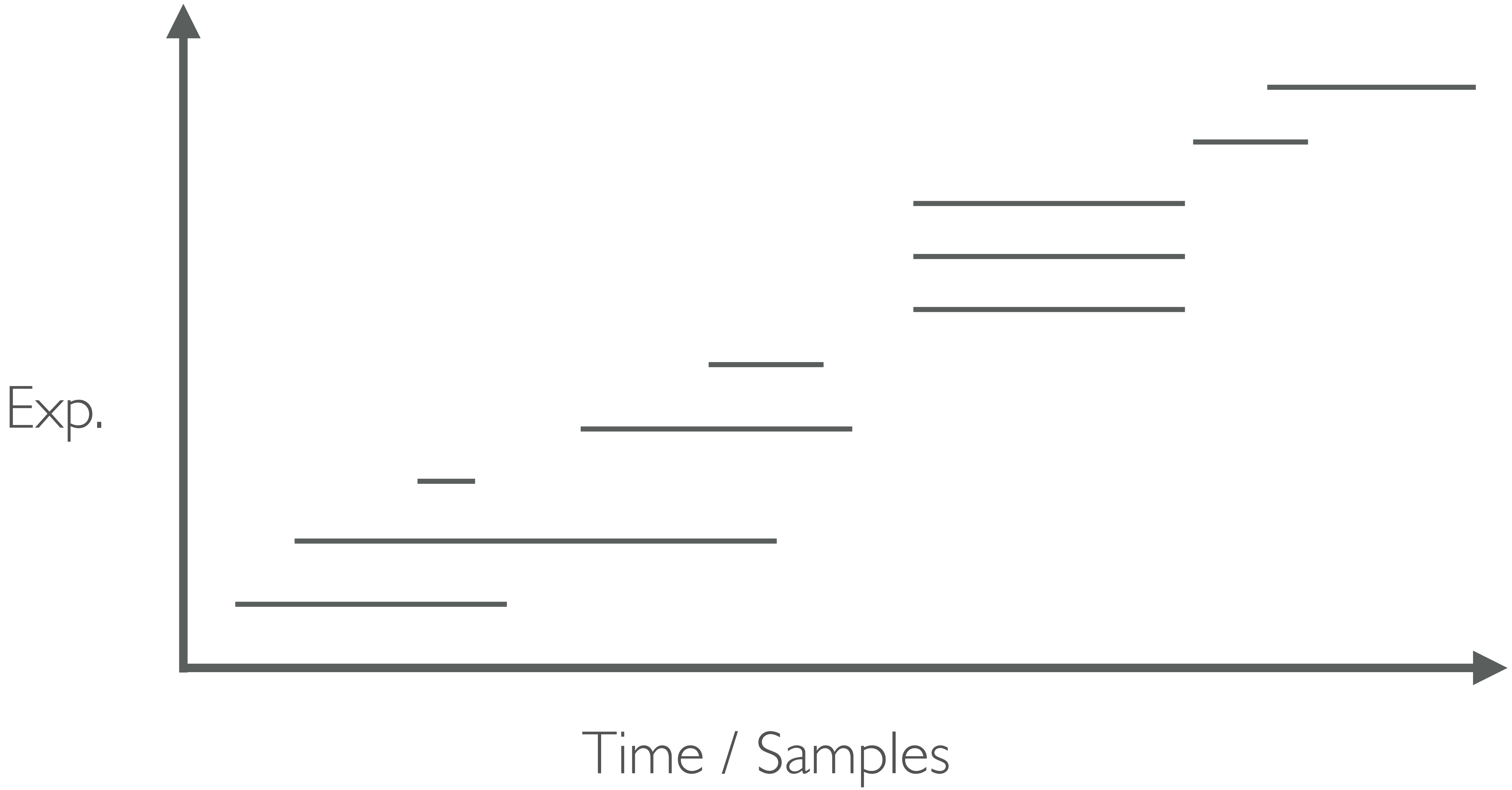
. . .

A new “doubly-sequential” perspective: a sequence of sequential experiments

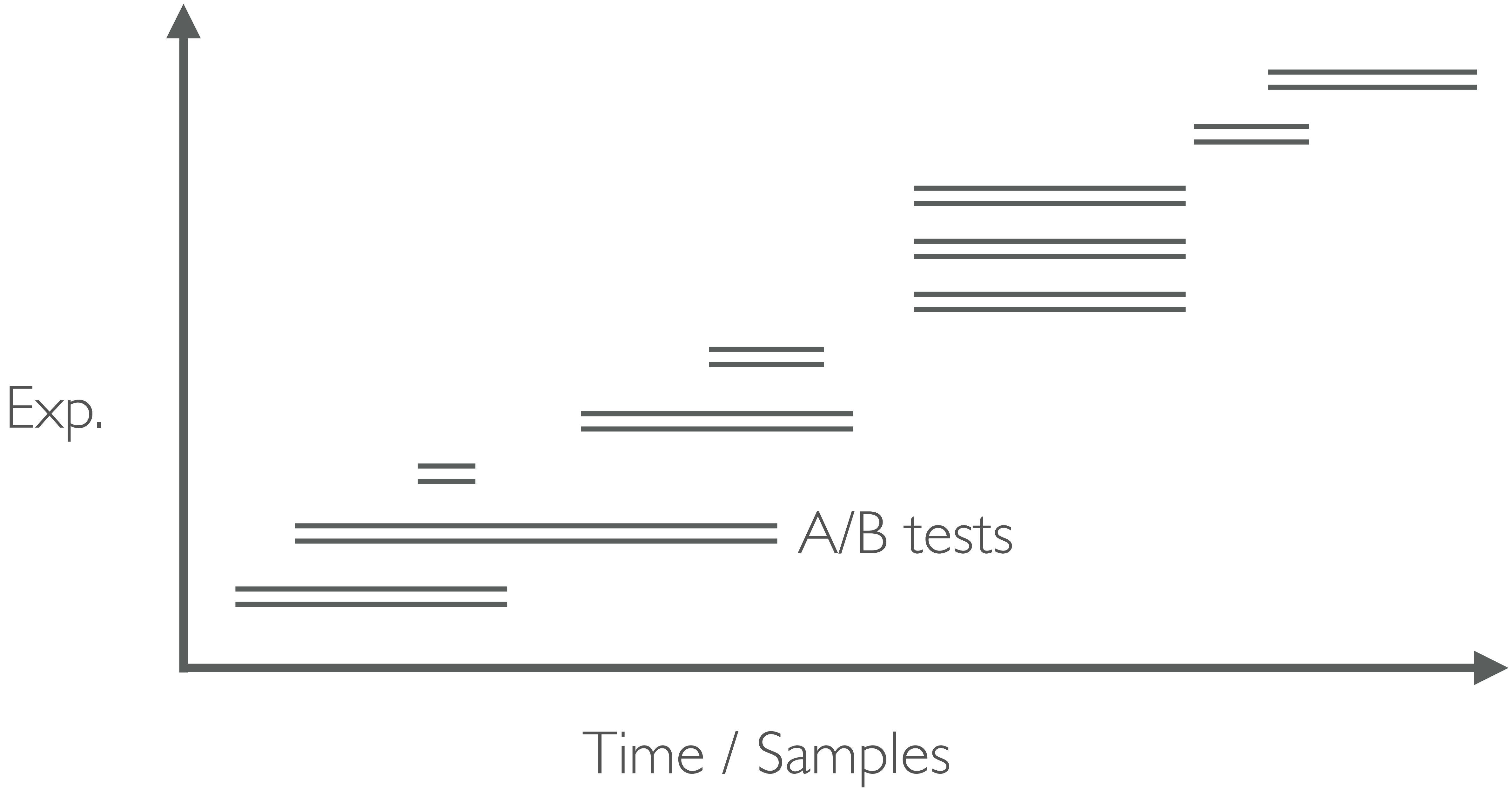
A new “doubly-sequential” perspective: a sequence of sequential experiments



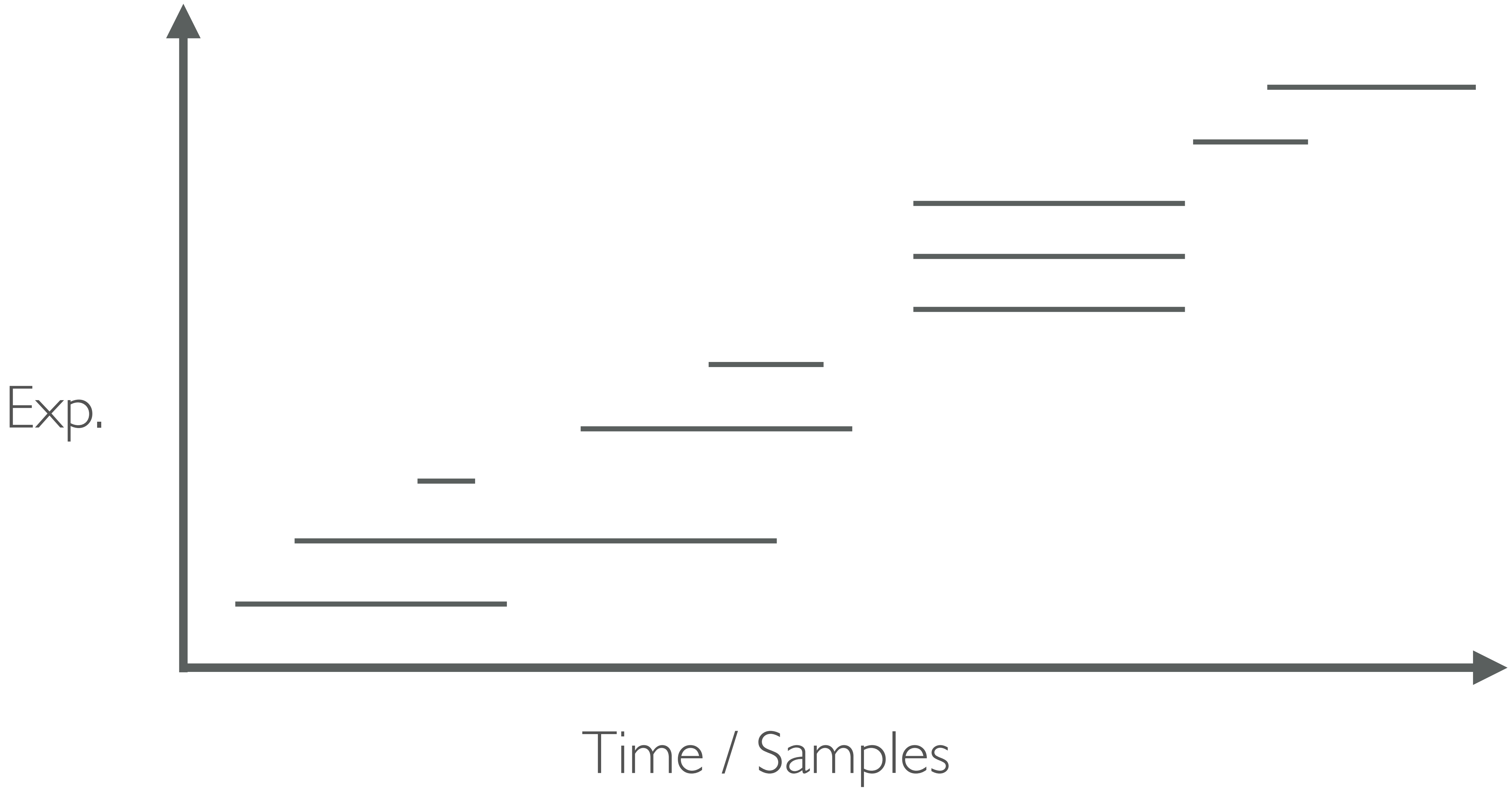
A new "doubly-sequential" perspective: a sequence of sequential experiments



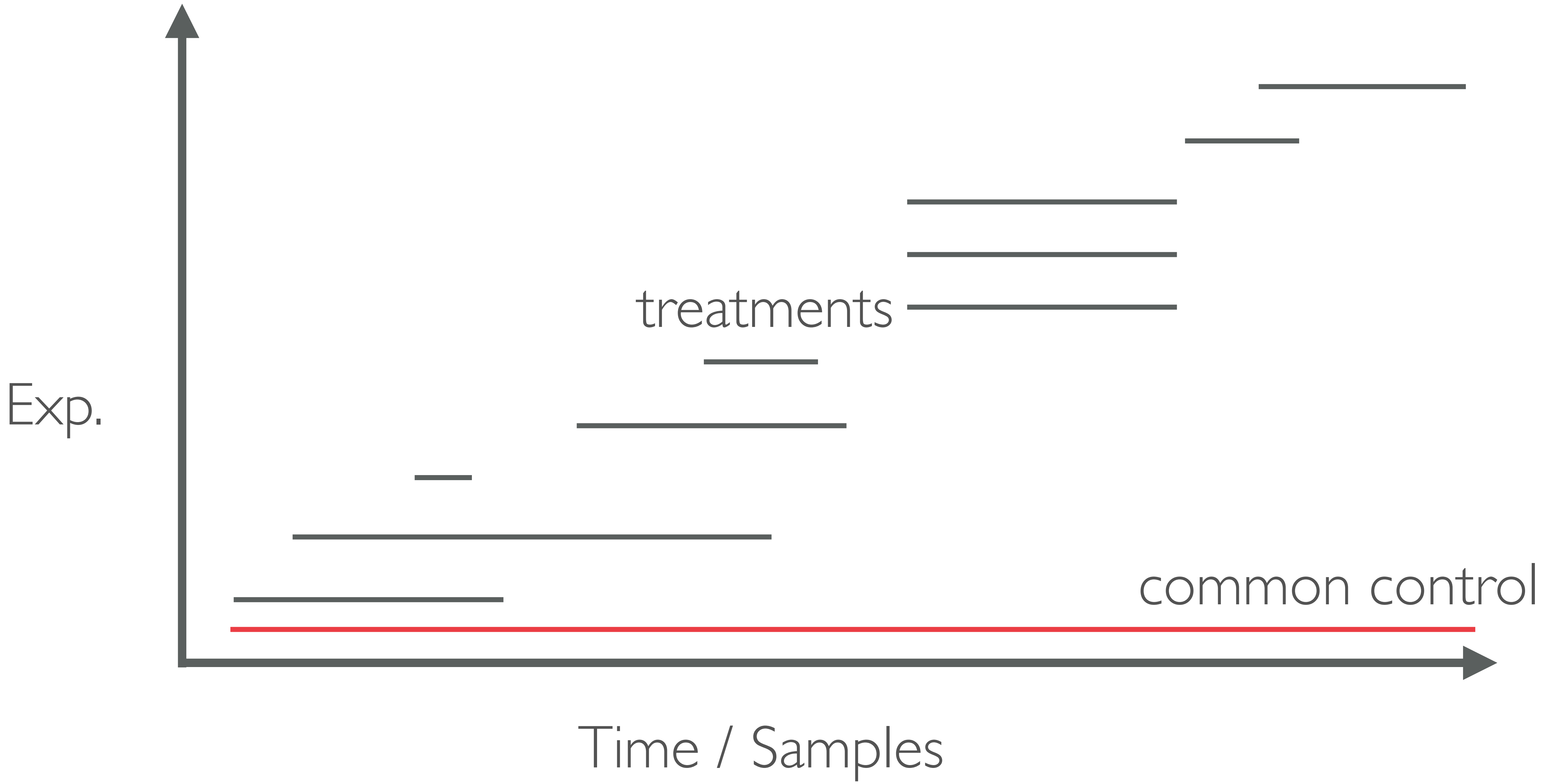
A new "doubly-sequential" perspective: a sequence of sequential experiments



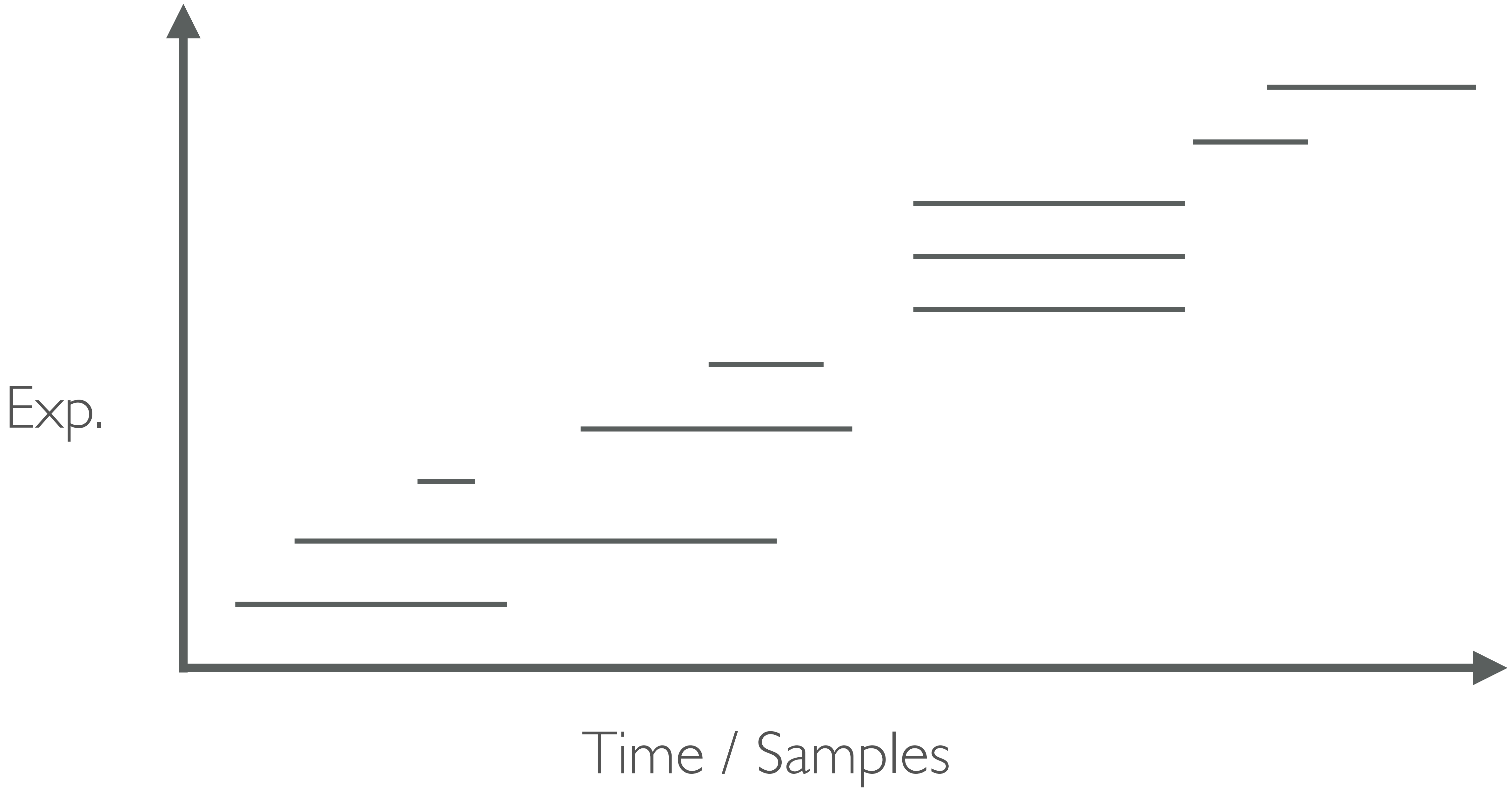
A new "doubly-sequential" perspective: a sequence of sequential experiments



A new "doubly-sequential" perspective: a sequence of sequential experiments



A new "doubly-sequential" perspective: a sequence of sequential experiments



Zrnic, Ramdas, Jordan '18
Yang, Ramdas, Jamieson, Wainwright '17

What kind of guarantees would we like for doubly sequential experimentation?

What kind of guarantees would we like for doubly sequential experimentation?

(a) ***inner sequential process (a single experiment)***

— correct inference when experiment ends (correct p-values for A/B test or correct confidence intervals for treatment effect)

What kind of guarantees would we like for doubly sequential experimentation?

(a) ***inner sequential process (a single experiment)***

— correct inference when experiment ends (correct p-values for A/B test or correct confidence intervals for treatment effect)

(b) ***outer sequential process (multiple experiments)***

— less clear (is error control on inner process enough?!)

Some existing problems in practice

Some potential issues within each experiment

Some potential issues across experiments

Many other concerns as well

Some existing problems in practice

Some potential issues within each experiment

(a) *continuous monitoring*

(b) *flexible experiment horizon*

(c) *arbitrary stopping (or continuation) rules*

Some potential issues across experiments

Many other concerns as well

Some existing problems in practice

Some potential issues within each experiment

(a) *continuous monitoring*

(b) *flexible experiment horizon*

(c) *arbitrary stopping (or continuation) rules*

Some potential issues across experiments

(a) *selection bias (multiplicity)*

(b) *dependence across experiments*

(c) *don't know future outcomes*

Many other concerns as well

Solutions for these issues

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Outer sequential process:

Part II

"false coverage rate" for estimation

(correspondingly, "false discovery rate" for testing)

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Outer sequential process:

Part II

"false coverage rate" for estimation

(correspondingly, "false discovery rate" for testing)

Modular solutions: fit well together

Many extensions to each piece

Part III

Part I

The INNER Sequential Process
(a single experiment)

[1 hour]

The “duality” between confidence intervals and p-values

Hypothesis testing is like
stochastic proof by contradiction.

Hypothesis testing is like stochastic proof by contradiction.

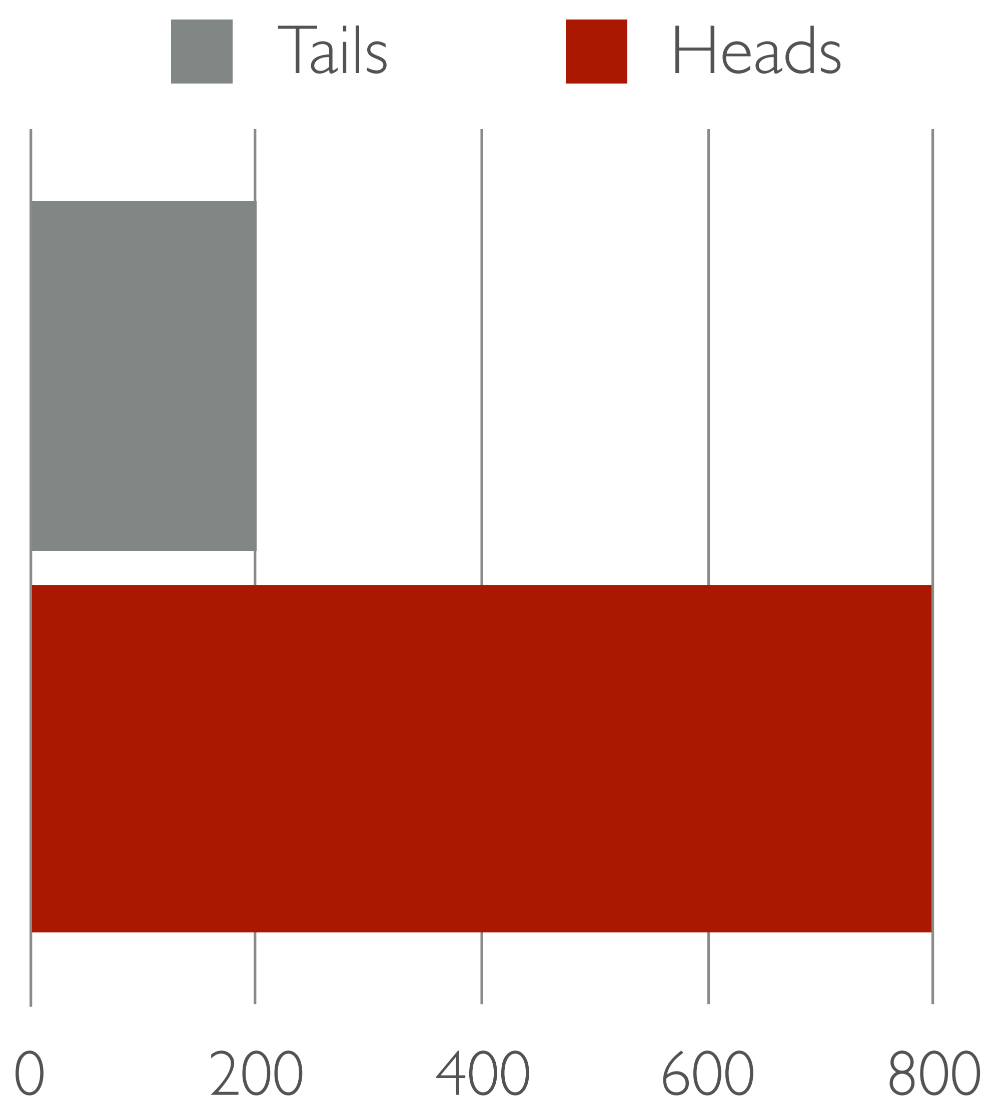
Hypothesis testing is like stochastic proof by contradiction.

Null hypothesis:
The coin is fair (bias = 0)

Alternative:
Coin is biased towards H

Hypothesis testing is like stochastic proof by contradiction.

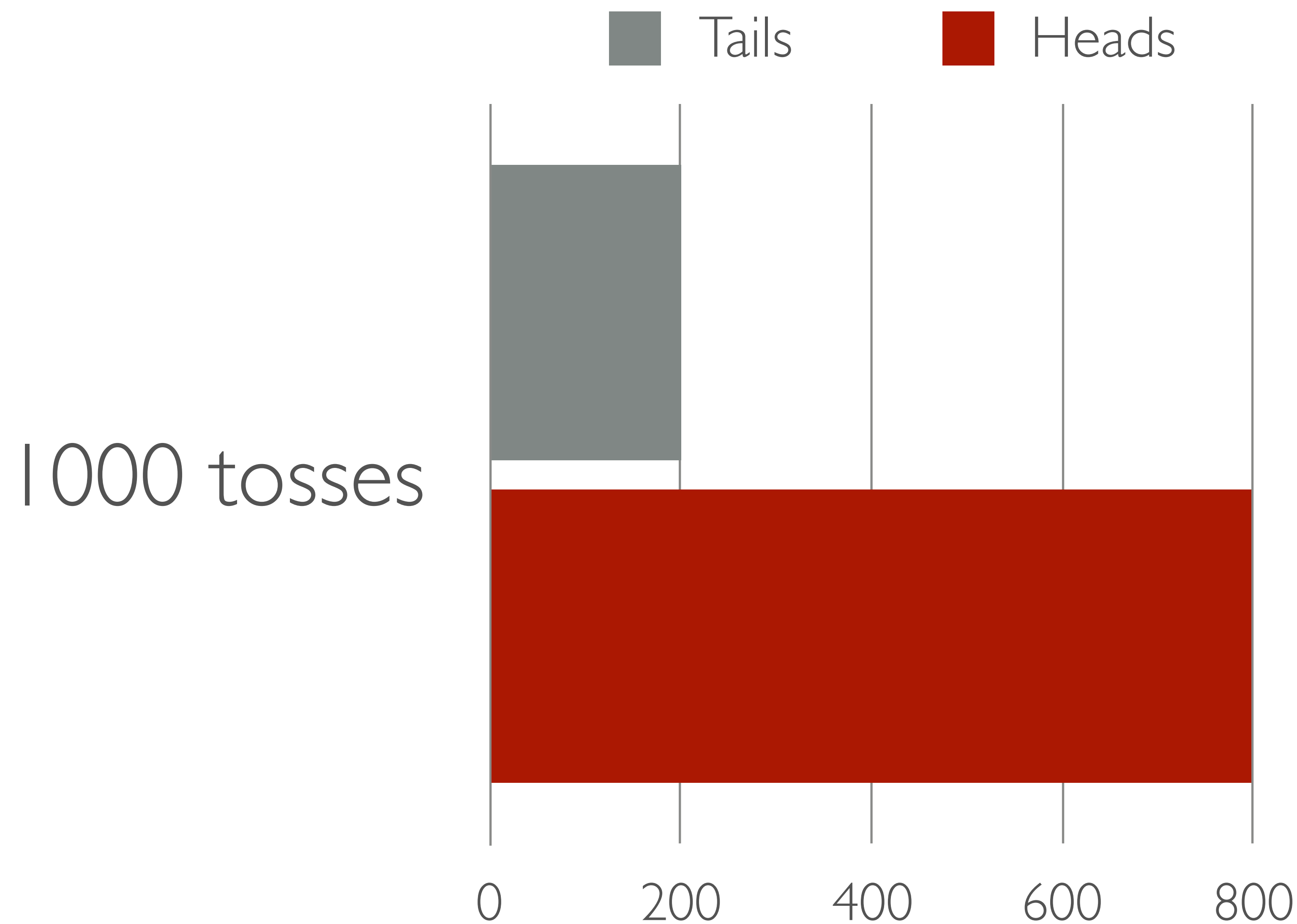
1000 tosses



Null hypothesis:
The coin is fair (bias = 0)

Alternative:
Coin is biased towards H

Hypothesis testing is like stochastic proof by contradiction.



Null hypothesis:
The coin is fair (bias = 0)

Alternative:
Coin is biased towards H

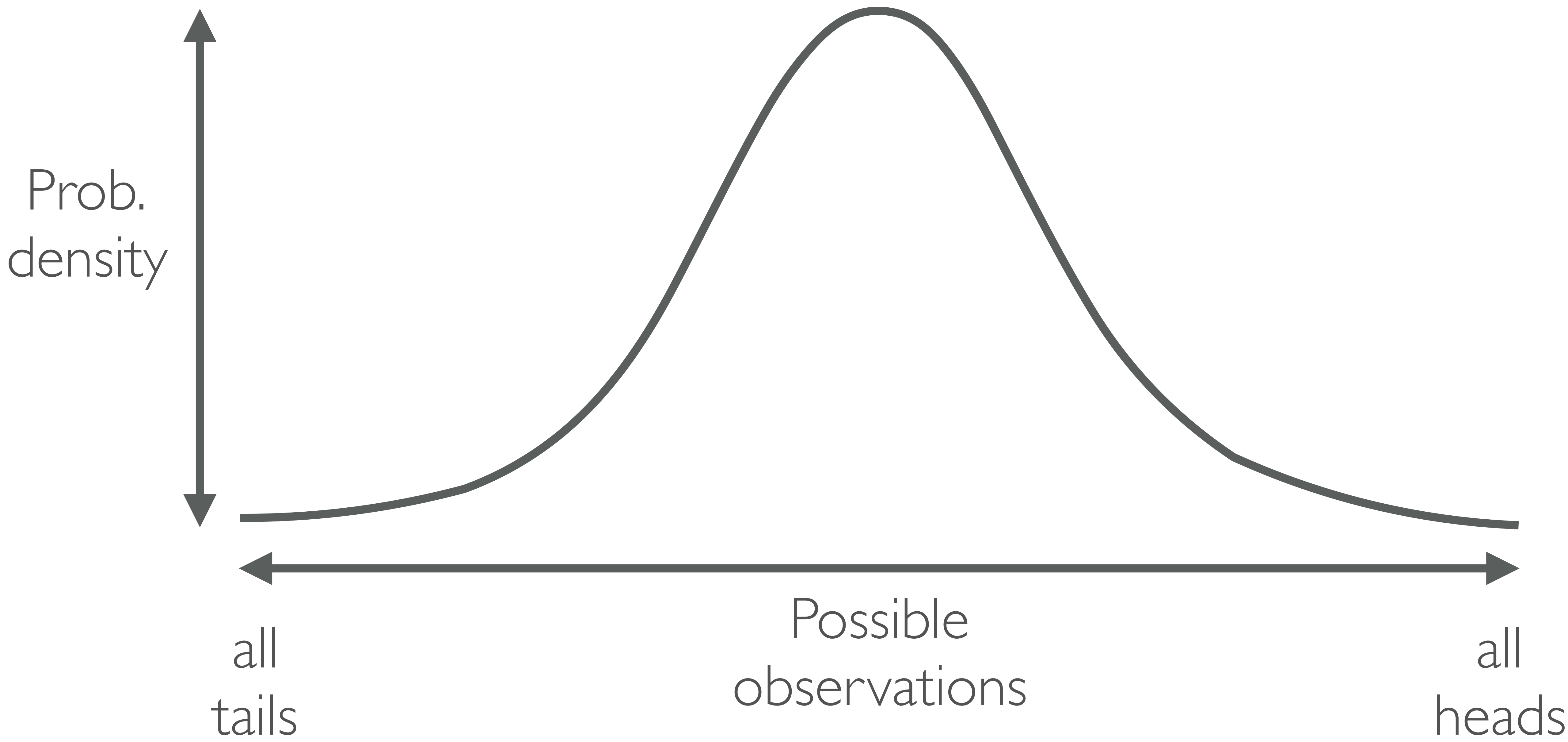
Apparent contradiction!
Should we reject the null hypothesis?

Calculate p-value

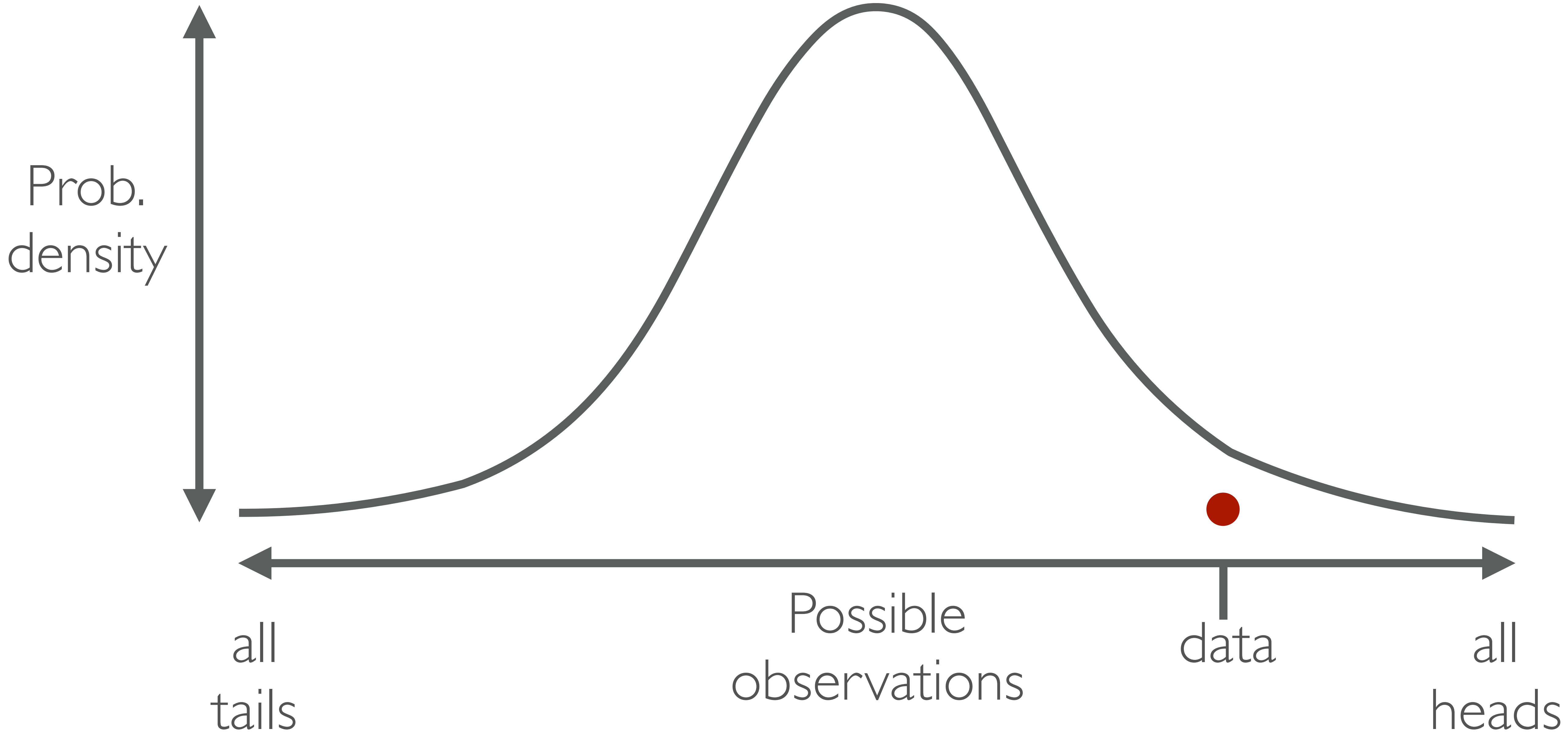
Calculate p-value

Calculate p-value

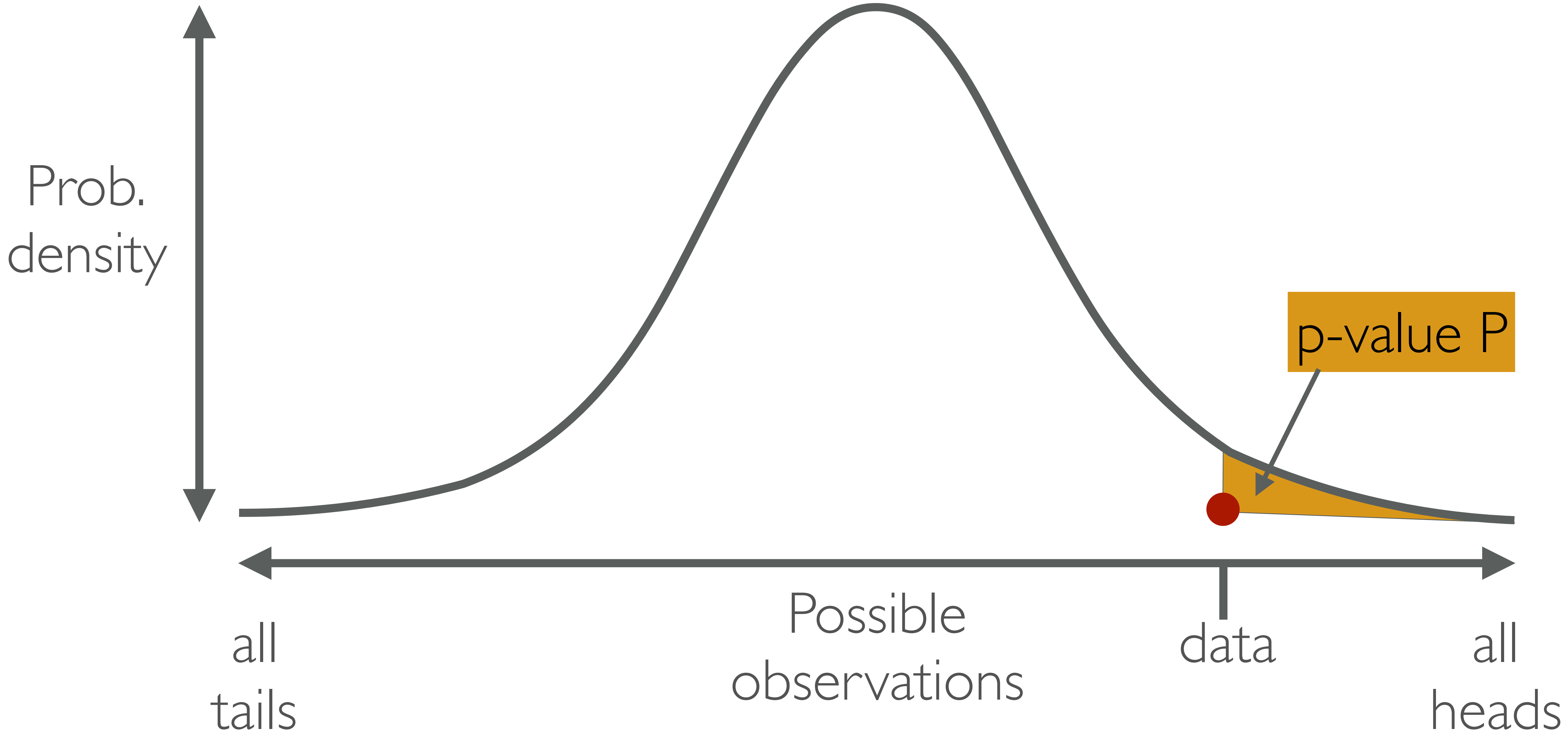
Calculate p-value



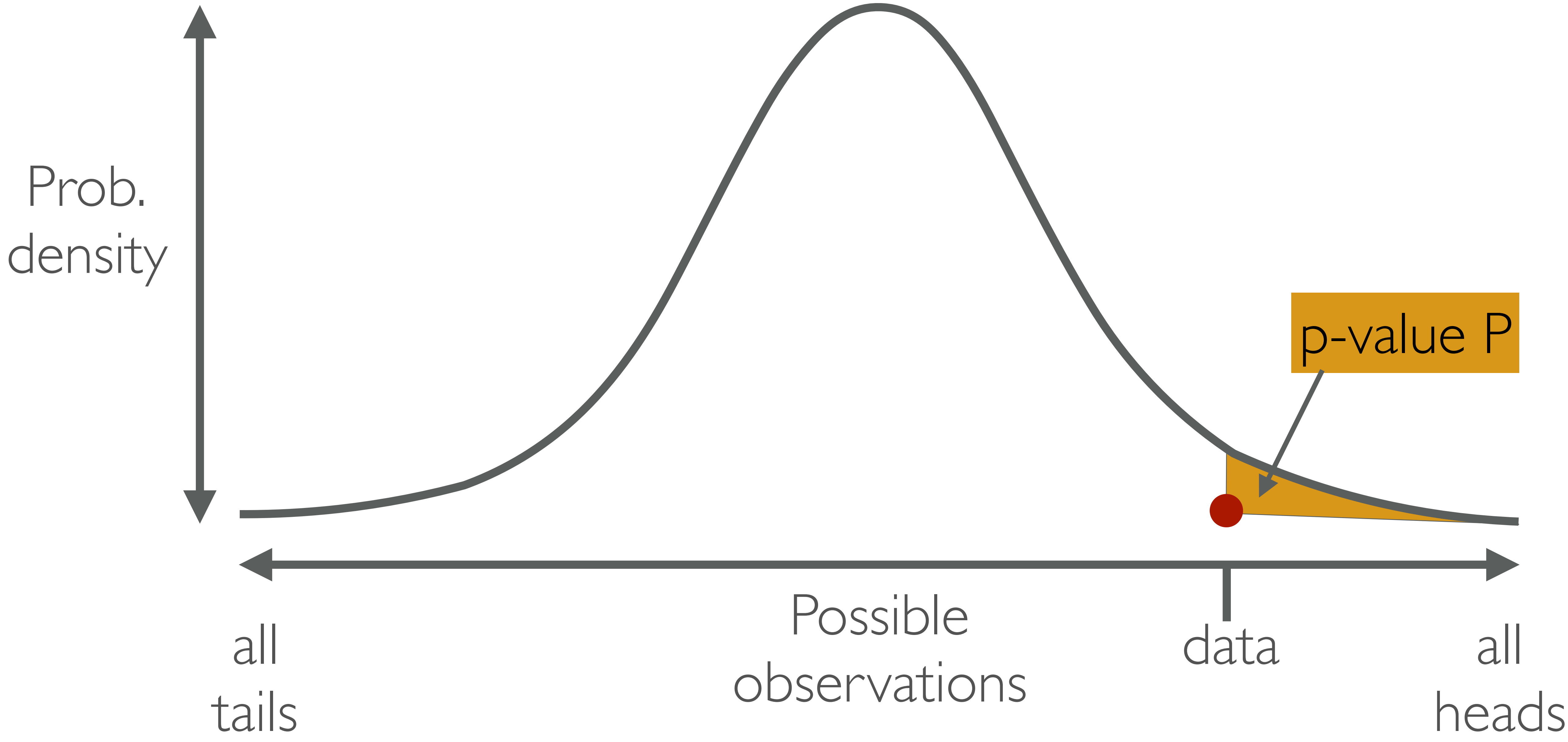
Calculate p-value



Calculate p-value

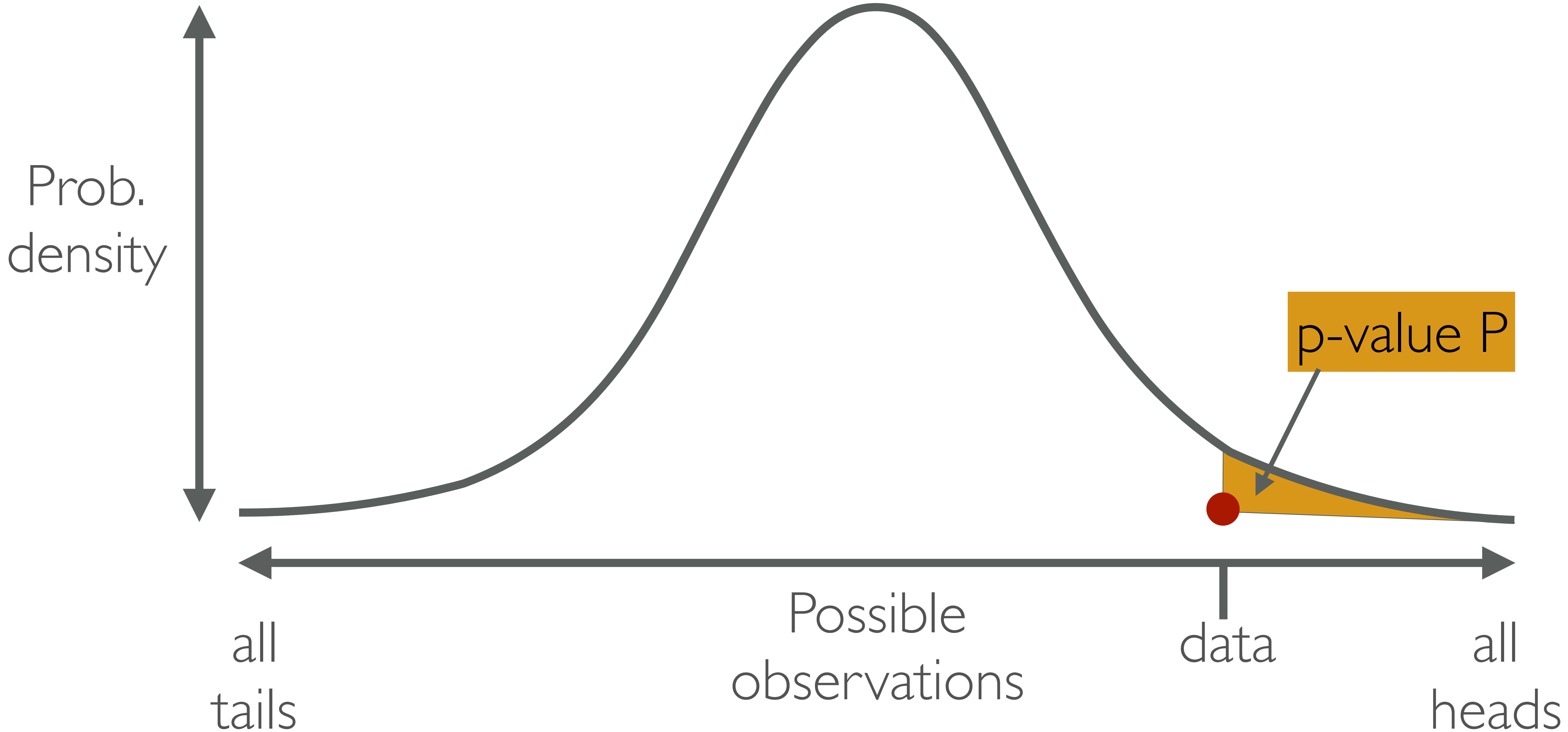


Calculate p-value



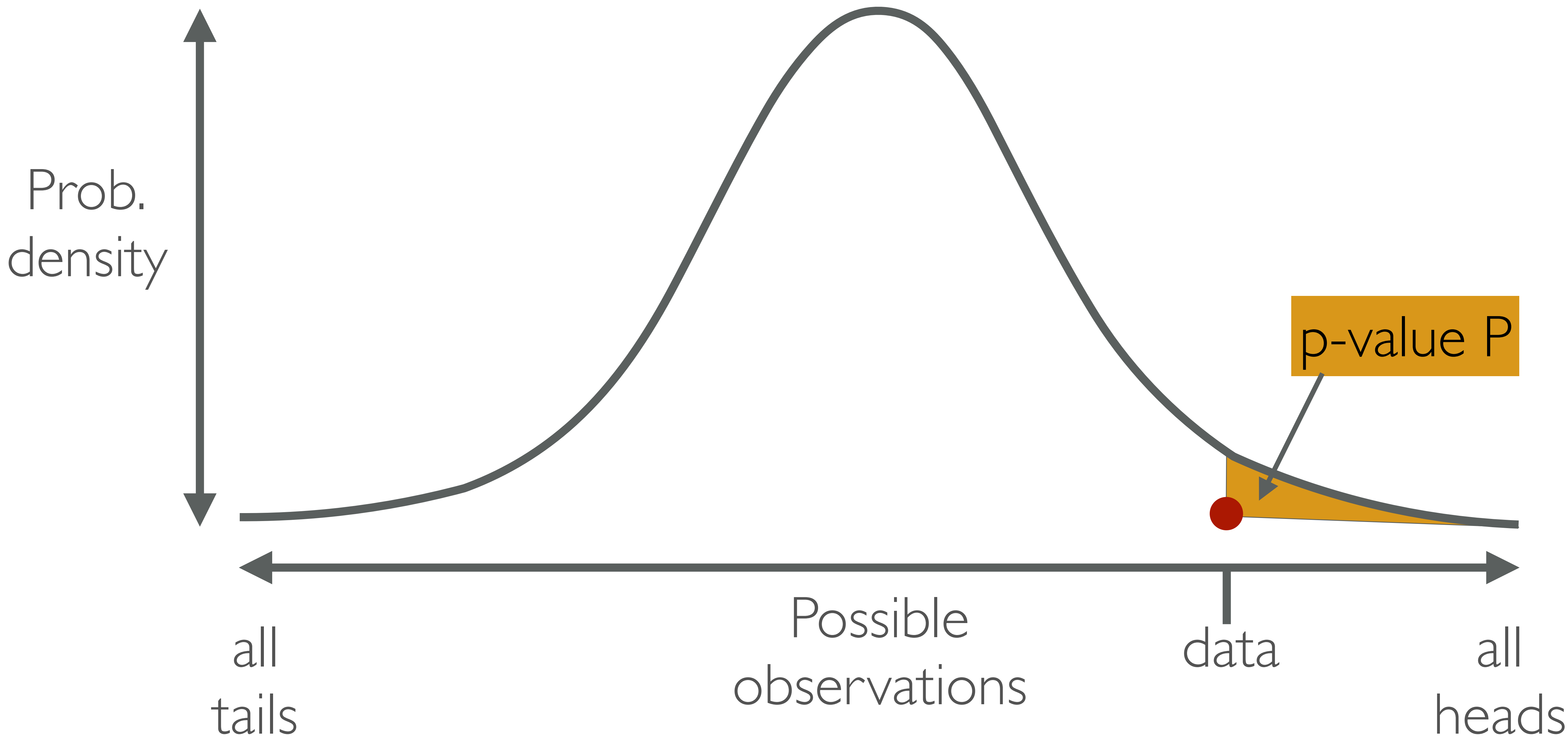
Reject null if $P \leq \alpha$

Calculate p-value



Reject null if $P \leq \alpha \quad \approx \#H - \#T \geq \sqrt{2N \log(1/\alpha)}$.

Calculate p-value



Reject null if $P \leq \alpha \quad \approx \#H - \#T \geq \sqrt{2N \log(1/\alpha)}$.

Then, $\mathbf{Pr}(\text{false positive}) \leq \alpha$.

An equivalent view via confidence intervals

An equivalent view via confidence intervals

Estimate the coin bias by $\widehat{\mu} := \dfrac{\#H - \#T}{N}$.

An equivalent view via confidence intervals

Estimate the coin bias by $\widehat{\mu} := \frac{\#H - \#T}{N}$.

An asymptotic $(1 - \alpha)$-CI for μ is given by $\left(\widehat{\mu} - \frac{z_{1-\alpha}}{\sqrt{N}}, \widehat{\mu} + \frac{z_{1-\alpha}}{\sqrt{N}}\right)$,

An equivalent view via confidence intervals

Estimate the coin bias by $\hat{\mu} := \frac{\#H - \#T}{N}$.

An asymptotic $(1-\alpha)$-CI for μ is given by $\left(\hat{\mu} - \frac{z_{1-\alpha}}{\sqrt{N}}, \hat{\mu} + \frac{z_{1-\alpha}}{\sqrt{N}}\right)$,

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of $N(0,1)$.

(appealing to the Central Limit Theorem)

An equivalent view via confidence intervals

Estimate the coin bias by $\hat{\mu} := \dfrac{\#H - \#T}{N}$.

An asymptotic $(1-\alpha)$-CI for μ is given by $\left(\hat{\mu} - \dfrac{z_{1-\alpha}}{\sqrt{N}}, \hat{\mu} + \dfrac{z_{1-\alpha}}{\sqrt{N}}\right)$,

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of $N(0,1)$.
(appealing to the Central Limit Theorem)

If this confidence interval does not contain 0,
we may be reasonably confident that the coin is biased,
and we may reject the null hypothesis.

An equivalent view via confidence intervals

Estimate the coin bias by $\widehat{\mu} := \dfrac{\#H - \#T}{N}$.

An asymptotic $(1-\alpha)$-CI for μ is given by $\left(\widehat{\mu} - \dfrac{z_{1-\alpha}}{\sqrt{N}}, \widehat{\mu} + \dfrac{z_{1-\alpha}}{\sqrt{N}}\right)$,

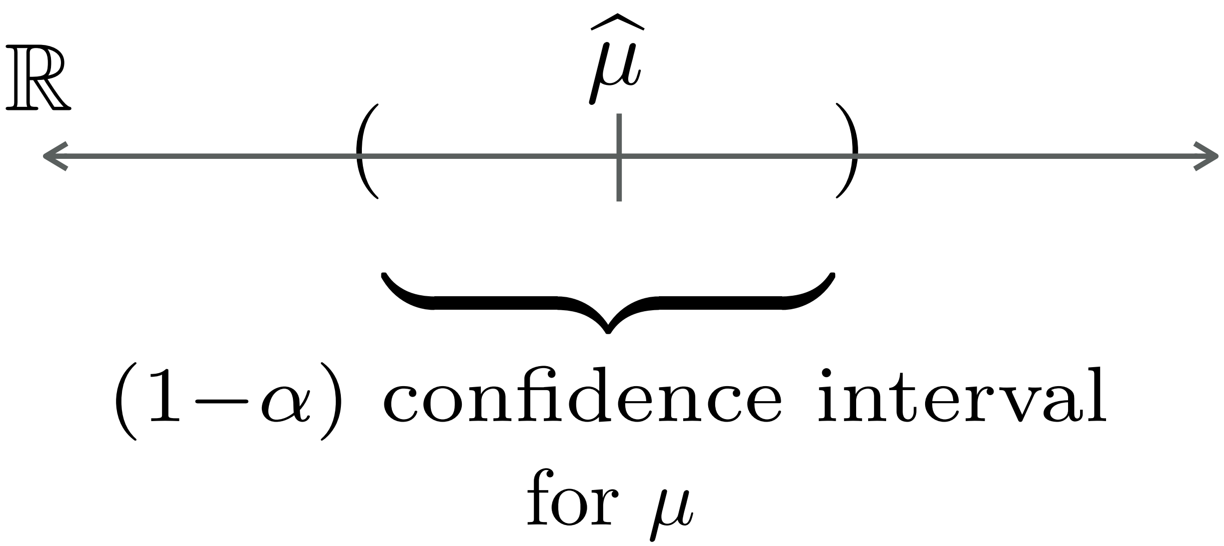
where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of $N(0,1)$.

(appealing to the Central Limit Theorem)

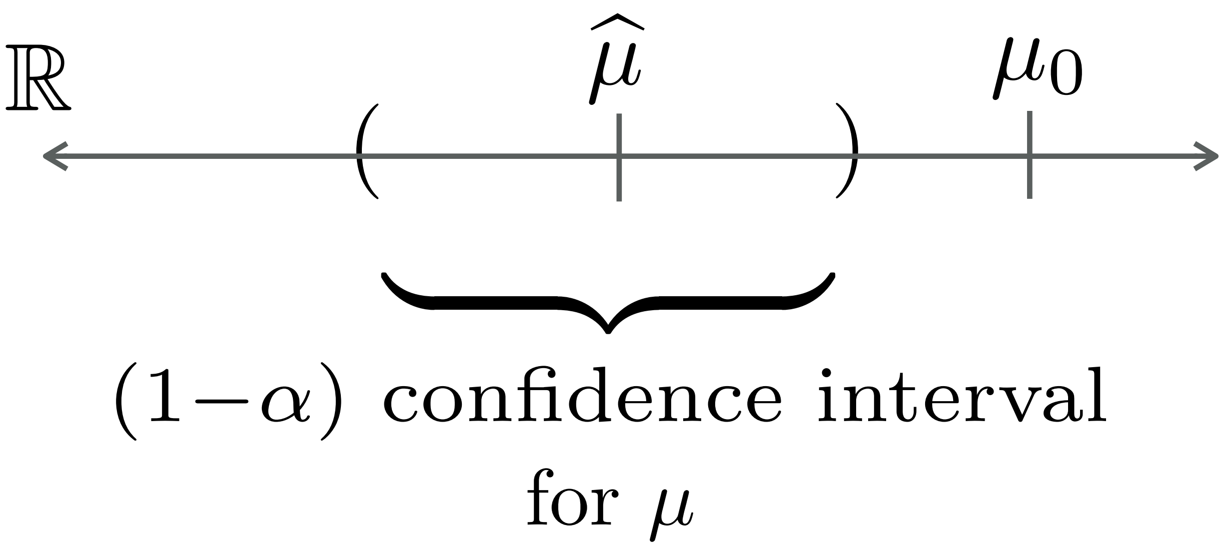
If this confidence interval does not contain 0,
we may be reasonably confident that the coin is biased,
and we may reject the null hypothesis.

$$\approx \#H - \#T \geq \sqrt{2N \log(1/\alpha)}\,.$$

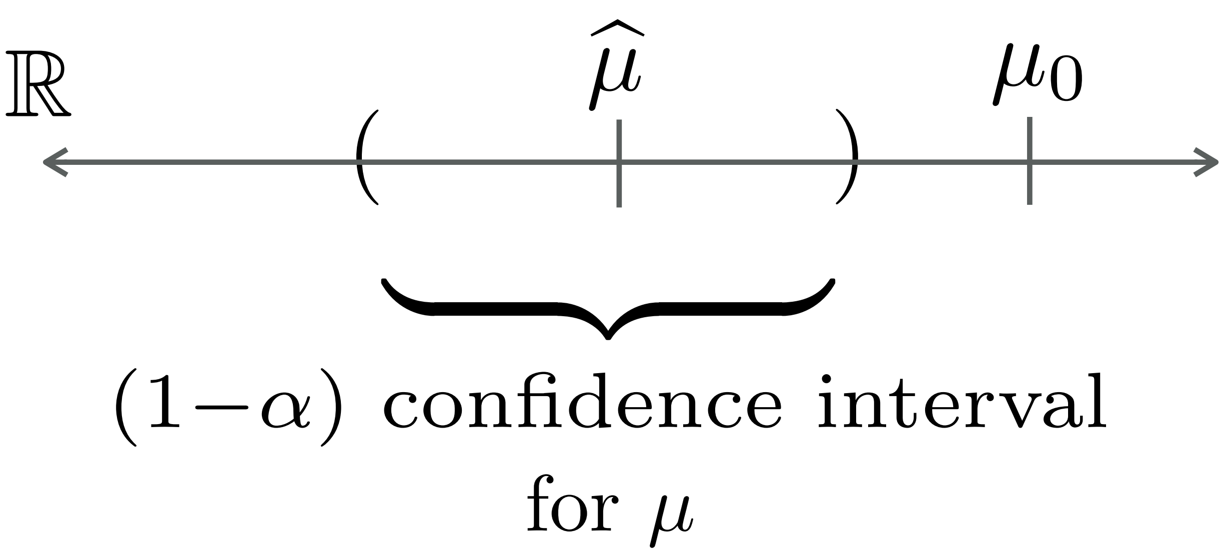
For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:



For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:

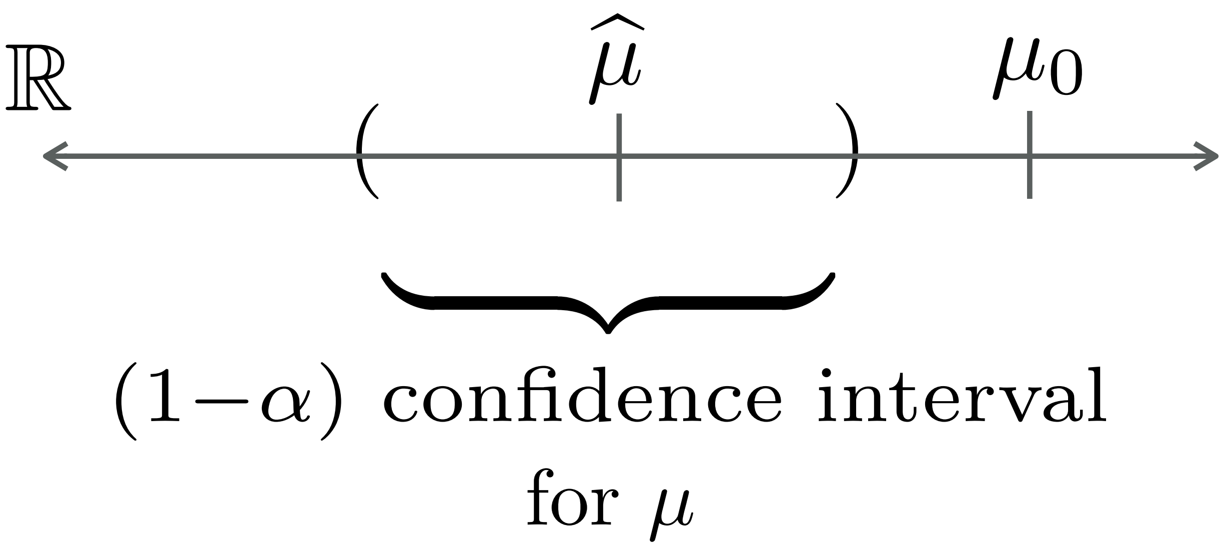


For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:



$\equiv$

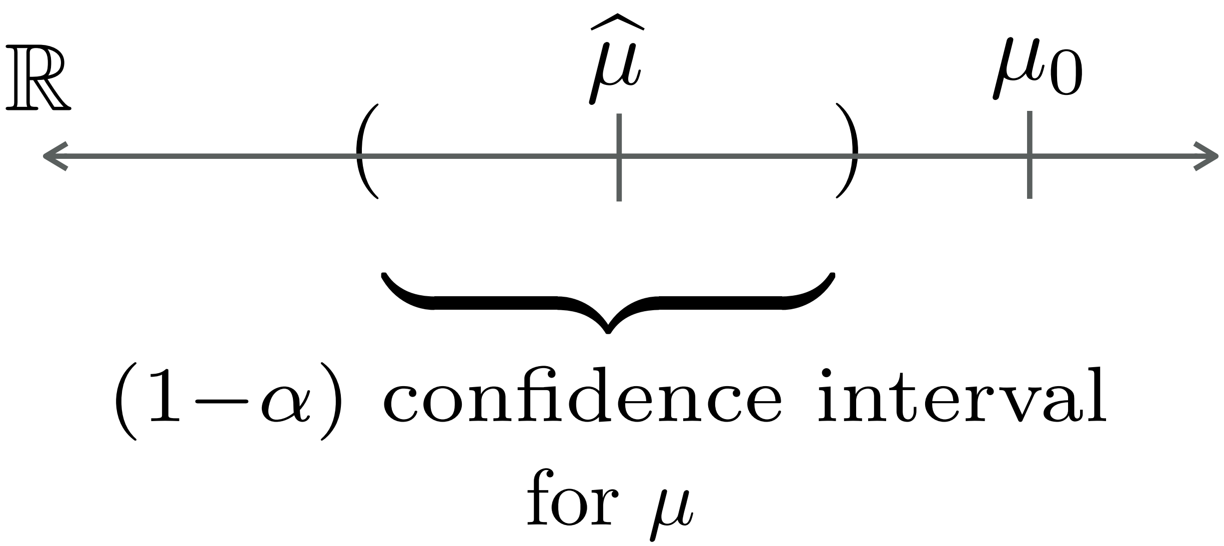
For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:



$\equiv$

For $H_0 : \mu = \mu_0$

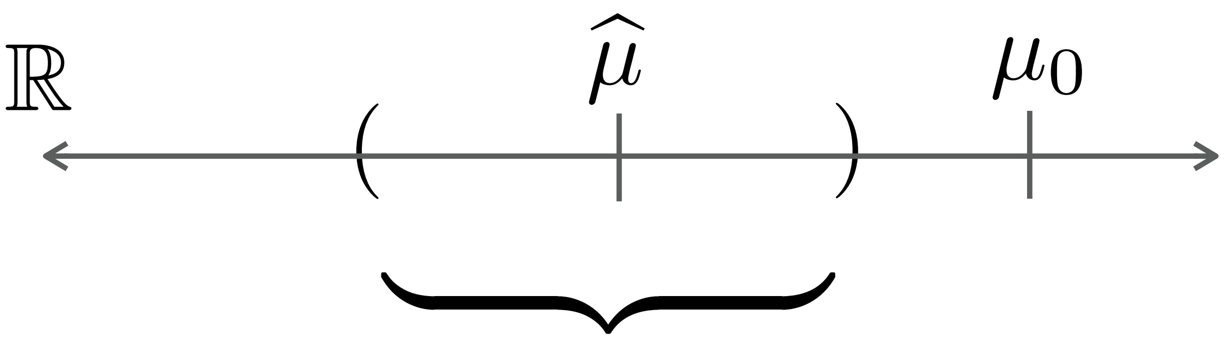
For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:



$\equiv$

For $H_0 : \mu = \mu_0$
we have p-value $P_{\mu_0} \leq \alpha$

For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:



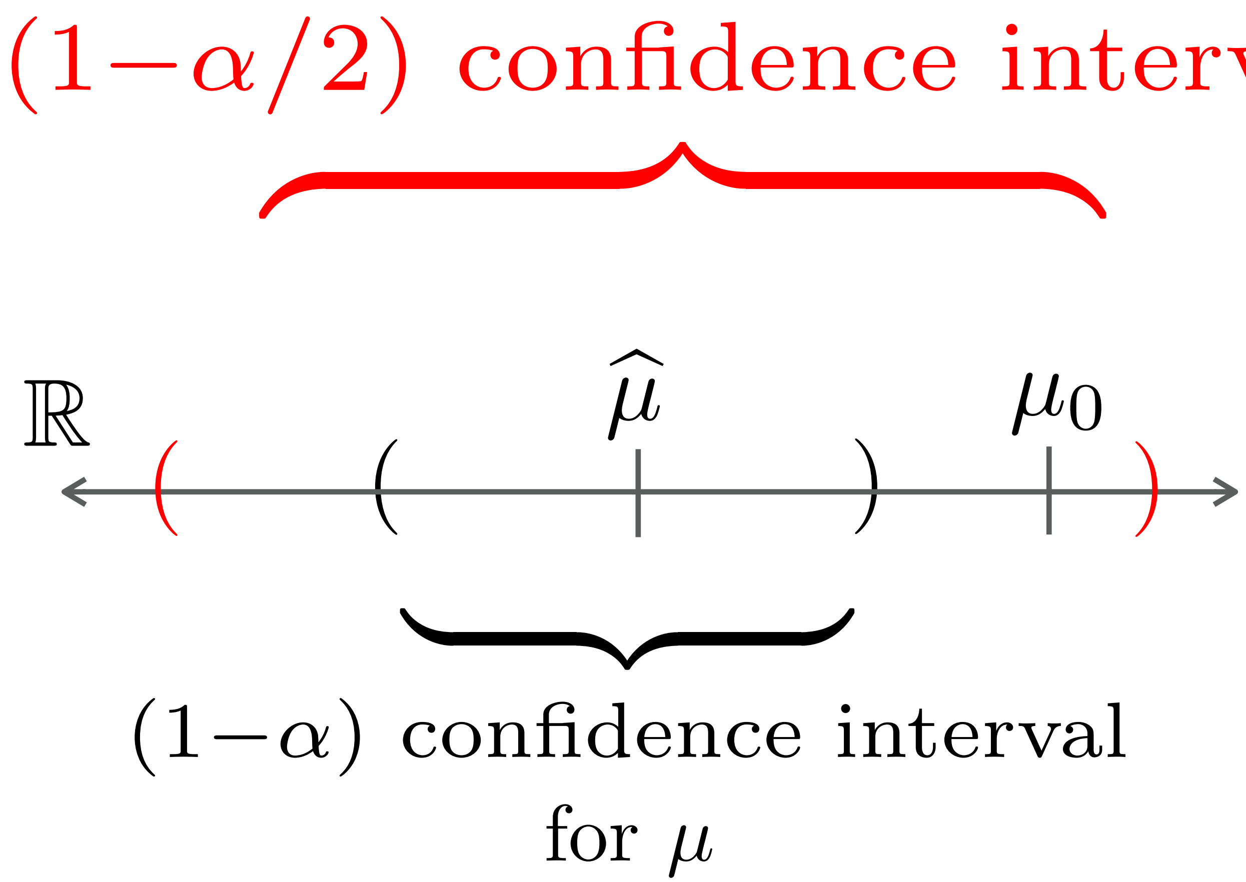
$(1-\alpha)$ confidence interval
for μ

$\equiv$

For $H_0 : \mu = \mu_0$
we have p-value $P_{\mu_0} \leq \alpha$

(we would reject the null
hypothesis at level α)

For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:

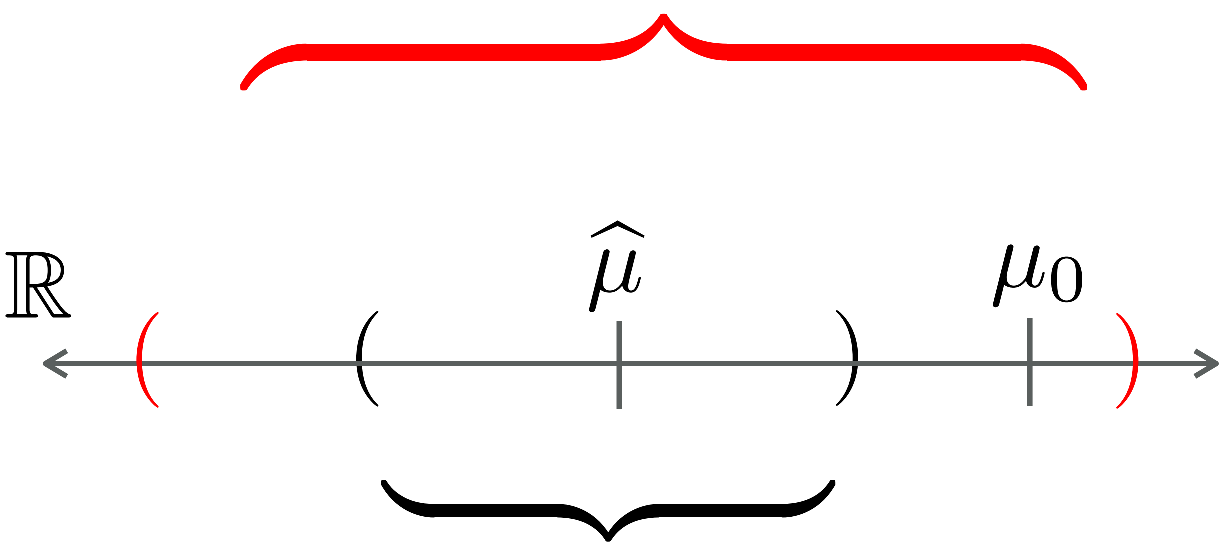


$\equiv$

For $H_0 : \mu = \mu_0$
we have p-value $P_{\mu_0} \leq \alpha$
(we would reject the null
hypothesis at level α)

For any parameter μ of interest,
with associated estimator $\widehat{\mu}$,
the following claim holds:

$(1-\alpha/2)$ confidence interval $\equiv$ $P_{\mu_0} > \alpha/2$



$(1-\alpha)$ confidence interval
for μ

$\equiv$

For $H_0 : \mu = \mu_0$
we have p-value $P_{\mu_0} \leq \alpha$
(we would reject the null
hypothesis at level α)

In summary, tests (p-values) and CIs are “dual”.

In summary, tests (p-values) and CIs are "dual".

family of tests for $\theta \rightarrow$ CI for θ

CI for $\theta \rightarrow$ family of tests for θ

In summary, tests (p-values) and CIs are "dual".

family of tests for θ $\rightarrow$ CI for θ

A $(1-\alpha)$-CI for a parameter θ is the set of all θ_0 such that the test for $H_0 : \theta = \theta_0$ has p-value larger than α.

CI for θ $\rightarrow$ family of tests for θ

In summary, tests (p-values) and CIs are “dual”.

family of tests for $\theta \rightarrow$ CI for θ

A $(1-\alpha)$-CI for a parameter θ is the set of all θ_0 such that the test for $H_0 : \theta = \theta_0$ has p-value larger than α.

CI for $\theta \rightarrow$ family of tests for θ

A p-value for testing the null $H_0 : \theta = \theta_0$ can be given by the smallest q for which the $(1-q)$-CI for θ fails to cover θ_0.

In summary, tests (p-values) and CIs are "dual".

family of tests for θ → CI for θ

A $(1-\alpha)$-CI for a parameter θ is the set of all θ_0 such that the test for $H_0: \theta = \theta_0$ has p-value larger than α.

CI for θ → family of tests for θ

A p-value for testing the null $H_0: \theta = \theta_0$ can be given by the smallest q for which the $(1-q)$-CI for θ fails to cover θ_0.

CI for θ → composite tests for θ

In summary, tests (p-values) and CIs are "dual".

family of tests for $\theta \rightarrow$ CI for θ

A $(1-\alpha)$-CI for a parameter θ is the set of all θ_0 such that the test for $H_0 : \theta = \theta_0$ has p-value larger than α.

CI for $\theta \rightarrow$ family of tests for θ

A p-value for testing the null $H_0 : \theta = \theta_0$ can be given by the smallest q for which the $(1-q)$-CI for θ fails to cover θ_0.

CI for $\theta \rightarrow$ composite tests for θ

A p-value for testing the null $H_0 : \theta \in \Theta_0$ can be given by the smallest q for which the $(1-q)$-CI for θ fails to intersect Θ_0.

In summary, tests (p-values) and CIs are "dual".

family of tests for $\theta \to$ CI for θ

A $(1-\alpha)$-CI for a parameter θ is the set of all θ_0 such that the test for $H_0 : \theta = \theta_0$ has p-value larger than α.

CI for $\theta \to$ family of tests for θ

A p-value for testing the null $H_0 : \theta = \theta_0$ can be given by the smallest q for which the $(1-q)$-CI for θ fails to cover θ_0.

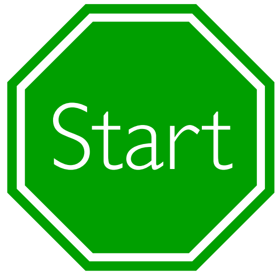
CI for $\theta \to$ composite tests for θ

A p-value for testing the null $H_0 : \theta \in \Theta_0$ can be given by the smallest q for which the $(1-q)$-CI for θ fails to intersect Θ_0.

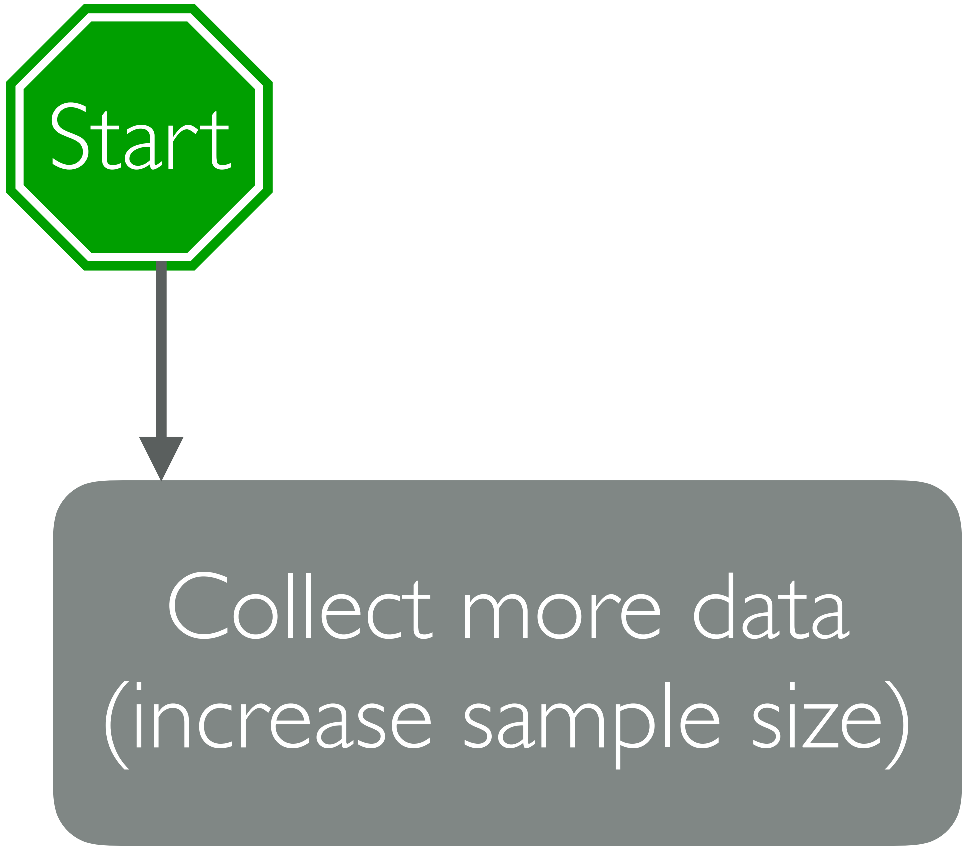
Both of them are useful tools to estimate uncertainty,
and like any other tool, they can be used well, or be misused.

However, commonly taught confidence intervals
and p-values are only valid (correctly control error)
if the sample size is fixed in advance.

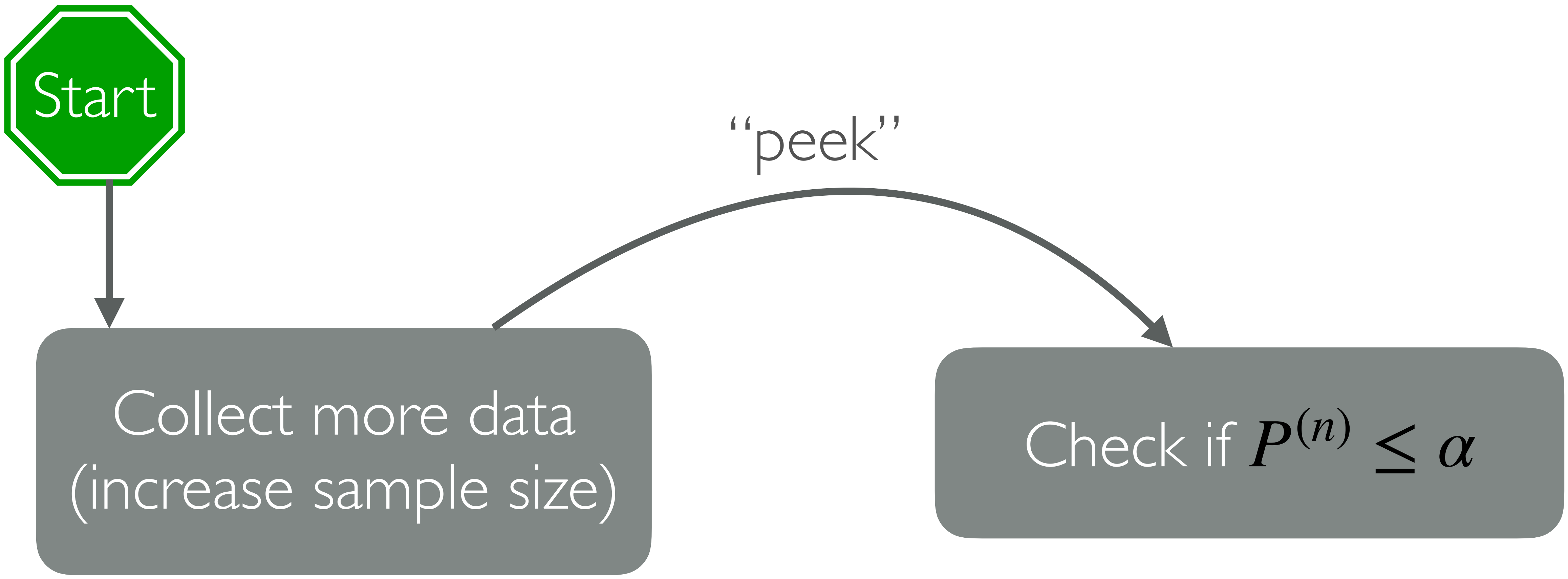
High-level caricature of an A/B-test



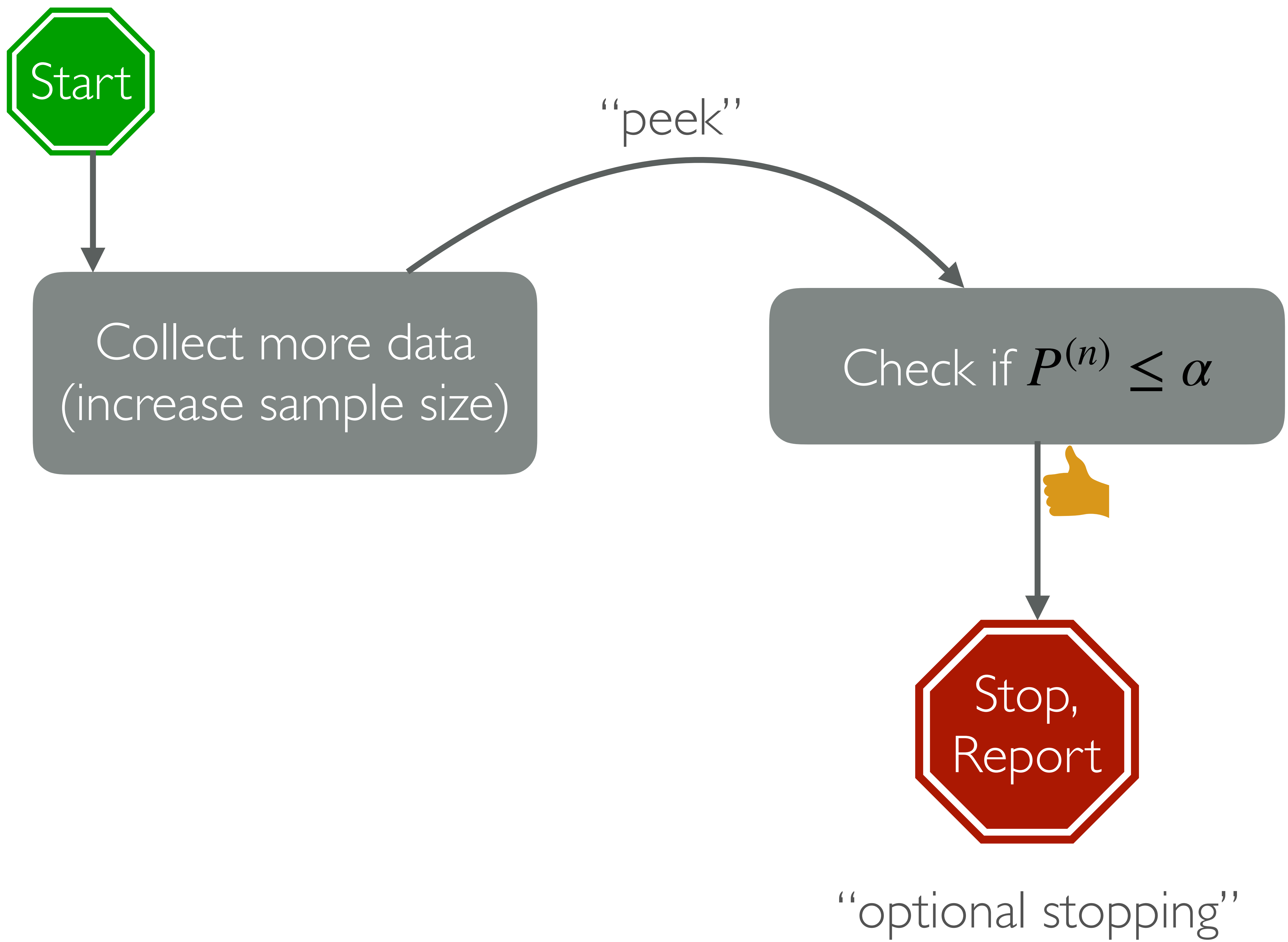
High-level caricature of an A/B-test



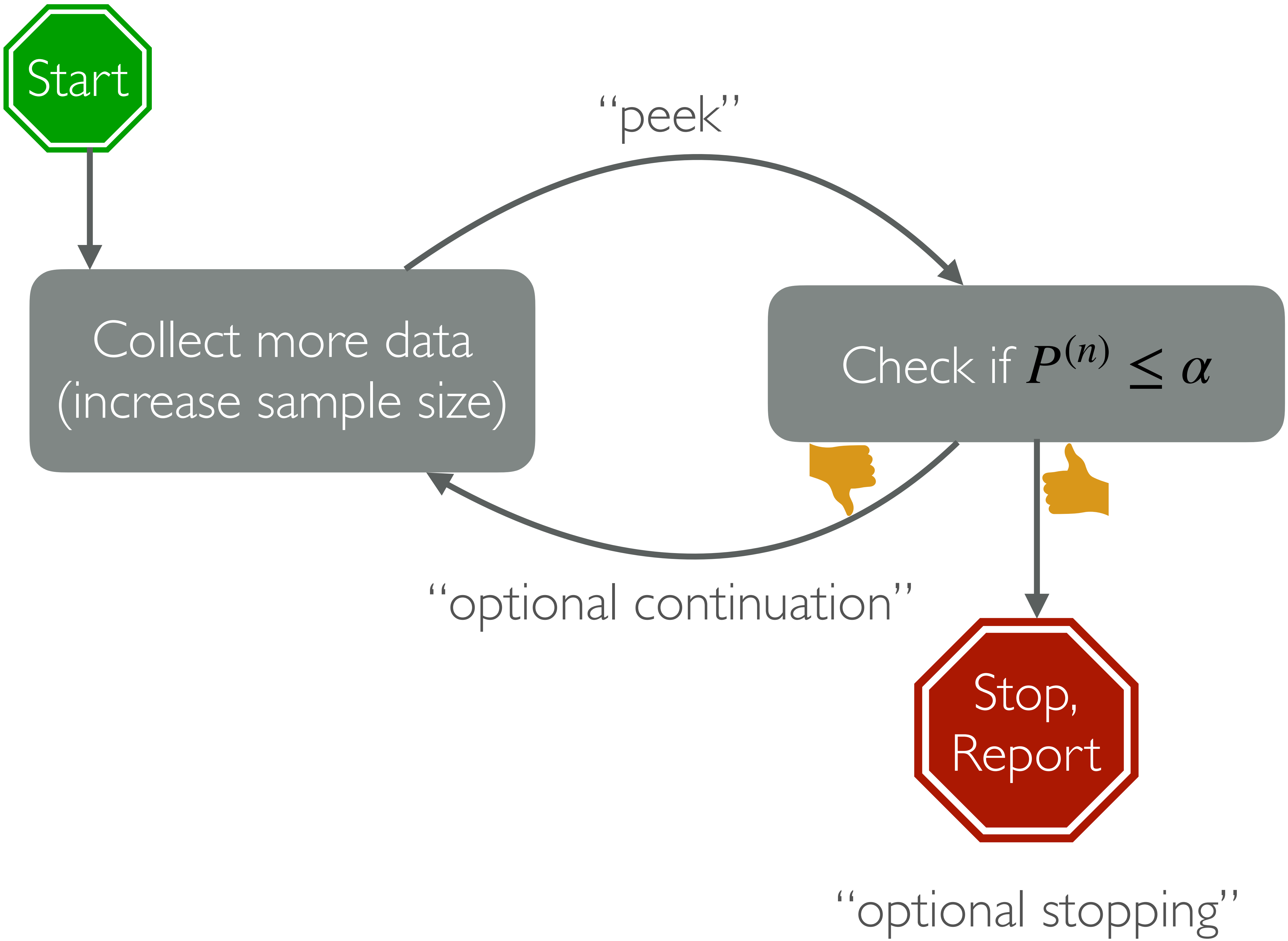
High-level caricature of an A/B-test



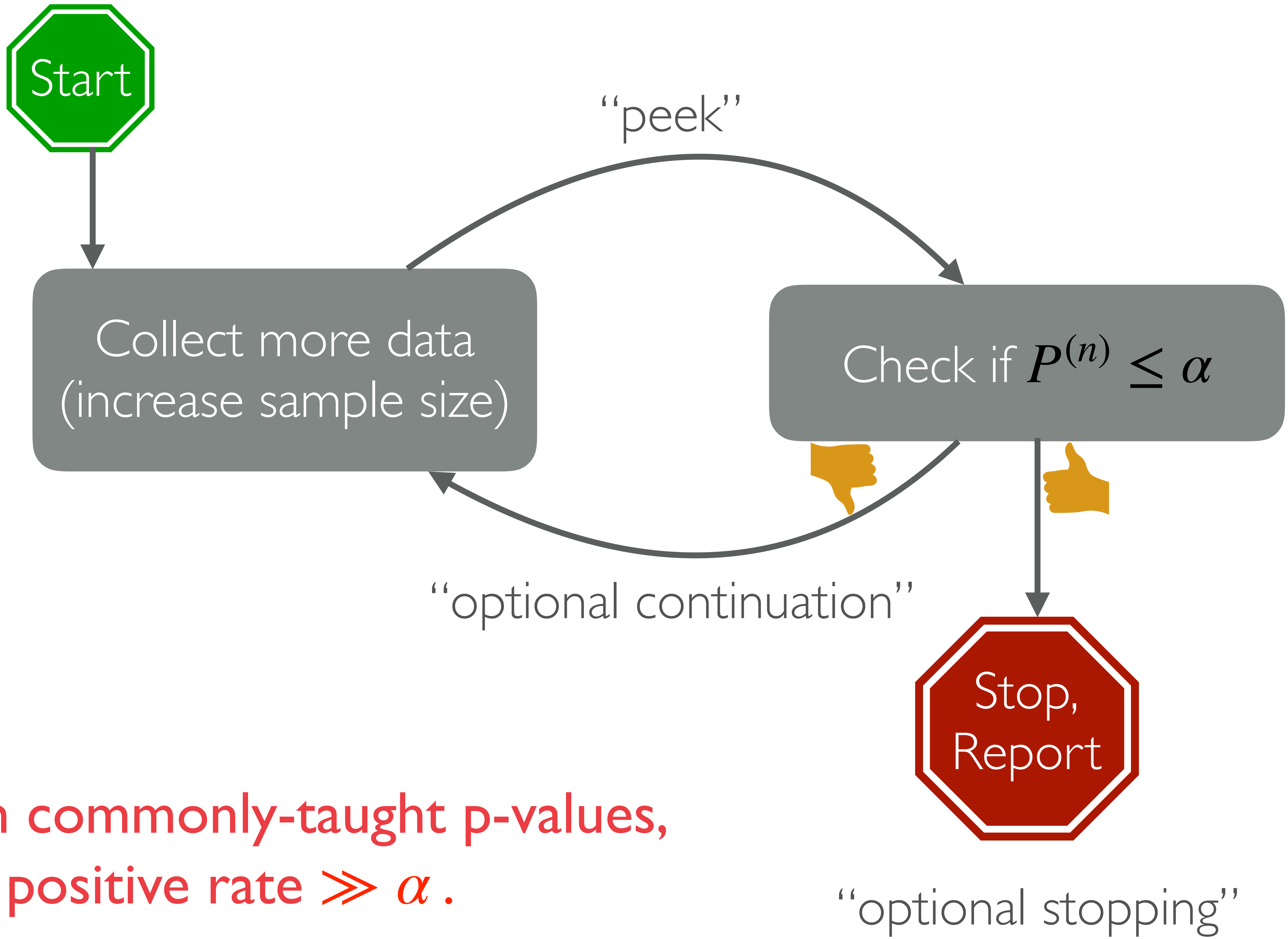
High-level caricature of an A/B-test



High-level caricature of an A/B-test



High-level caricature of an A/B-test

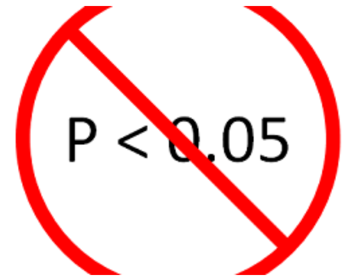


With commonly-taught p-values,
false positive rate $\gg \alpha$.

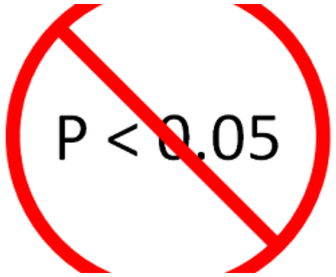

NOW
NOW
NOW
NOW
NOW
NOW
NOW
NOW


NOW
NOW
NOW
NOW
NOW
NOW
NOW
NOW

After 10 people

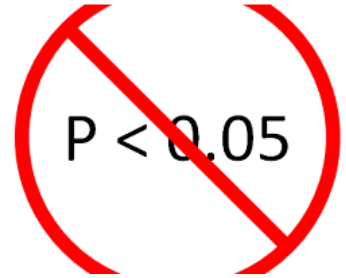


After 10 people

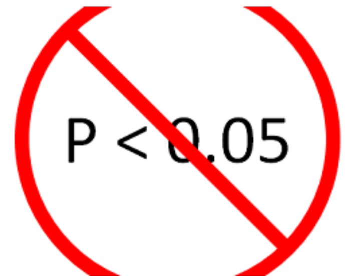


After 10 people

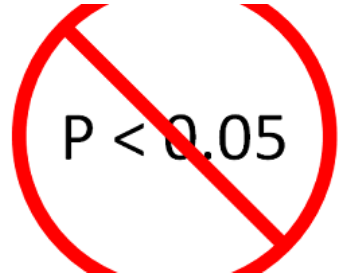
After 284 people



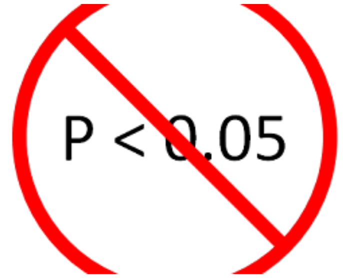
After 10 people



After 284 people

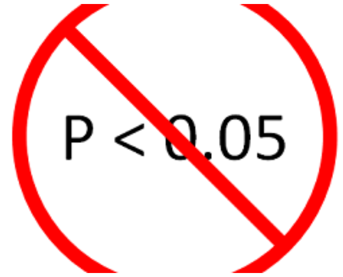


After 10 people

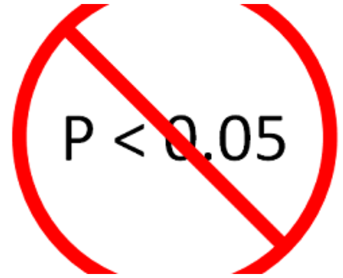


After 284 people

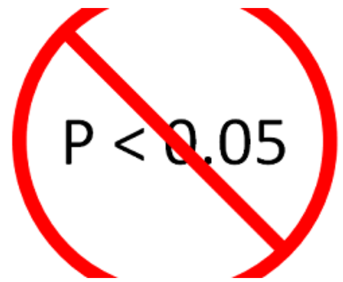
After 1214 people



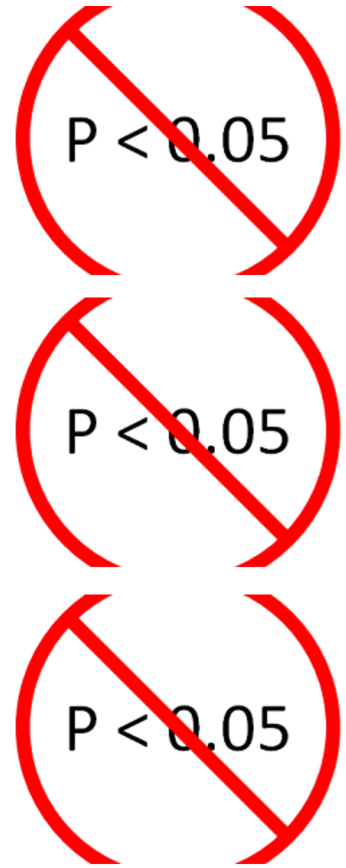
After 10 people



After 284 people



After 1214 people

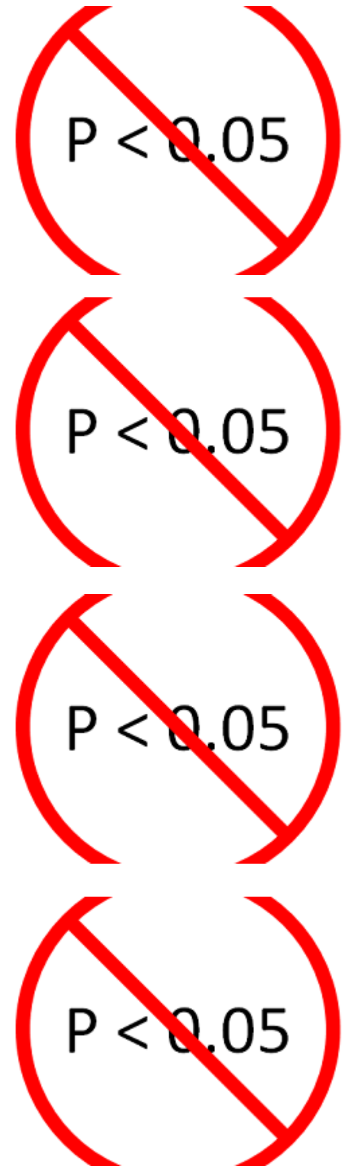


After 10 people

After 284 people

After 1214 people

After 2398 people

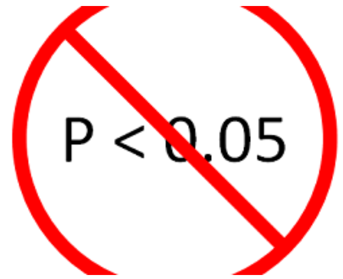


After 10 people

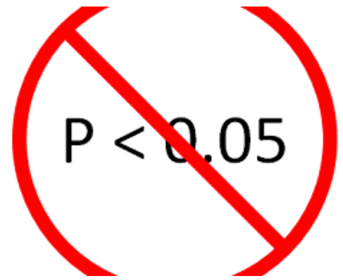
After 284 people

After 1214 people

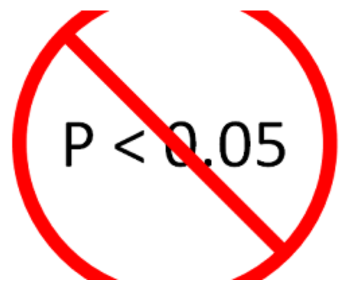
After 2398 people



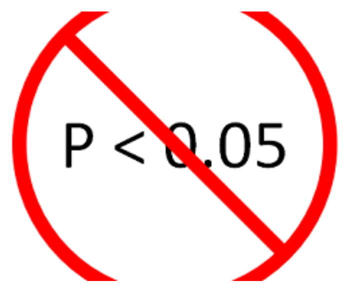
After 10 people



After 284 people



After 1214 people



After 2398 people

After 7224 people

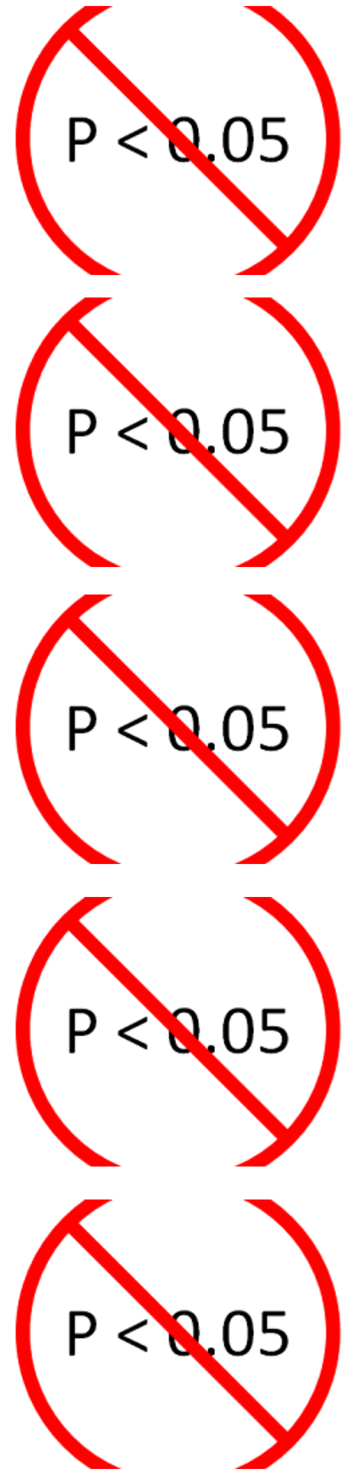
~~$P < 0.05$~~ After 10 people

~~$P < 0.05$~~ After 284 people

~~$P < 0.05$~~ After 1214 people

~~$P < 0.05$~~ After 2398 people

~~$P < 0.05$~~ After 7224 people



After 10 people

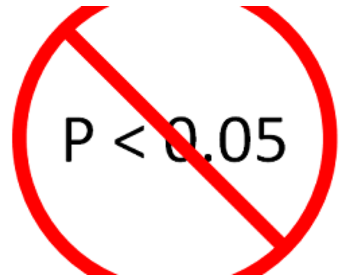
After 284 people

After 1214 people

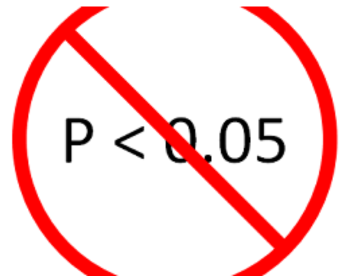
After 2398 people

After 7224 people

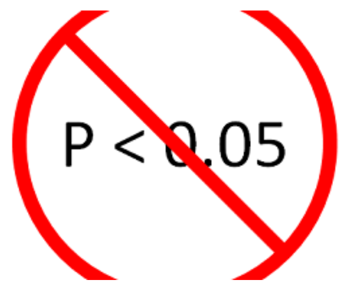
After 11,219 people, STOP!



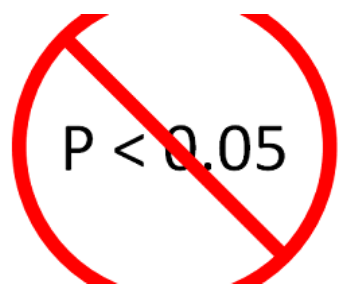
After 10 people



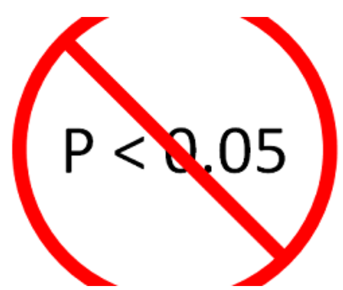
After 284 people



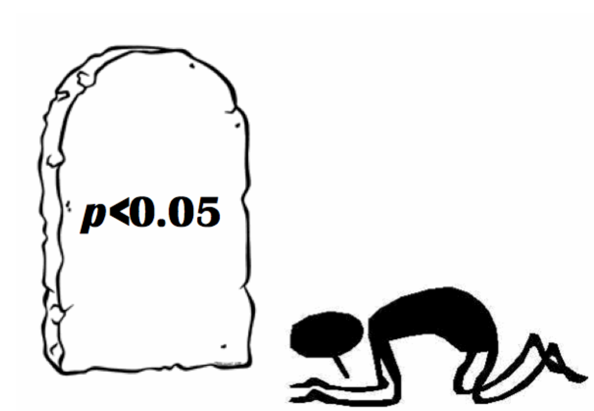
After 1214 people



After 2398 people



After 7224 people



After 11,219 people, STOP!

Let $P^{(n)}$ be a classical p-value (eg: t-test), calculated using the first n samples.

Let $P^{(n)}$ be a classical p-value (eg: t-test), calculated using the first n samples.

Under the null hypothesis (no treatment effect),

$$\forall n \geq 1, \qquad \underbrace{\Pr(P^{(n)} \leq \alpha)}_{\text{prob. of false positive}} \qquad \leq \alpha \,.$$

Let $P^{(n)}$ be a classical p-value (eg: t-test), calculated using the first n samples.

Under the null hypothesis (no treatment effect),

$$\forall n \geq 1, \qquad \underbrace{\Pr(P^{(n)} \leq \alpha)}_{\text{prob. of false positive}} \qquad \leq \alpha\,.$$

Let τ be the stopping time of the experiment.

Often, τ depends on data, eg: $\tau := \min\{n \in \mathbb{N} : P_n \leq \alpha\}$.

Let $P^{(n)}$ be a classical p-value (eg: t-test), calculated using the first n samples.

Under the null hypothesis (no treatment effect),

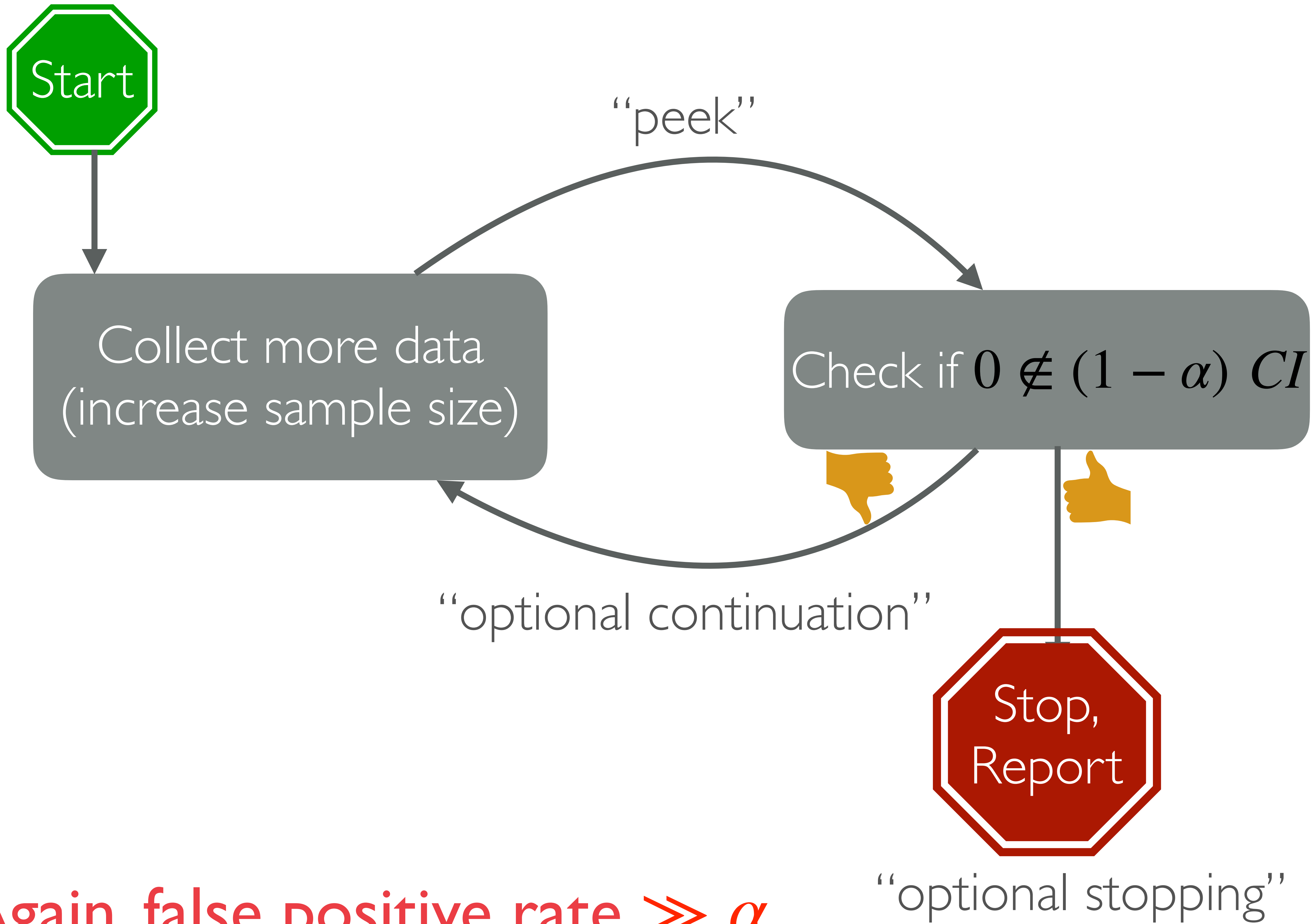
$$\forall n \geq 1, \qquad \underbrace{\Pr(P^{(n)} \leq \alpha)}_{\text{prob. of false positive}} \qquad \leq \alpha\,.$$

Let τ be the stopping time of the experiment.

Often, τ depends on data, eg: $\tau := \min\{n \in \mathbb{N} : P_n \leq \alpha\}$.

Unfortunately, $\Pr(P^{(\tau)} \leq \alpha) \nleq \alpha$.
In other words, $\Pr(\exists n \in \mathbb{N} : P^{(n)} \leq \alpha) \gg \alpha$.

Same problem with confidence interval (CI)



Again, false positive rate $\gg \alpha$.

Let $(L^{(n)}, U^{(n)})$ be any classical $(1 - \alpha)$ CI, calculated using the first n samples (eg: CLT).

Let $(L^{(n)}, U^{(n)})$ be any classical $(1-\alpha)$ CI, calculated using the first n samples (eg: CLT).

When trying to estimate the treatment effect θ,

$$\forall n \geq 1,\ \underbrace{\Pr(\theta \in (L^{(n)}, U^{(n)}))}_{\text{prob. of coverage}} \geq 1 - \alpha\,.$$

Let $(L^{(n)}, U^{(n)})$ be any classical $(1-\alpha)$ CI, calculated using the first n samples (eg: CLT).

When trying to estimate the treatment effect θ,

$$\forall n \geq 1,\ \underbrace{\Pr(\theta \in (L^{(n)}, U^{(n)}))}_{\text{prob. of coverage}} \geq 1 - \alpha\,.$$

Let τ be the stopping time of the experiment.

Again, τ may depend on data, eg: $\tau := \min\{n \in \mathbb{N} : L^{(n)} > 0\}$.

Unfortunately, $\Pr(\theta \in (L^{(\tau)}, U^{(\tau)})) \ngeq 1 - \alpha\,.$

Let $(L^{(n)}, U^{(n)})$ be any classical $(1-\alpha)$ CI,
calculated using the first n samples (eg: CLT).

When trying to estimate the treatment effect θ,

$$\forall n \geq 1,\ \underbrace{\Pr(\theta \in (L^{(n)}, U^{(n)}))}_{\text{prob. of coverage}} \geq 1-\alpha\,.$$

Let τ be the stopping time of the experiment.

Again, τ may depend on data, eg: $\tau := \min\{n \in \mathbb{N} : L^{(n)} > 0\}$.

Unfortunately, $\Pr(\theta \in (L^{(\tau)}, U^{(\tau)})) \ngeq 1-\alpha\,.$

In other words, $\Pr(\forall n \geq 1 : \theta \in (L^{(n)}, U^{(n)})) \ll 1-\alpha\,.$

usually $= 0$.

Solution: "confidence sequence"
(aka "anytime confidence intervals")

or "sequential p-values" for testing
(aka "always-valid p-values")

A "confidence sequence" for a parameter θ is a sequence of confidence intervals (L_n, U_n) with a uniform (simultaneous) coverage guarantee.

$$\mathbb{P}(\forall n \geq 1 : \theta \in (L_n, U_n)) \geq 1 - \alpha\,.$$

Sample size

A “confidence sequence” for a parameter θ is a sequence of confidence intervals (L_n, U_n) with a uniform (simultaneous) coverage guarantee.

$$\mathbb{P}(\forall n \geq 1 : \theta \in (L_n, U_n)) \geq 1 - \alpha\,.$$

Sample size

Darling, Robbins ’67, ‘68
Lai ’76, ’84
Howard, Ramdas, McAuliffe, Sekhon ’18

Example: tracking the mean of a Gaussian or Bernoulli from i.i.d. observations.

$$X_1, X_2, \ldots \sim N(\theta, 1) \text{ or } Ber(\theta)$$

Example: tracking the mean of a Gaussian or Bernoulli from i.i.d. observations.

$$X_1, X_2, \ldots \sim N(\theta,1) \text{ or } Ber(\theta)$$

Producing a confidence *interval* at a fixed time is elementary statistics (~100 years old).

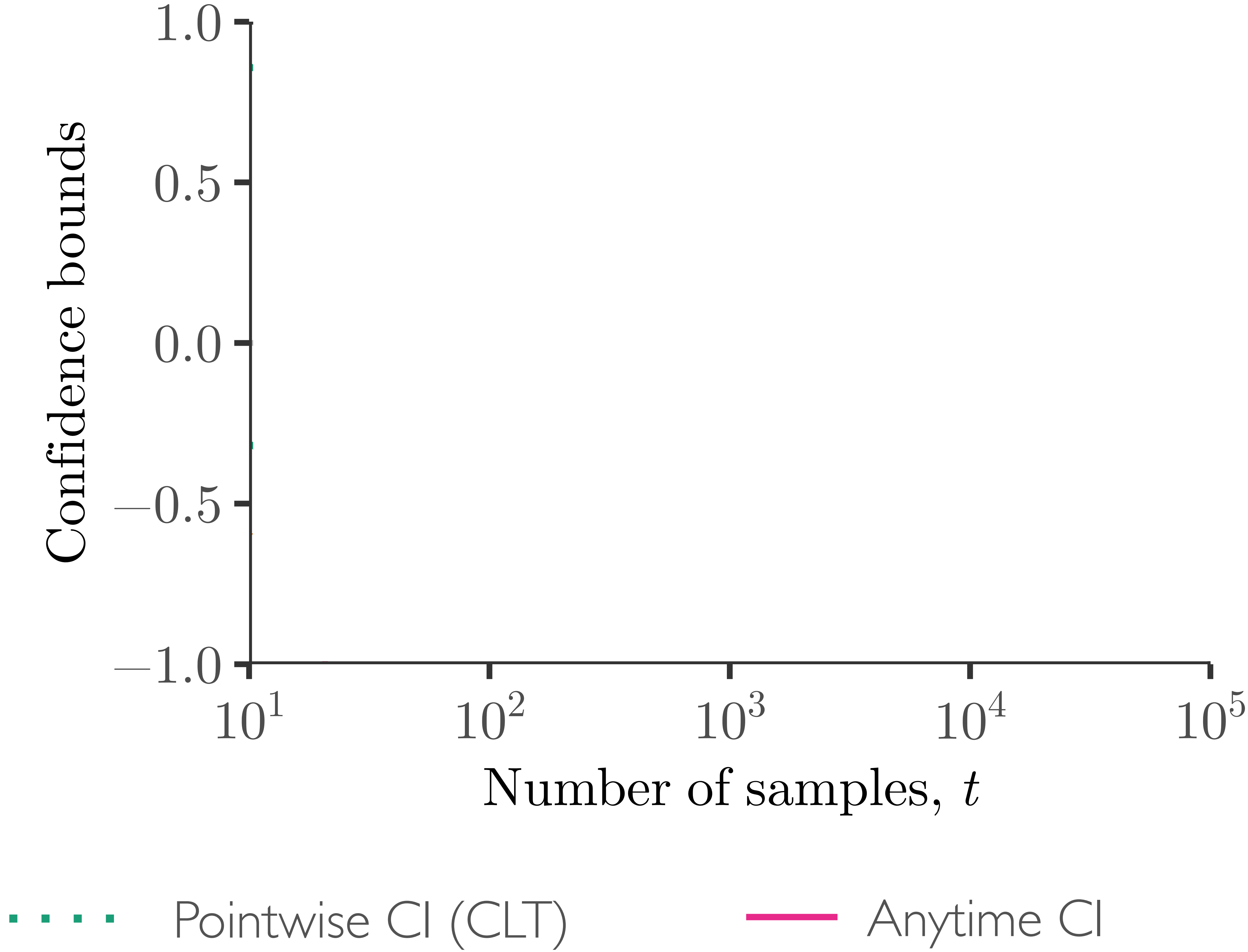
Example: tracking the mean of a Gaussian or Bernoulli from i.i.d. observations.

$$X_1, X_2, \ldots \sim N(\theta, 1) \text{ or } Ber(\theta)$$

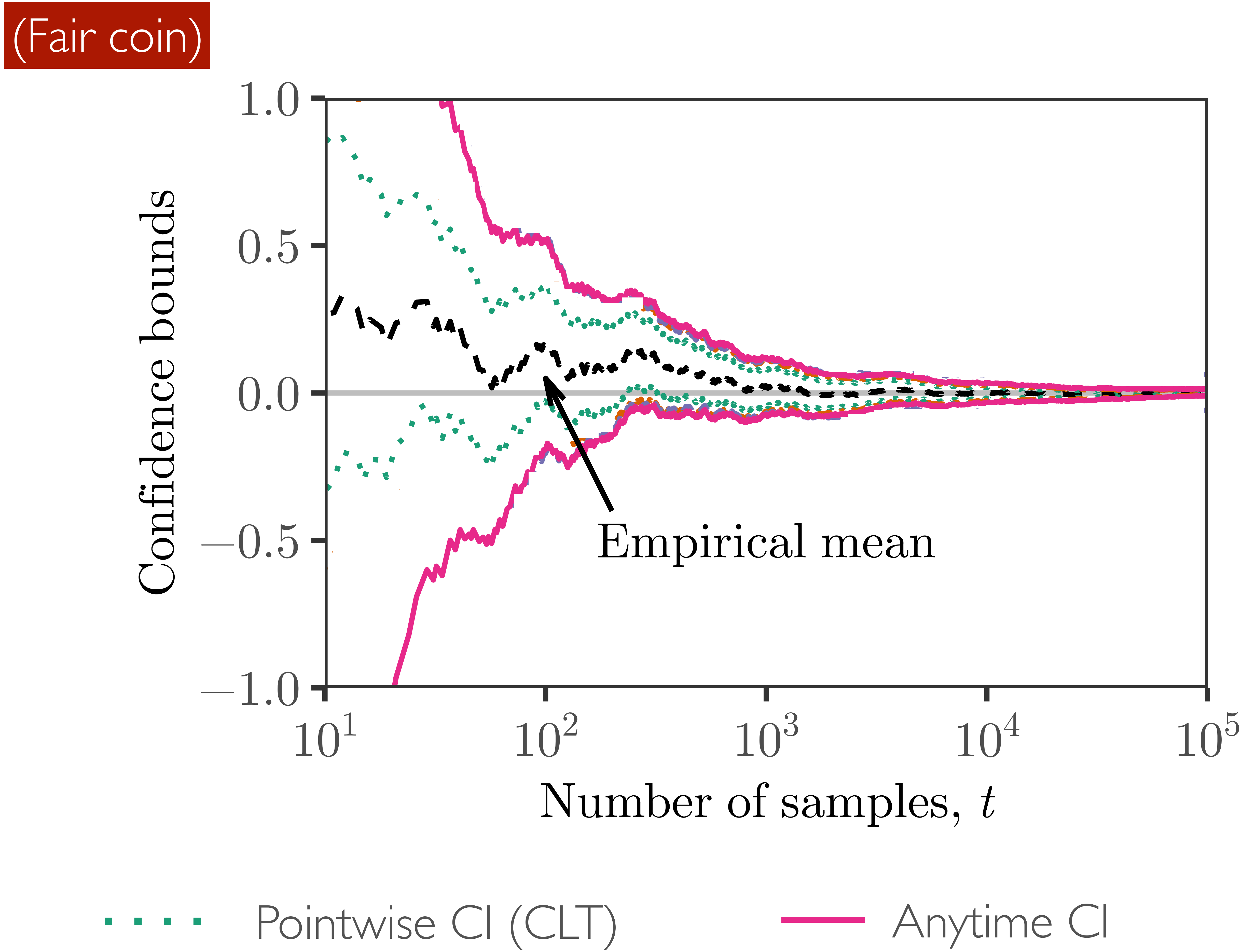
Producing a confidence *interval* at a fixed time is elementary statistics (~100 years old).

How do we produce a confidence sequence? (which is like a confidence band over time)

(Fair coin)



1.0
0.5
0.0
−0.5
−1.0
Confidence bounds
10^1
10^2
10^3
10^4
10^5
Number of samples, t
Pointwise CI (CLT)
Anytime CI



(Fair coin)
1.0
0.5
0.0
−0.5
−1.0
Confidence bounds
Empirical mean
10^1
10^2
10^3
10^4
10^5
Number of samples, t
Pointwise CI (CLT)
Anytime CI

Eg: If X_i is 1-subGaussian, then

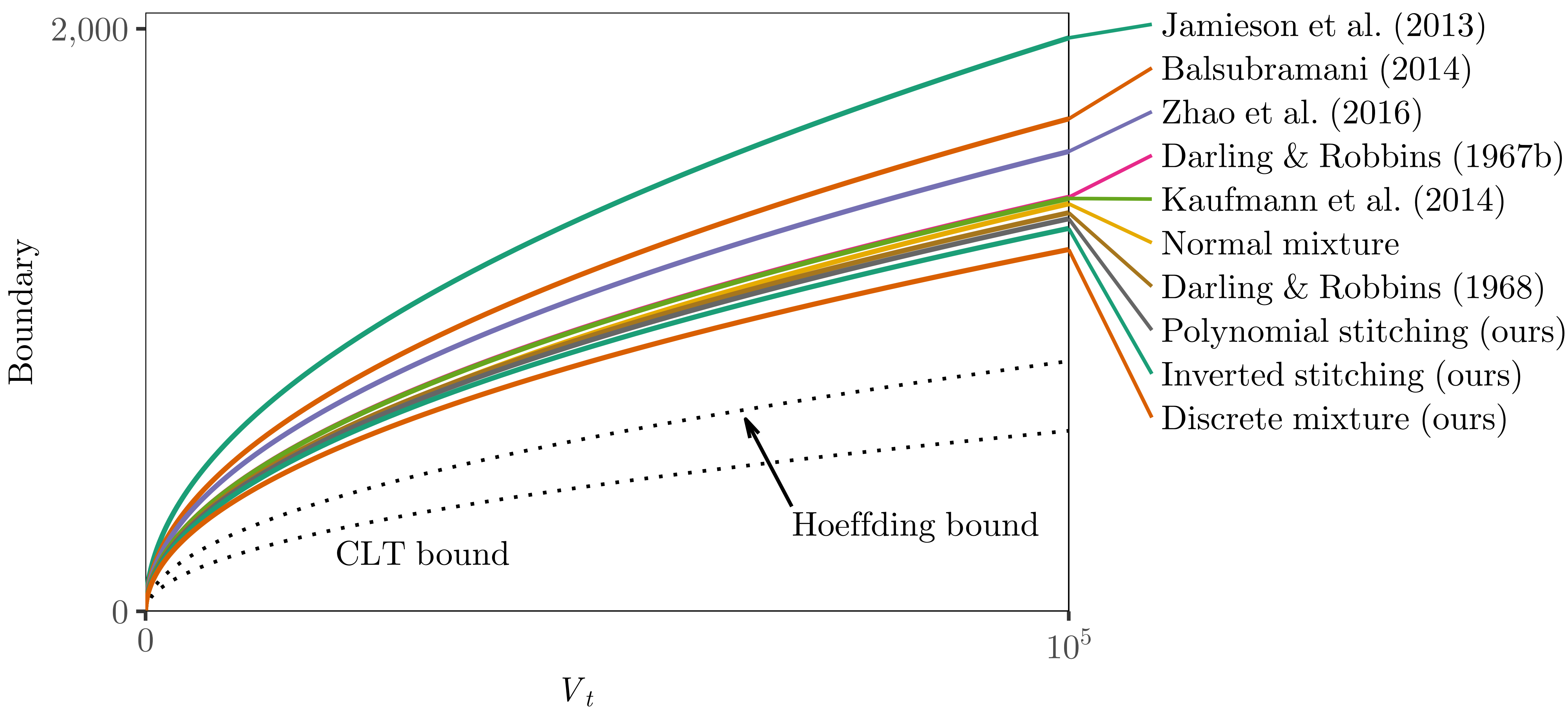
$$\frac{\sum_{i=1}^{n} X_i}{n} \pm 1.71\sqrt{\frac{\log\log(2n) + 0.72\log(5.19/\alpha)}{n}}$$

is a $(1-\alpha)$ confidence sequence.

Eg: If X_i is 1-subGaussian, then

$$\frac{\sum_{i=1}^{n} X_i}{n} \pm 1.71\sqrt{\frac{\log\log(2n) + 0.72\log(5.19/\alpha)}{n}}$$

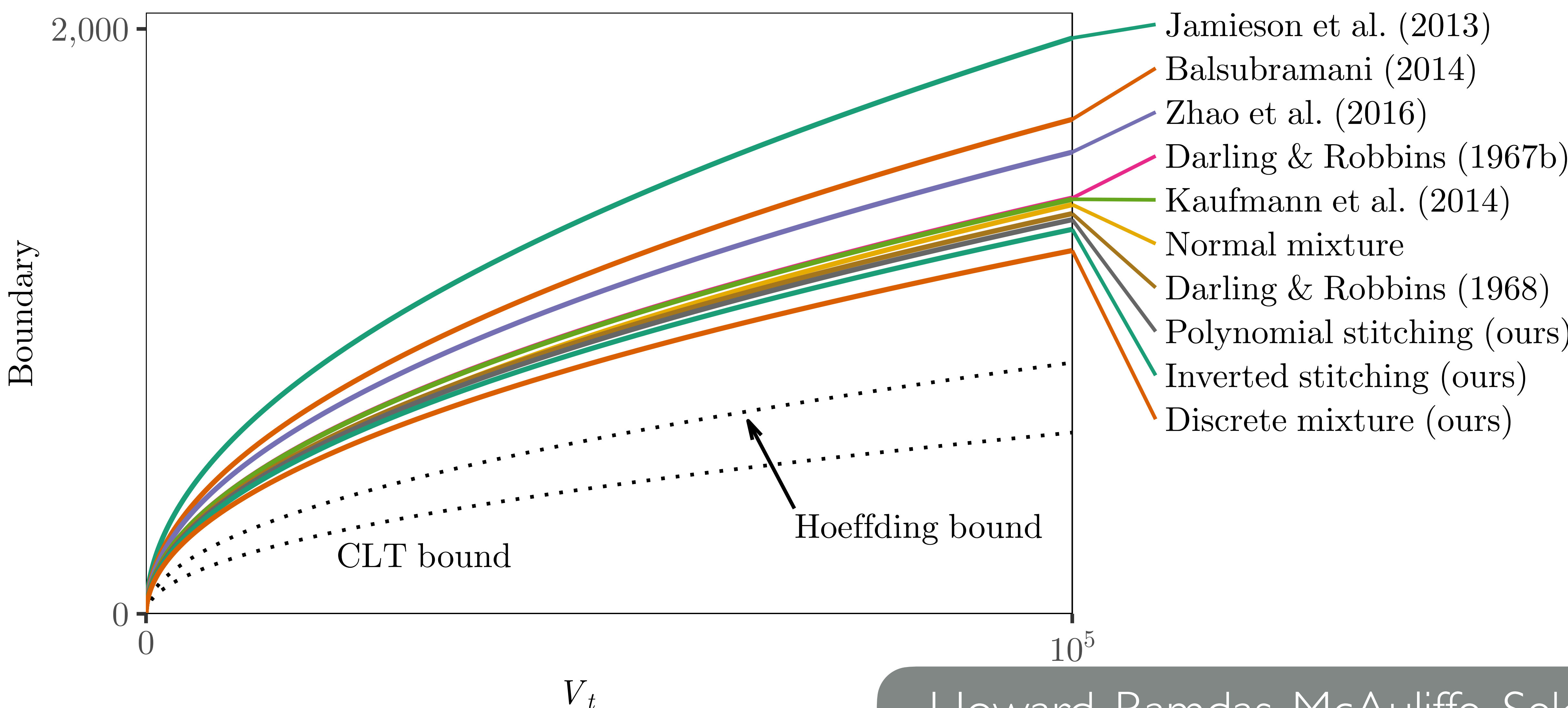
is a $(1-\alpha)$ confidence sequence.



Eg: If X_i is 1-subGaussian, then

$$\frac{\sum_{i=1}^{n} X_i}{n} \pm 1.71\sqrt{\frac{\log\log(2n) + 0.72\log(5.19/\alpha)}{n}}$$

is a $(1-\alpha)$ confidence sequence.



Howard, Ramdas, McAuliffe, Sekhon '18

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} \{\theta \notin (L_n, U_n)\}\Big) \leq \alpha\,.$$

$$\mathbb{P}\Big(\bigcup_{n \in \mathbb{N}} \{\theta \notin (L_n, U_n)\}\Big) \leq \alpha\,.$$

Some implications:

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \{\theta \notin (L_n, U_n)\}\right) \leq \alpha\,.$$

Some implications:

1. Valid inference at any time, even stopping times:

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \{\theta \notin (L_n, U_n)\}\right) \leq \alpha\,.$$

Some implications:

1. Valid inference at any time, even stopping times:

   For any stopping time τ : $\mathbb{P}(\theta \notin (L_\tau, U_\tau)) \leq \alpha\,.$

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}}\{\theta \notin (L_n, U_n)\}\Big) \le \alpha\,.$$

Some implications:

1. Valid inference at any time, even stopping times:

   For any stopping time τ : $\mathbb{P}(\theta \notin (L_\tau, U_\tau)) \le \alpha\,.$

2. Valid post-hoc inference (in hindsight):

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} \{\theta \notin (L_n, U_n)\}\Big) \leq \alpha \,.$$

Some implications:

1. Valid inference at any time, even stopping times:

   For any stopping time τ : $\mathbb{P}(\theta \notin (L_\tau, U_\tau)) \leq \alpha$.

2. Valid post-hoc inference (in hindsight):

   For any random time T : $\mathbb{P}(\theta \notin (L_T, U_T)) \leq \alpha$.

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} \{\theta \notin (L_n, U_n)\}\Big) \le \alpha\,.$$

Some implications:

1. Valid inference at any time, even stopping times:

   For any stopping time τ : $\mathbb{P}(\theta \notin (L_\tau, U_\tau)) \le \alpha\,.$

2. Valid post-hoc inference (in hindsight):

   For any random time T : $\mathbb{P}(\theta \notin (L_T, U_T)) \le \alpha\,.$

3. No pre-specified sample size:

   can extend or stop experiments adaptively.

The same duality between confidence intervals and p-values also holds in the sequential setting: "confidence sequences" are dual to "always valid p-values".

Duality between anytime p-value and CI

Duality between anytime p-value and CI

Define a set of null values $\mathscr{H}_0$ for θ.

Duality between anytime p-value and CI

Define a set of null values $\mathscr{H}_0$ for θ.

Let $P^{(n)} := \inf\{\alpha :$ the $(1-\alpha)$ $CI^{(n)}$ does not intersect $\mathscr{H}_0\}$

Duality between anytime p-value and CI

Define a set of null values $\mathcal{H}_0$ for θ.

Let $P^{(n)} := \inf\{\alpha :$ the $(1-\alpha)$ $CI^{(n)}$ does not intersect $\mathcal{H}_0\}$

If $CI^{(n)}$ is a pointwise CI then $P^{(n)}$ is a classical p-value.

For all fixed times n, $\underbrace{\Pr(P^{(n)} \leq \alpha)}_{\text{prob. of false positive}} \leq \alpha$.

Duality between anytime p-value and CI

Define a set of null values $\mathscr{H}_0$ for θ.

Let $P^{(n)} := \inf\{\alpha :$ the $(1-\alpha)$ $CI^{(n)}$ does not intersect $\mathscr{H}_0\}$

If $CI^{(n)}$ is a pointwise CI then $P^{(n)}$ is a classical p-value.

For all fixed times n, $\underbrace{\Pr(P^{(n)} \leq \alpha)}_{\text{prob. of false positive}} \leq \alpha$.

If $CI^{(n)}$ is an anytime CI then $P^{(n)}$ is an always-valid p-value.

For all stopping times τ, $\Pr(P^{(\tau)} \leq \alpha) \leq \alpha$.

For all data-dependent times T, $\Pr(P^{(T)} \leq \alpha) \leq \alpha$.

Relationship to Sequential Probability Ratio Test

Given a stream of data $X_1, X_2, \ldots \sim f_\theta$, suppose we want to test a null hypothesis $H_0 : \theta = \theta_0$ against an alternative hypothesis $H_1 : \theta = \theta_1$.

Wald '48

Relationship to Sequential Probability Ratio Test

Given a stream of data $X_1, X_2, \ldots \sim f_\theta$, suppose we want to test a null hypothesis $H_0 : \theta = \theta_0$ against an alternative hypothesis $H_1 : \theta = \theta_1$.

Wald's SPRT (or SLRT) calculates a probability/likelihood ratio:

$$L^{(n)} := \frac{\prod_{i=1}^{n} f_1(X_i)}{\prod_{i=1}^{n} f_0(X_i)},$$

and rejects when $L^{(n)} > 1/\alpha$. Can also use prior/mixture over θ_1.

Wald '48

Relationship to Sequential Probability Ratio Test

Given a stream of data $X_1, X_2, \ldots \sim f_\theta$, suppose we want to test a null hypothesis $H_0 : \theta = \theta_0$ against an alternative hypothesis $H_1 : \theta = \theta_1$.

Wald's SPRT (or SLRT) calculates a probability/likelihood ratio:

$$L^{(n)} := \frac{\prod_{i=1}^{n} f_1(X_i)}{\prod_{i=1}^{n} f_0(X_i)},$$

and rejects when $L^{(n)} > 1/\alpha$. Can also use prior/mixture over θ_1.

Equivalently, define $P^{(n)} = 1/L^{(n)}$. Then $P^{(n)}$ is an always-valid p-value.

Wald '48

Relationship to Sequential Probability Ratio Test

Given a stream of data $X_1, X_2, \ldots \sim f_\theta$, suppose we want to test a null hypothesis $H_0 : \theta = \theta_0$ against an alternative hypothesis $H_1 : \theta = \theta_1$.

Wald's SPRT (or SLRT) calculates a probability/likelihood ratio:

$$L^{(n)} := \frac{\prod_{i=1}^{n} f_1(X_i)}{\prod_{i=1}^{n} f_0(X_i)},$$

and rejects when $L^{(n)} > 1/\alpha$. Can also use prior/mixture over θ_1.

Equivalently, define $P^{(n)} = 1/L^{(n)}$. Then $P^{(n)}$ is an always-valid p-value.

(And inverting it defines a confidence sequence.)

Wald '48

Can construct confidence sequences
(and hence always valid p-values)
in a wide variety of *nonparametric* settings
(eg: random variables that are
bounded, or subGaussian, or subexponential)

Howard, Ramdas, McAuliffe, Sekhon '18

Solutions for these issues

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Outer sequential process:

Part II

"false coverage rate" for estimation

(correspondingly, "false discovery rate" for testing)

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Outer sequential process:

Part II

"false coverage rate" for estimation

(correspondingly, "false discovery rate" for testing)

Modular solutions: fit well together

Many extensions to each piece

Part III

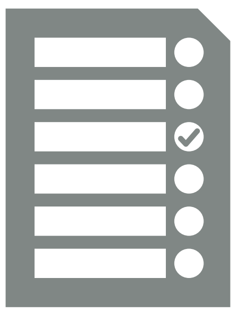
Part II

The OUTER Sequential Process (a sequence of experiments)

[40 mins]

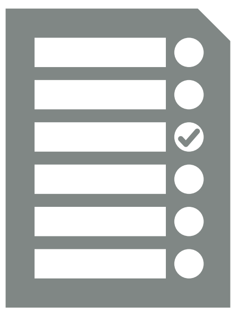
Quick recap of **A/B testing**

A :

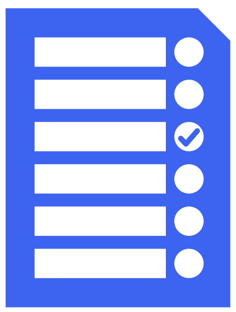


Quick recap of **A/B testing**

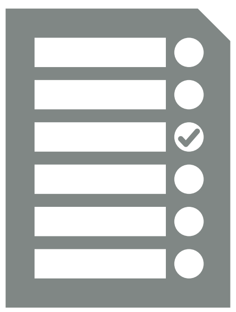
A :

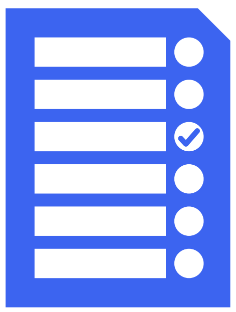


B :



Quick recap of **A/B testing**

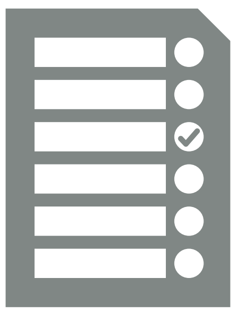
A : 

B : 

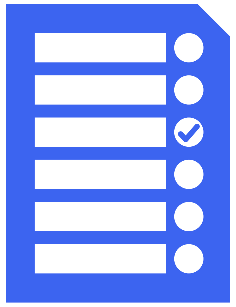
Null hypothesis:
A is at least
as good as B.

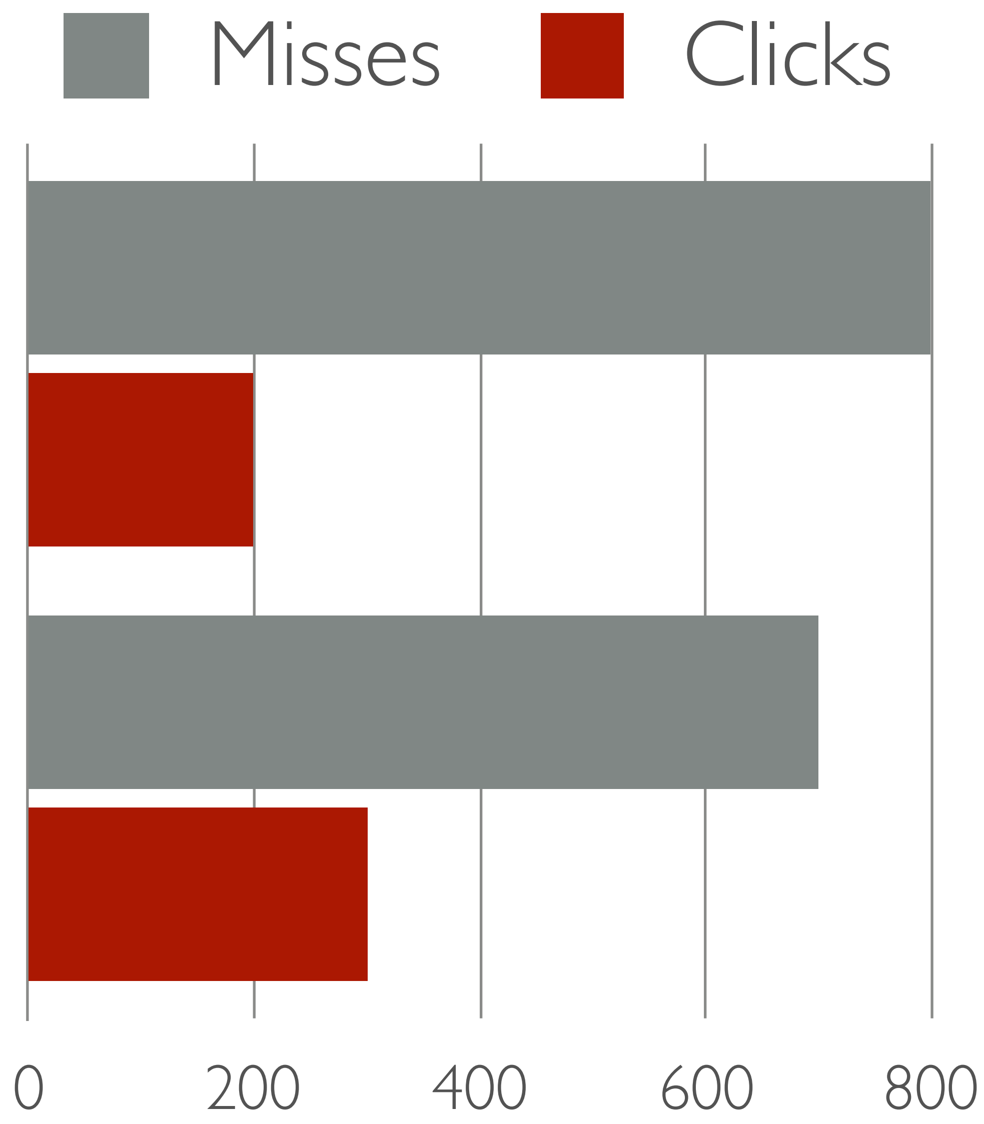
Quick recap of **A/B testing**

A :



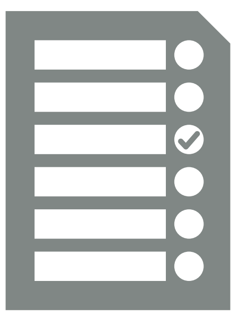
B :

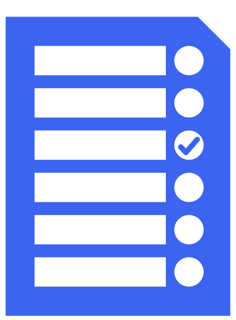


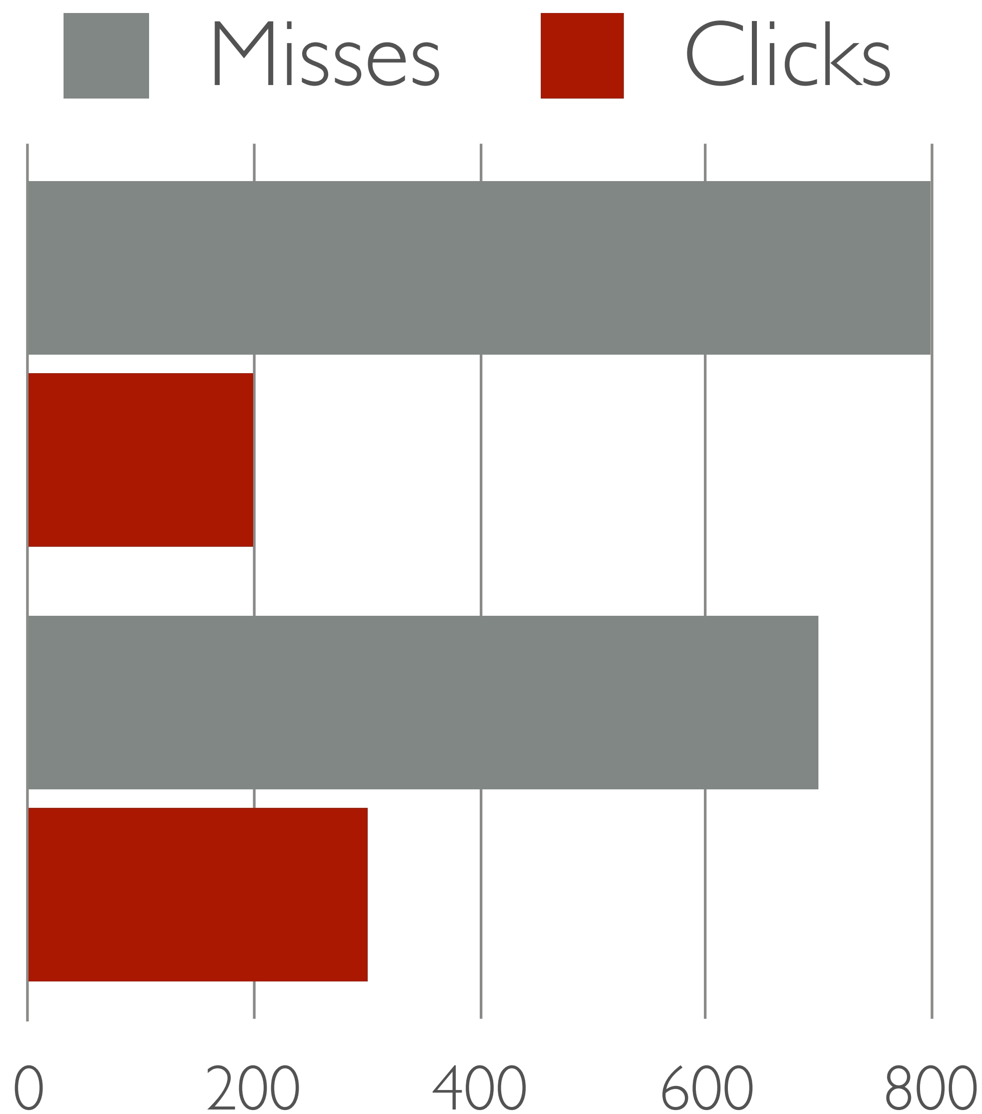


Null hypothesis:
A is at least
as good as B.

Quick recap of **A/B testing**

A : 

B : 

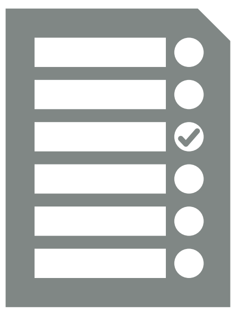


Null hypothesis:
A is at least
as good as B.

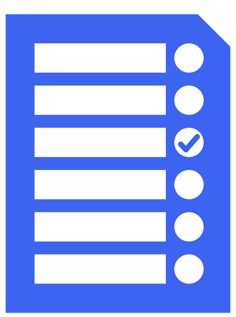
Calculate p-value:
P = Pr(observed data or more extreme, assuming null is true)

Quick recap of **A/B testing**

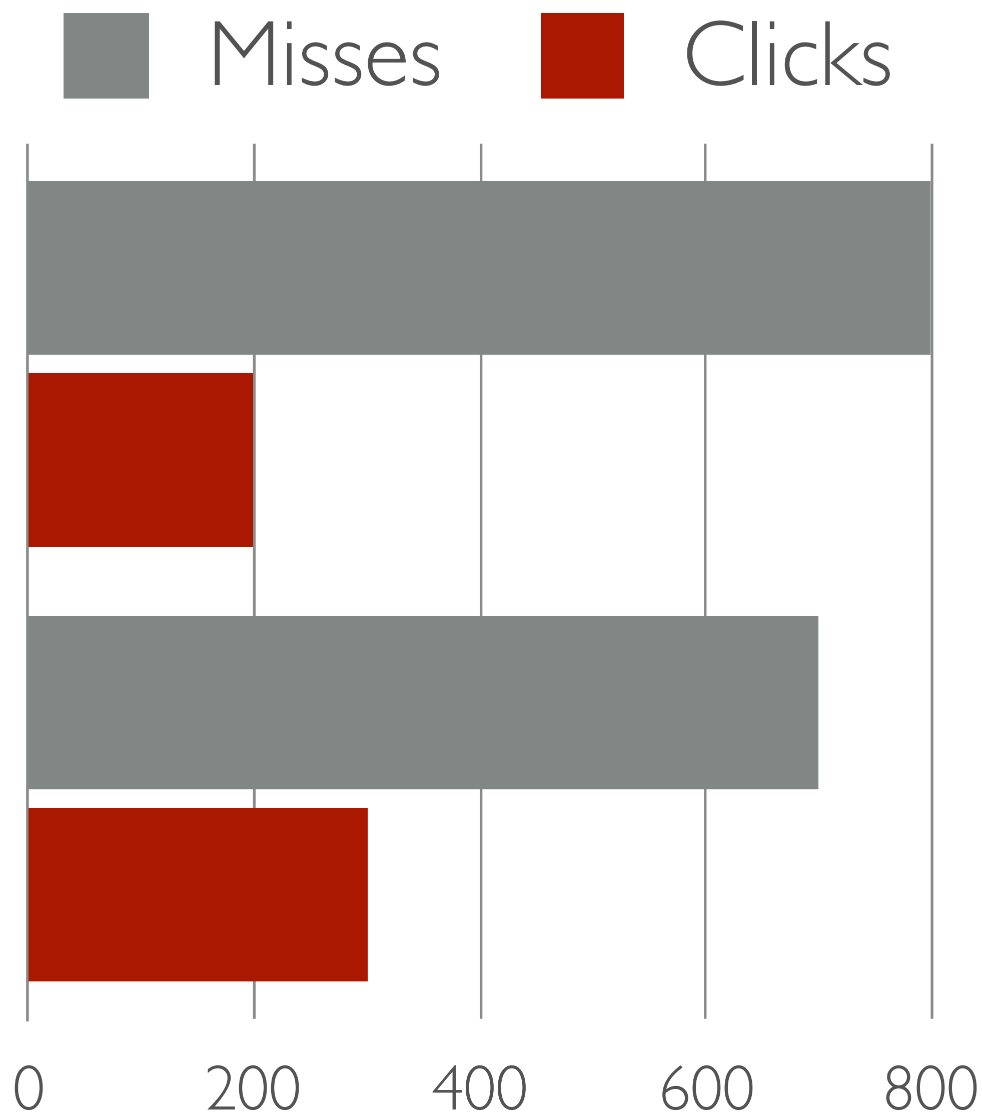
A :



B :



Null hypothesis:
A is at least
as good as B.

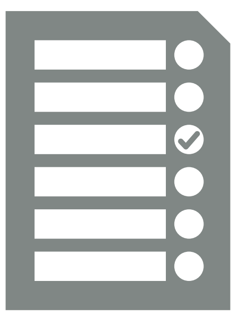


Calculate p-value:
P = Pr(observed data or more extreme, assuming null is true)

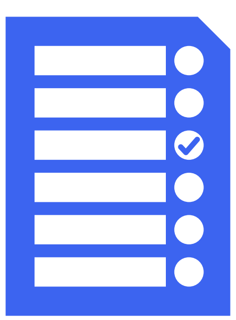
Decision rule :
if $P \leq \alpha$, **then we reject the null ("discovery").**
We **change A to B,** ensuring that
type-1 error $\leq \alpha$ **.**

Quick recap of **A/B testing**

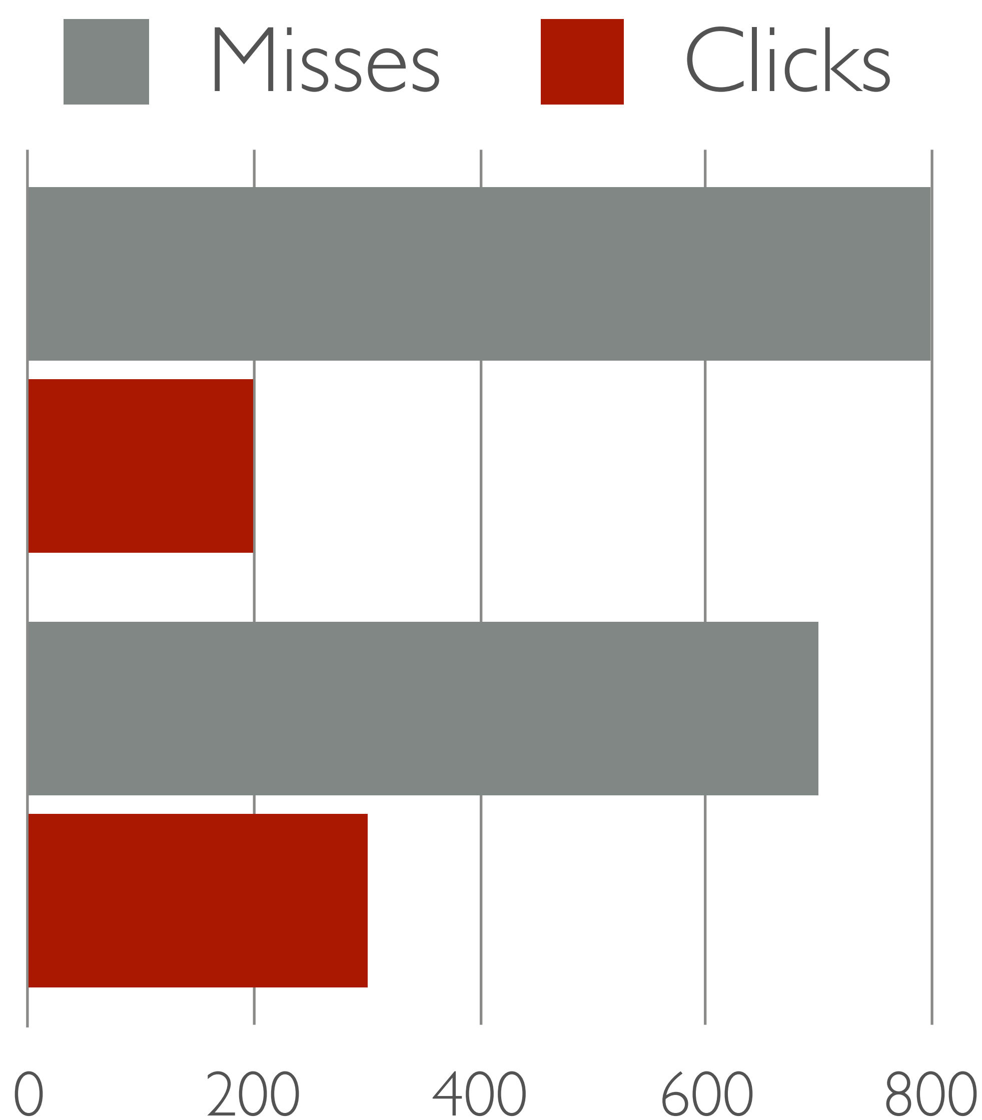
A :



B :



Null hypothesis:
A is at least as good as B.



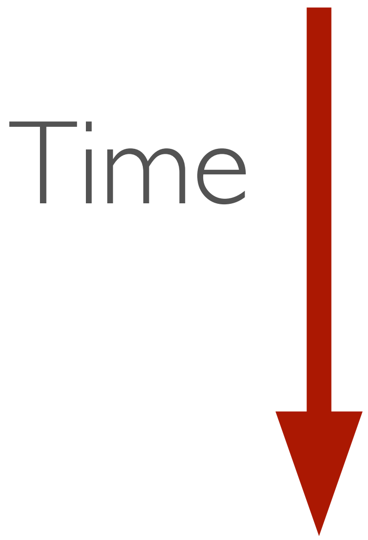
Calculate p-value:
P = Pr(observed data or more extreme, assuming null is true)

Decision rule :
if $P \leq \alpha$, **then we reject the null ("discovery").**
We **change A to B,** ensuring that **type-I error** $\leq \alpha$.

a wrong rejection of the null is a false discovery and implies **a bad change from A to B.**

Reality: internet companies run thousands of different (independent) A/B tests over time.

Reality: internet companies run thousands of different (independent) A/B tests over time.



Reality: internet companies run thousands of different (independent) A/B tests over time.

Reality: internet companies run thousands of different (independent) A/B tests over time.

Decision rule:

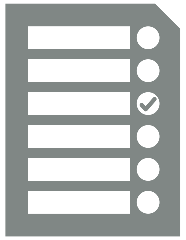
Color

Time

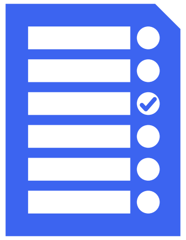
Reality: internet companies run thousands of different (independent) A/B tests over time.

Decision rule:

$P_1 \leq \alpha?$



vs.



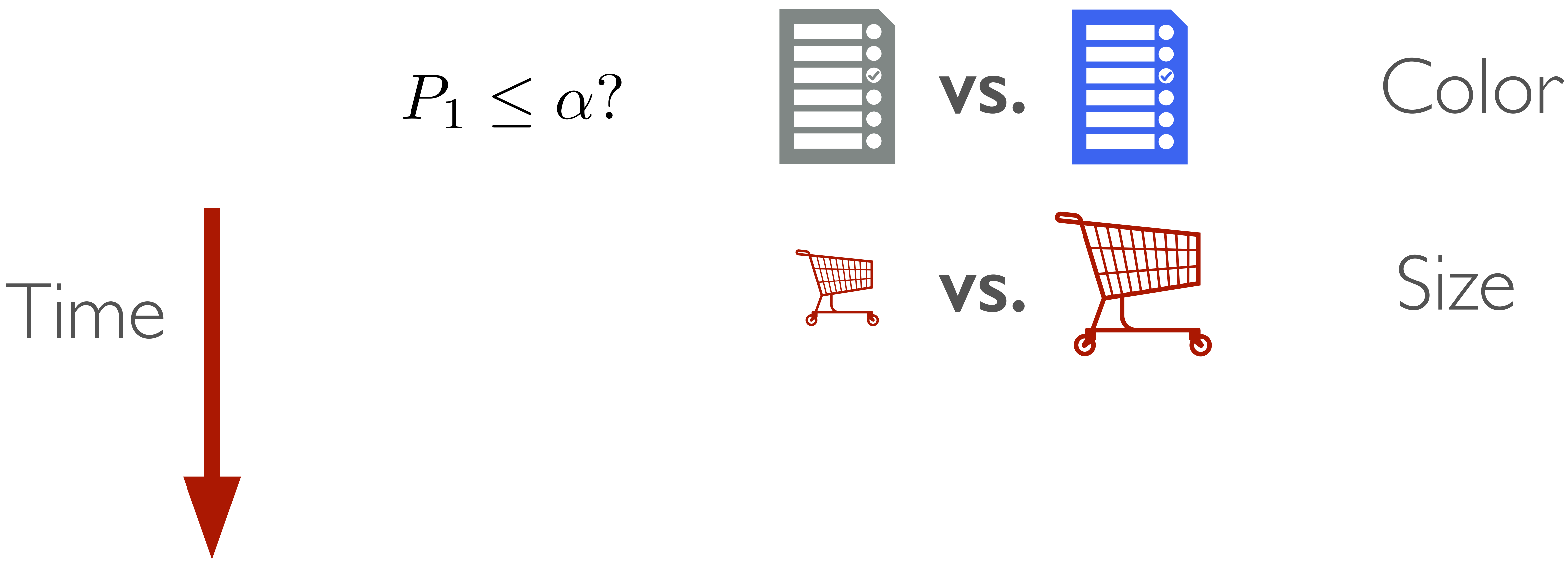
Color

Time

Reality: internet companies run thousands of different (independent) A/B tests over time.

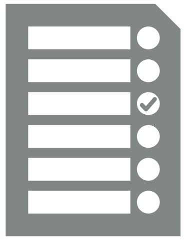
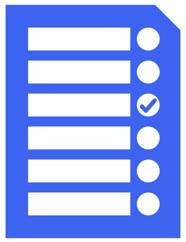
Decision rule:

$$P_1 \leq \alpha?$$



Reality: internet companies run thousands of different (independent) A/B tests over time.

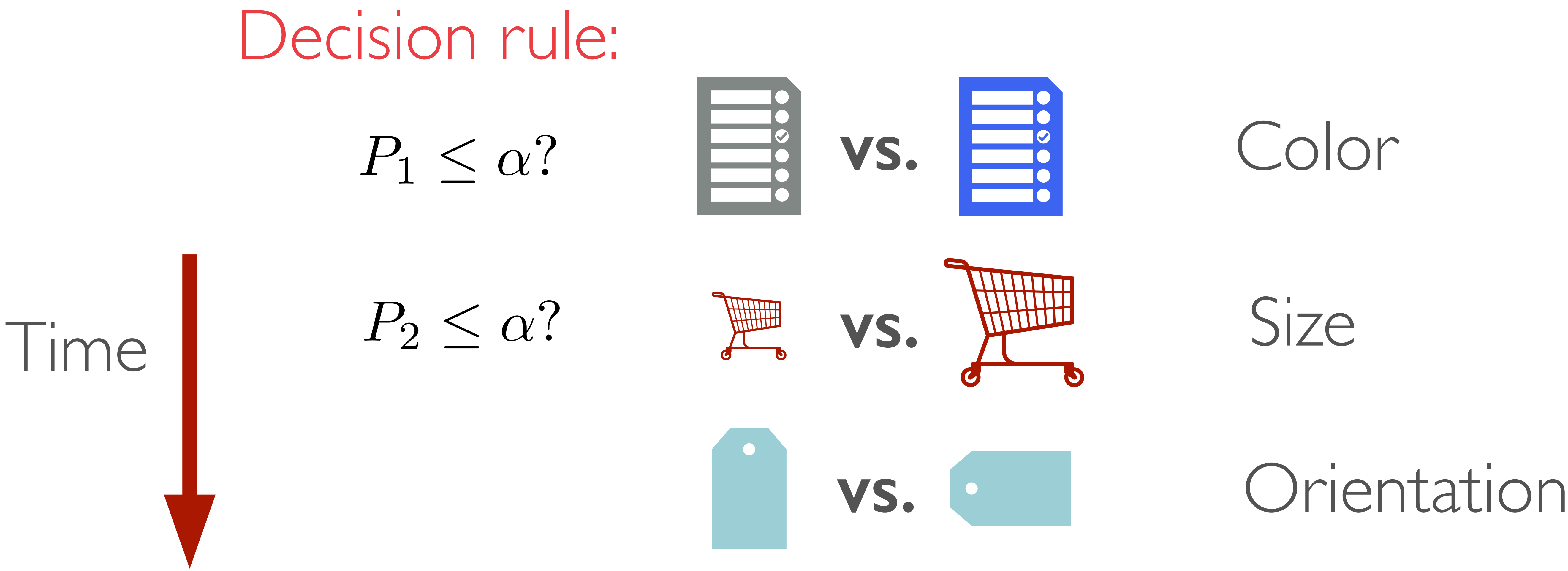
Decision rule:

$P_1 \leq \alpha?$  **vs.**  Color

Time

$P_2 \leq \alpha?$  **vs.**  Size

Reality: internet companies run thousands of different (independent) A/B tests over time.



Reality: internet companies run thousands of different (independent) A/B tests over time.

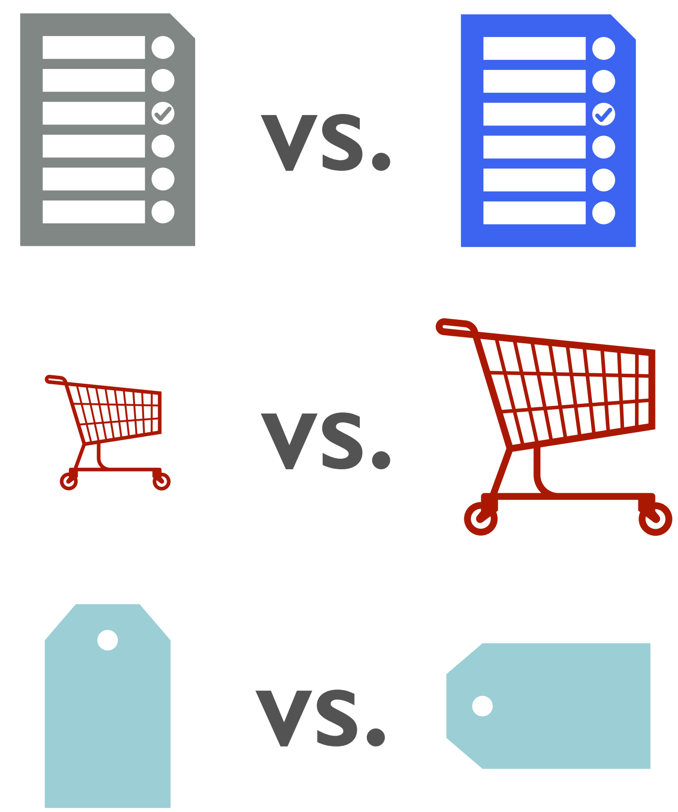
Decision rule:

Time ↓

$P_1 \leq \alpha?$ — Color

$P_2 \leq \alpha?$ — Size

$P_3 \leq \alpha?$ — Orientation



Reality: internet companies run thousands of different (independent) A/B tests over time.

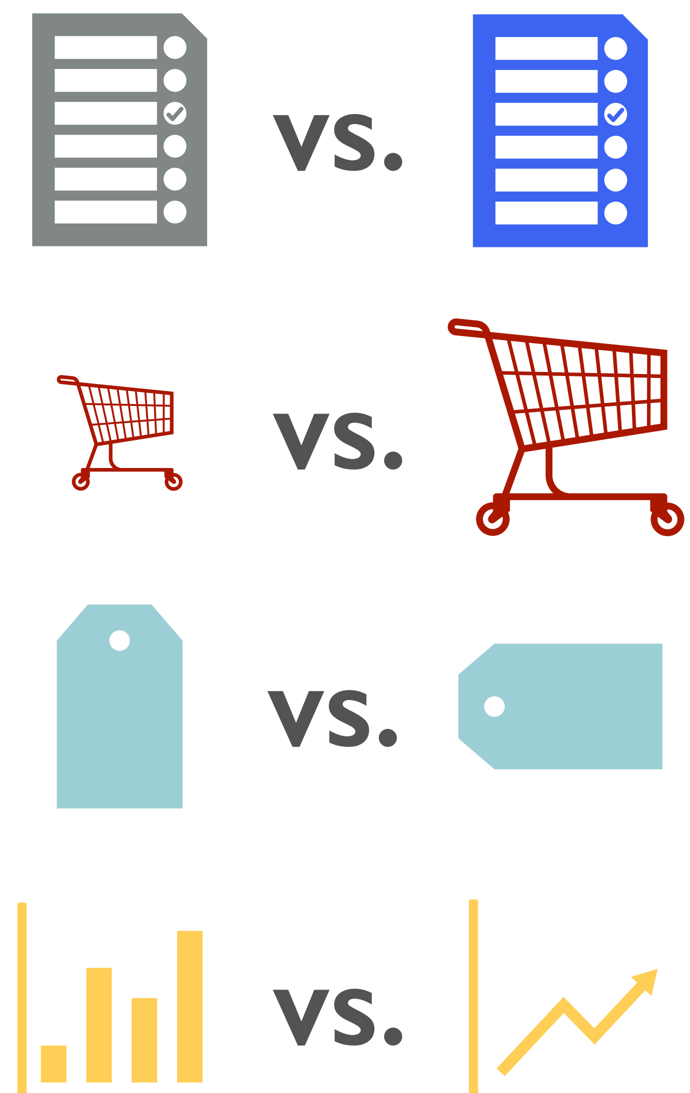
Decision rule:

Time ↓

$P_1 \leq \alpha?$

$P_2 \leq \alpha?$

$P_3 \leq \alpha?$



Color

Size

Orientation

Style

Reality: internet companies run thousands of different (independent) A/B tests over time.

Decision rule:

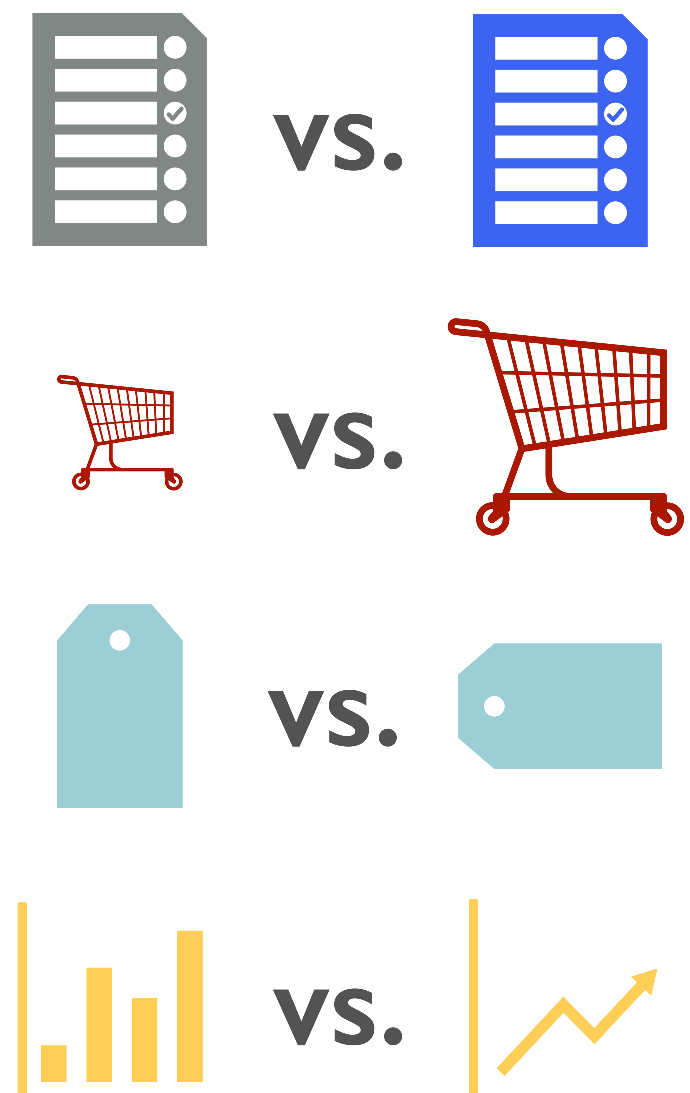
Time

$P_1 \leq \alpha?$ Color

$P_2 \leq \alpha?$ Size

$P_3 \leq \alpha?$ Orientation

$P_4 \leq \alpha?$ Style



Reality: internet companies run thousands of different (independent) A/B tests over time.

Decision rule:

Time ↓

$P_1 \leq \alpha?$ — Color

$P_2 \leq \alpha?$ — Size

$P_3 \leq \alpha?$ — Orientation

$P_4 \leq \alpha?$ — Style

Logo

Reality: internet companies run thousands of different (independent) A/B tests over time.

Decision rule:

Time

$P_1 \leq \alpha?$ — Color

$P_2 \leq \alpha?$ — Size

$P_3 \leq \alpha?$ — Orientation

$P_4 \leq \alpha?$ — Style

$P_5 \leq \alpha?$ — Logo

Reality: internet companies run thousands of different (independent) A/B tests over time.

Decision rule:

Time

$P_1 \leq \alpha?$

$P_2 \leq \alpha?$

$P_3 \leq \alpha?$

$P_4 \leq \alpha?$

$P_5 \leq \alpha?$

Color

Size

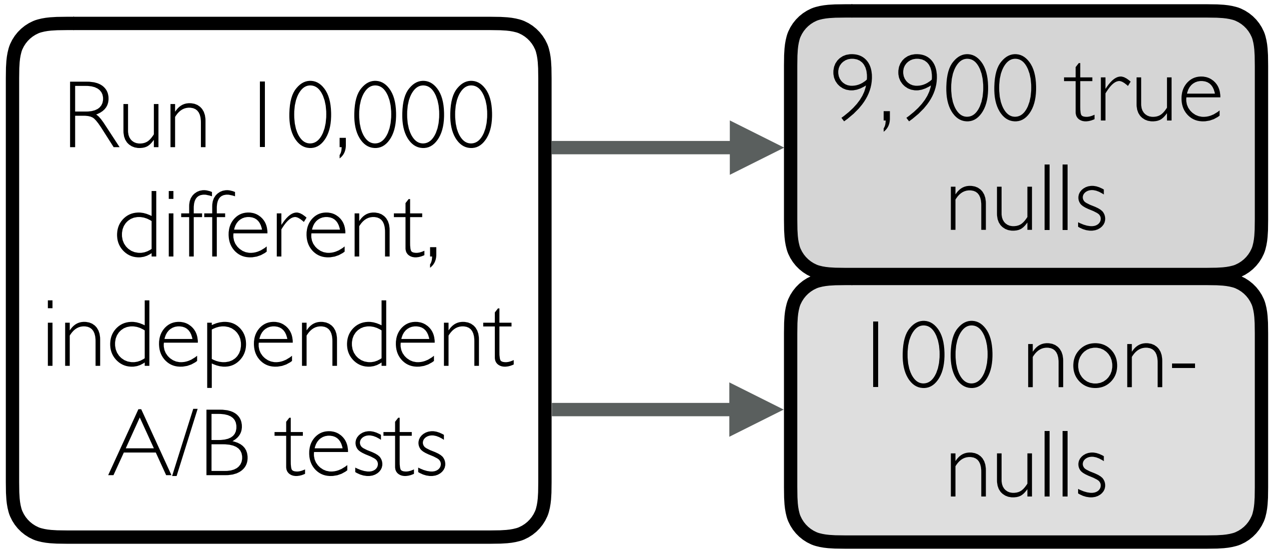
Orientation

Style

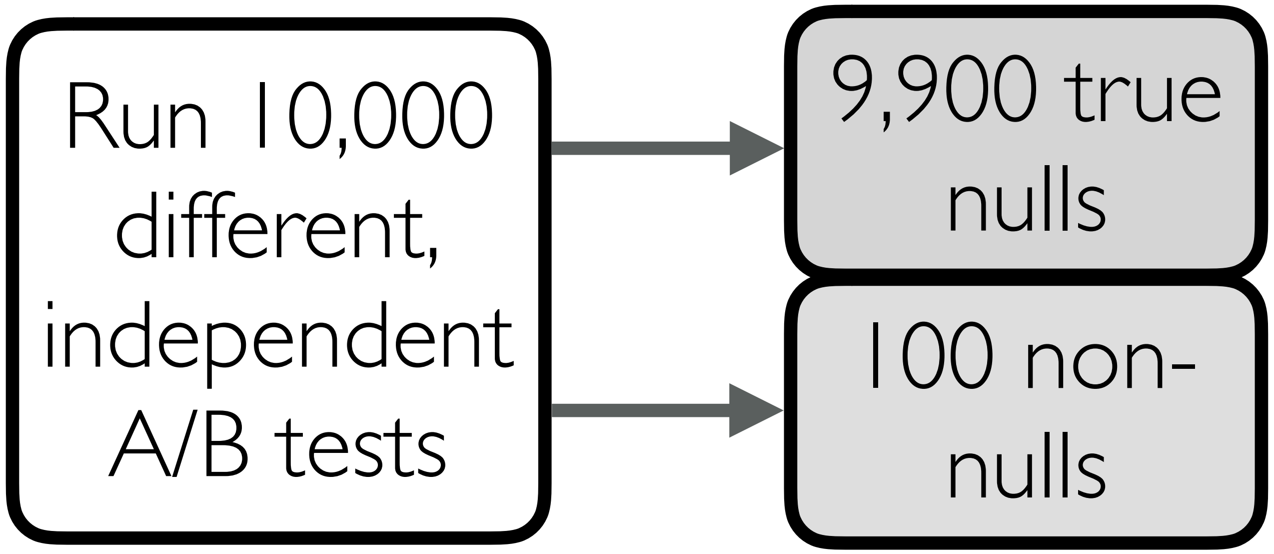
Logo

Problem!

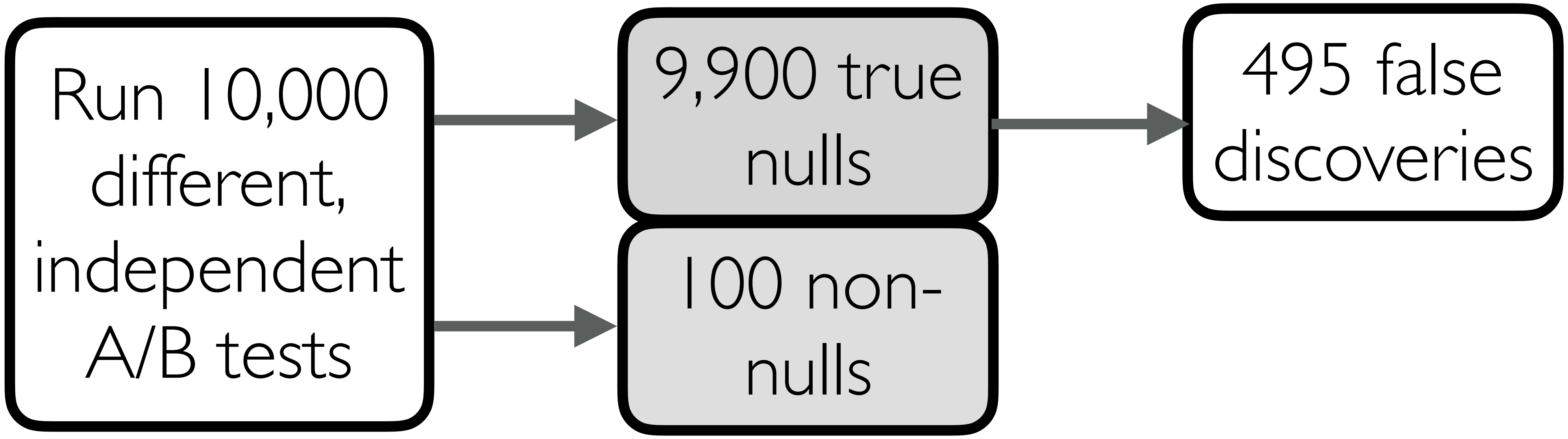
Run 10,000 different, independent A/B tests

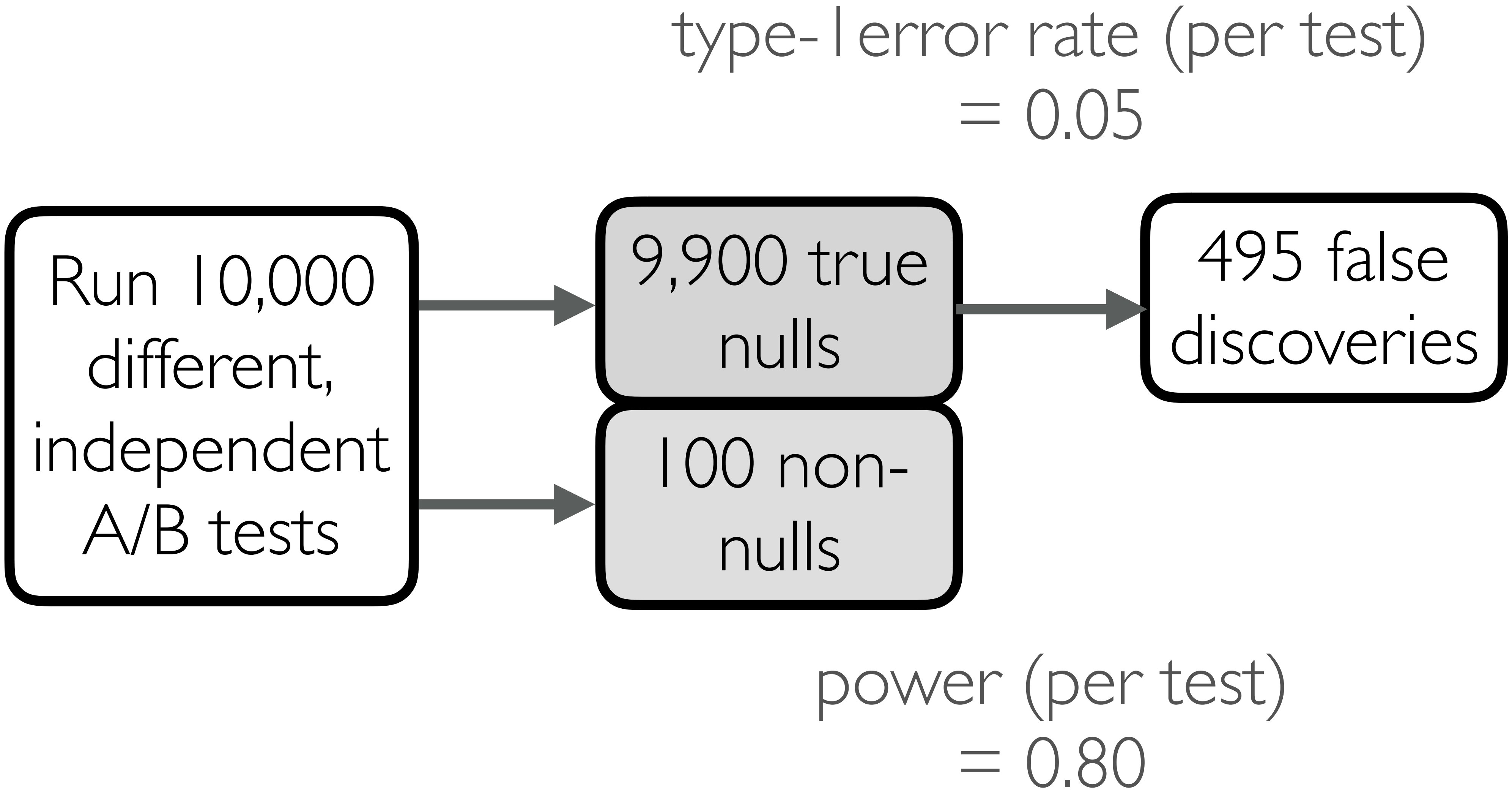

Run 10,000 different, independent A/B tests
9,900 true nulls
100 non-nulls

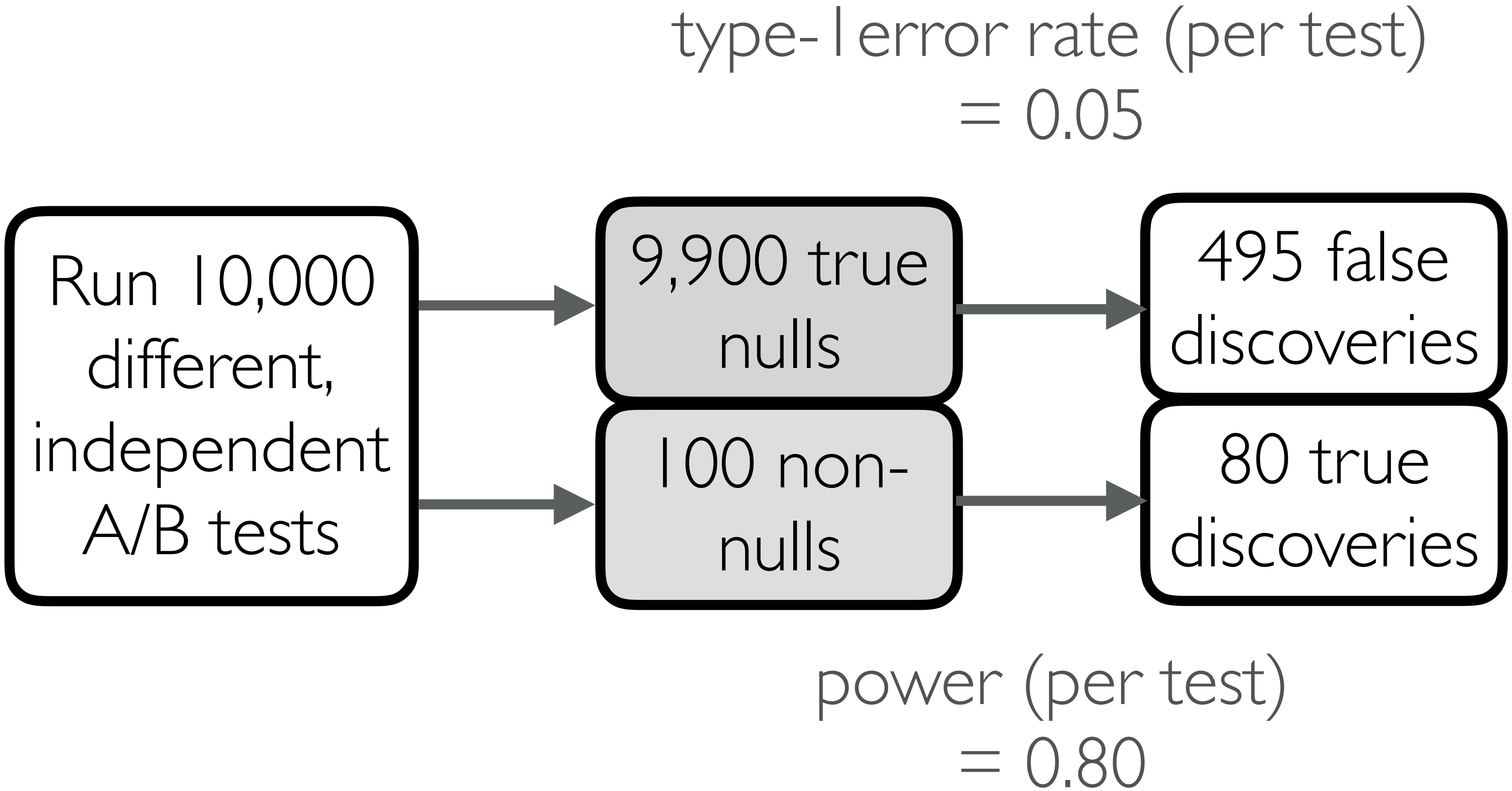
type-I error rate (per test)
= 0.05

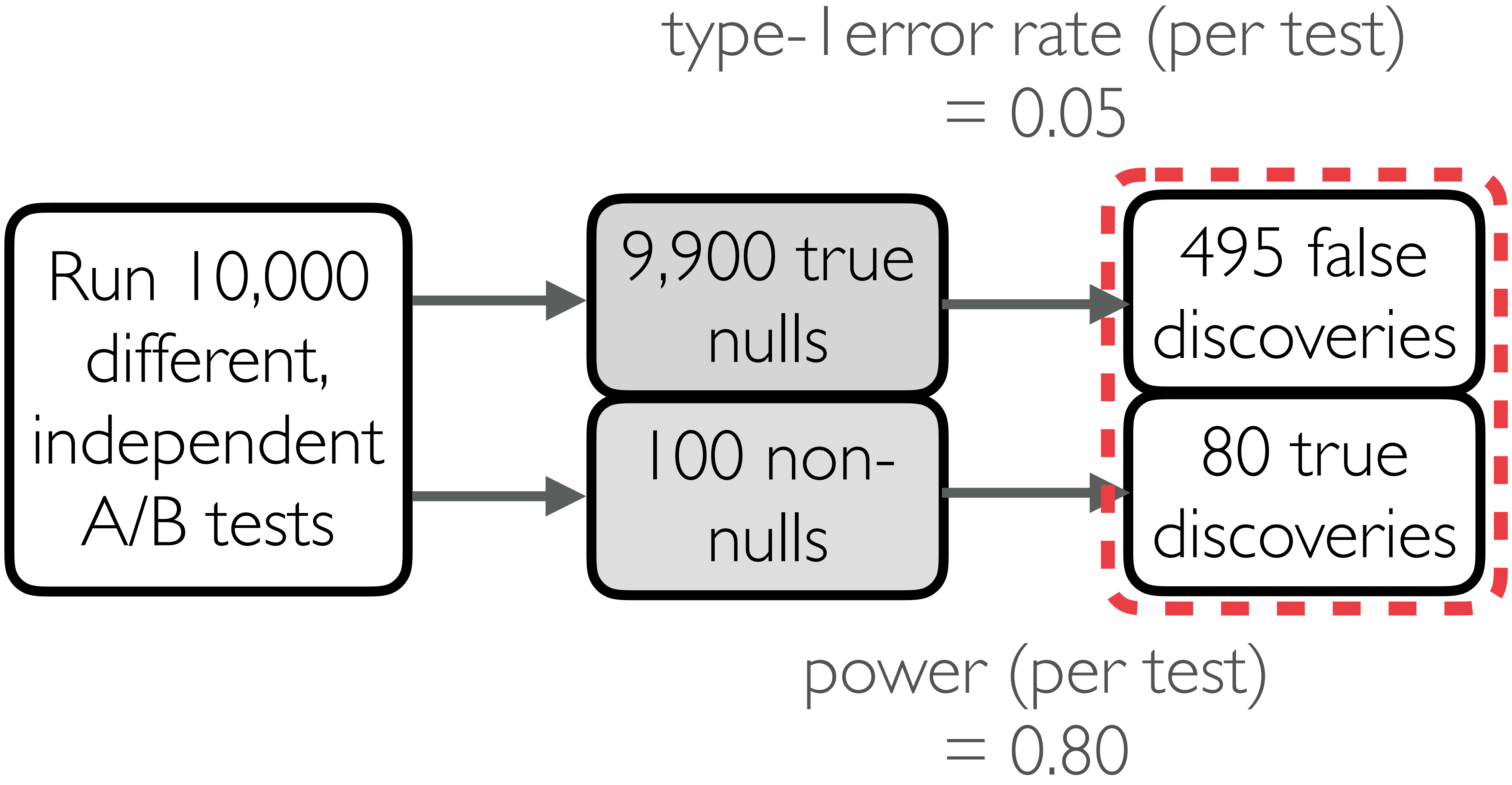


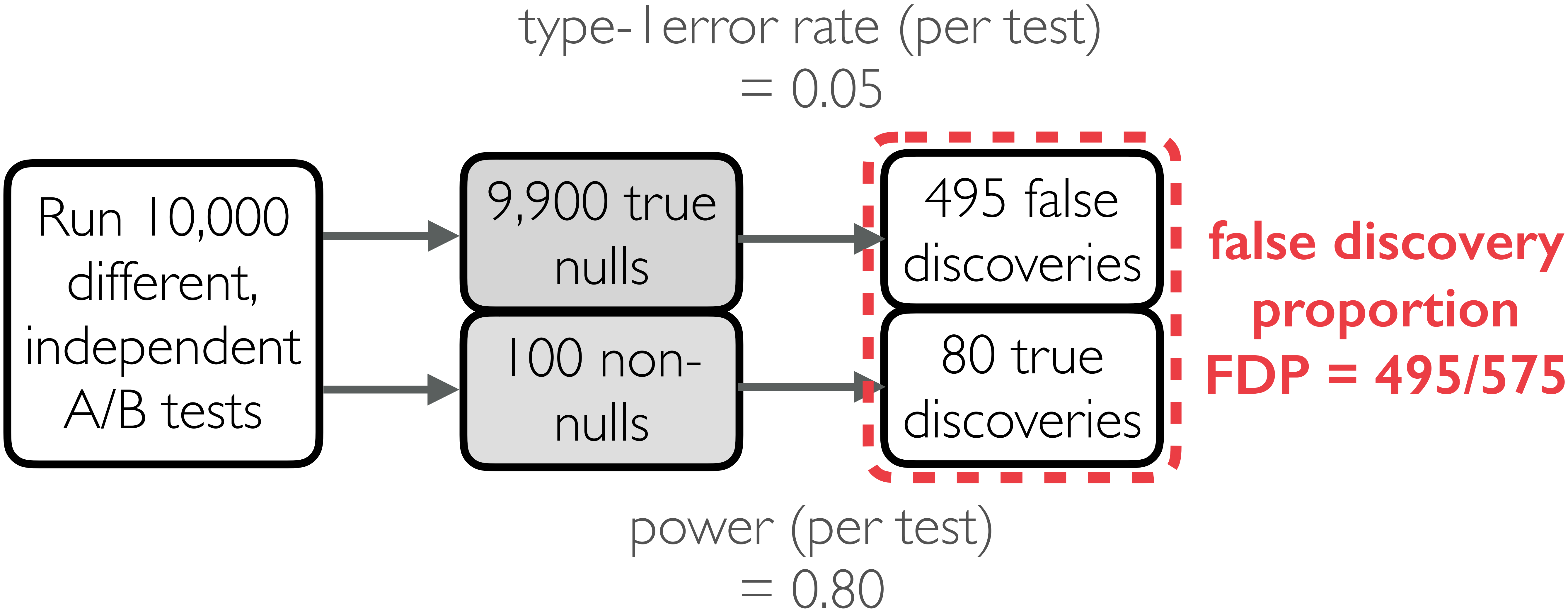
type-I error rate (per test)
= 0.05



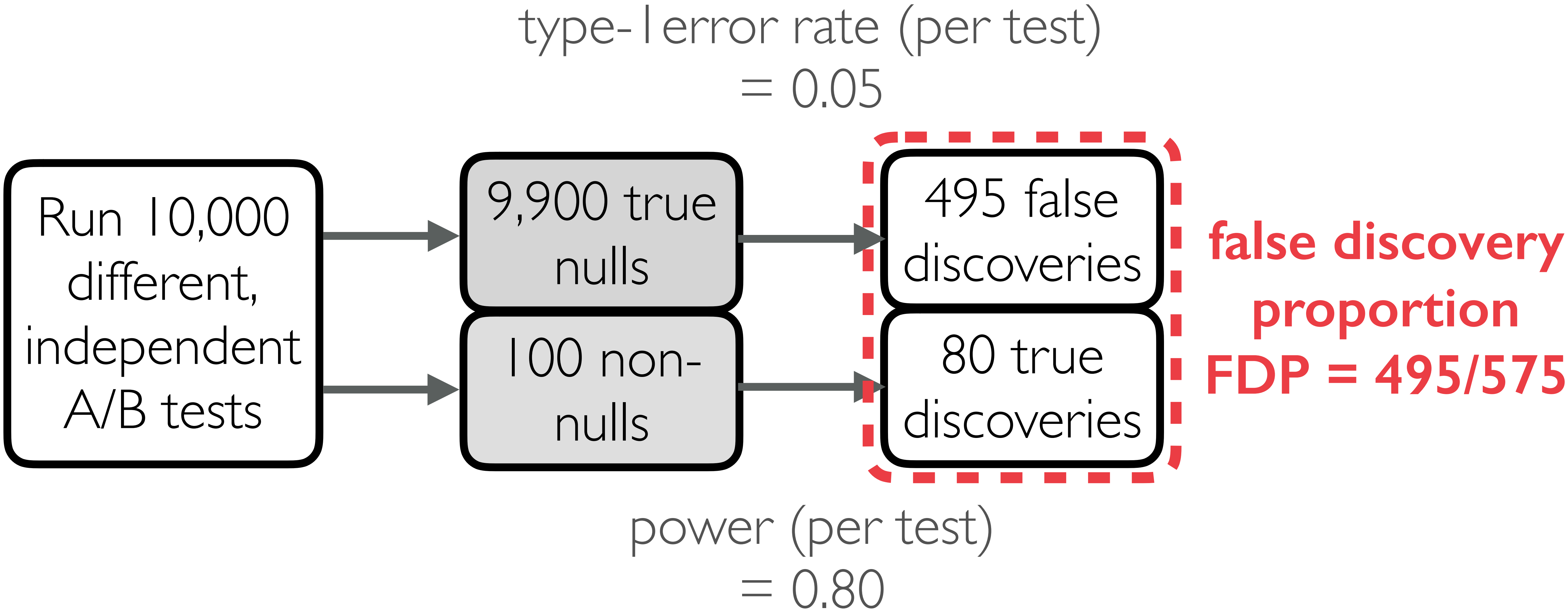

type-I error rate (per test)
= 0.05
Run 10,000 different, independent A/B tests
9,900 true nulls
100 non-nulls
495 false discoveries
power (per test)
= 0.80


type-I error rate (per test)
= 0.05
Run 10,000 different, independent A/B tests
9,900 true nulls
100 non-nulls
495 false discoveries
80 true discoveries
power (per test)
= 0.80


type-I error rate (per test)
= 0.05
Run 10,000 different, independent A/B tests
9,900 true nulls
100 non-nulls
495 false discoveries
80 true discoveries
power (per test)
= 0.80

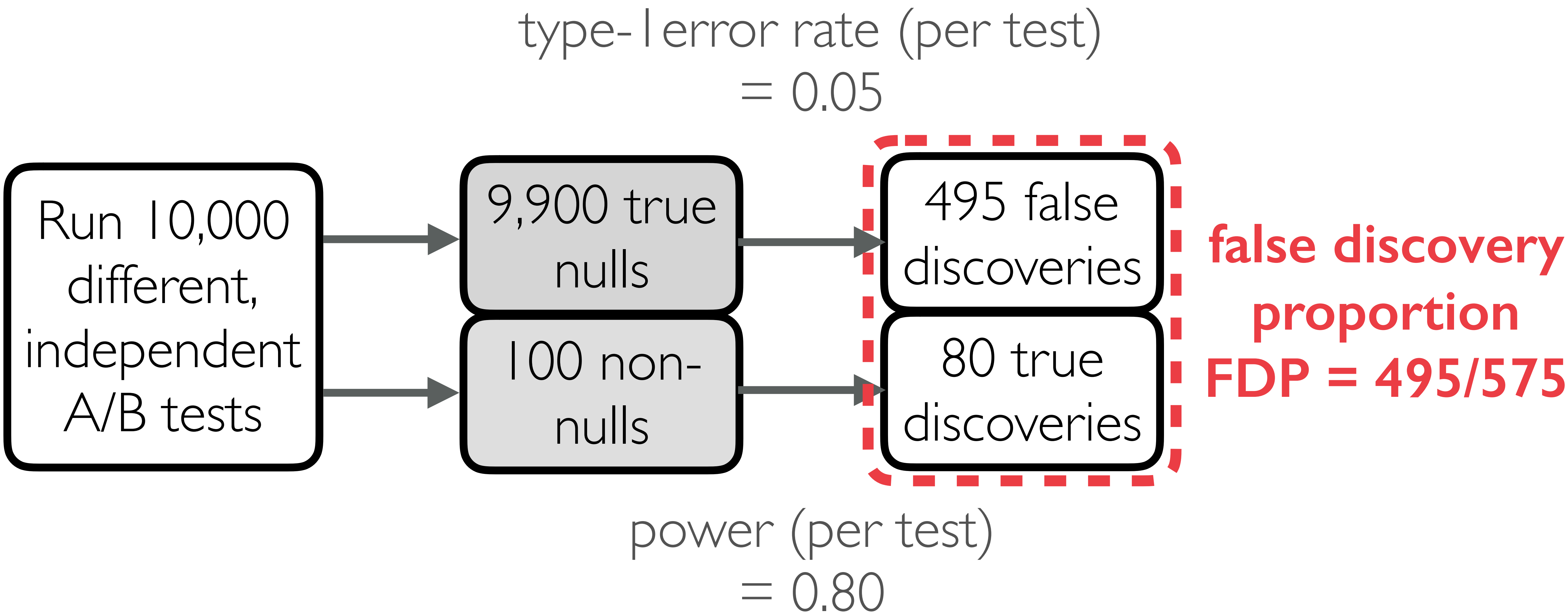


$$\text{FDP} = \frac{\#\ \text{false discoveries}}{\#\ \text{discoveries}}$$



$$\text{FDP} = \frac{\#\text{ false discoveries}}{\#\text{ discoveries}}$$

$$\text{FDR} = \mathbb{E}[\text{FDP}]$$



$$\mathrm{FDP} = \frac{\#\ \text{false discoveries}}{\#\ \text{discoveries}}$$

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$$

Summary: FDR can be larger than per-test error rate.
(even if hypotheses, tests, data are independent)

Given a possibly infinite sequence of independent tests (p-values), can we guarantee control of the FDR in a fully online fashion?

Foster-Stine '08
Aharoni-Rosset '14
Javanmard-Montanari '16
Ramdas-Yang-Wainwright-Jordan '17
Ramdas-Zrnic-Wainwright-Jordan '18
Tian-Ramdas '19

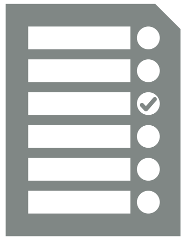
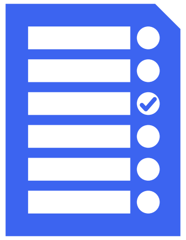
The aim of online FDR procedures

Decision rule:

Time

The aim of online FDR procedures

Decision rule:

 **vs.** 

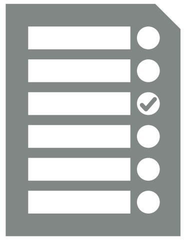
Color

Time

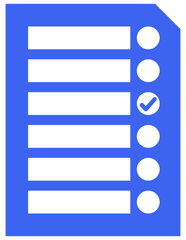
The aim of online FDR procedures

Decision rule:

$P_1 \leq \alpha_1$?



vs.



Color

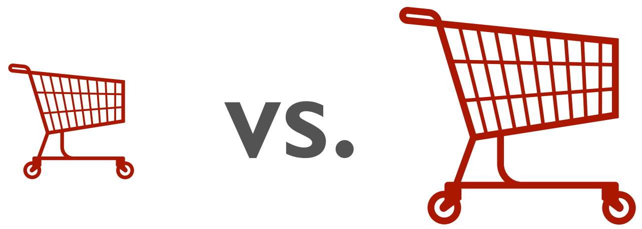
Time

The aim of online FDR procedures

Decision rule:

$P_1 \leq \alpha_1$?

Color

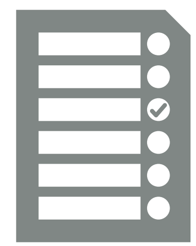
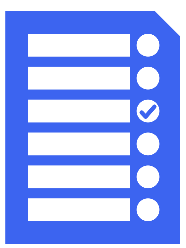


Size

Time

The aim of online FDR procedures

Decision rule:

$P_1 \leq \alpha_1$?  **vs.**  Color

$P_2 \leq \alpha_2$?  **vs.**  Size

Time

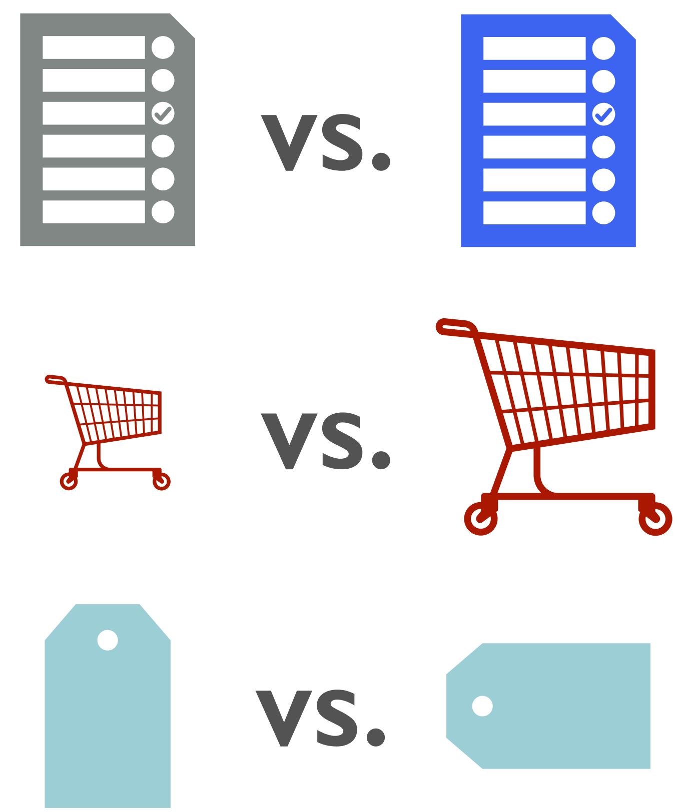
The aim of online FDR procedures

Decision rule:

$P_1 \leq \alpha_1$?

$P_2 \leq \alpha_2$?

Time



Color

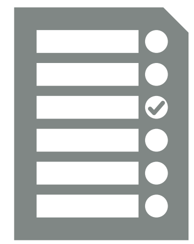
Size

Orientation

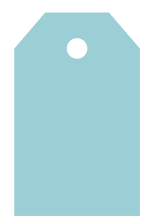
The aim of online FDR procedures

Decision rule:

Time

$P_1 \leq \alpha_1$?  **vs.**  Color

$P_2 \leq \alpha_2$?  **vs.**  Size

$P_3 \leq \alpha_3$?  **vs.**  Orientation

The aim of online FDR procedures

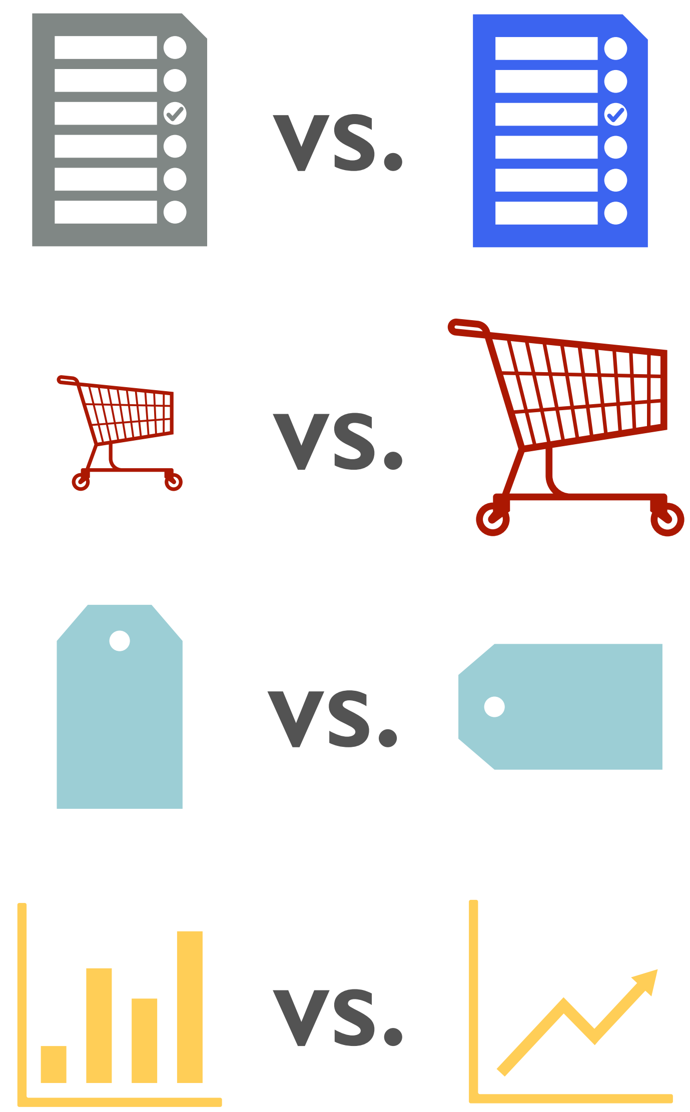
Decision rule:

Time

$P_1 \leq \alpha_1$?

$P_2 \leq \alpha_2$?

$P_3 \leq \alpha_3$?



Color

Size

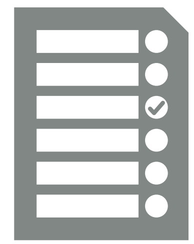
Orientation

Style

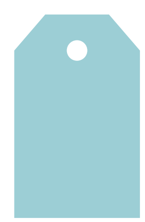
The aim of online FDR procedures

Decision rule:

Time

$P_1 \leq \alpha_1$?  **vs.**  Color

$P_2 \leq \alpha_2$?  **vs.**  Size

$P_3 \leq \alpha_3$?  **vs.**  Orientation

$P_4 \leq \alpha_4$? **vs.** Style

The aim of online FDR procedures

Decision rule:

Time

$P_1 \leq \alpha_1$?

$P_2 \leq \alpha_2$?

$P_3 \leq \alpha_3$?

$P_4 \leq \alpha_4$?

Color

Size

Orientation

Style

Logo

The aim of online FDR procedures

Decision rule:

Time

$P_1 \leq \alpha_1$?

$P_2 \leq \alpha_2$?

$P_3 \leq \alpha_3$?

$P_4 \leq \alpha_4$?

$P_5 \leq \alpha_5$?

Color

Size

Orientation

Style

Logo

The aim of online FDR procedures

Decision rule:

$P_1 \leq \alpha_1$?

$P_2 \leq \alpha_2$?

$P_3 \leq \alpha_3$?

$P_4 \leq \alpha_4$?

$P_5 \leq \alpha_5$?

Time

Color

Size

Orientation

Style

Logo

How do we set each error level to control FDR at any time?

One of the most famous offline FDR methods is the "Benjamini-Hochberg" (BH) method

Offline FDR methods do not control the FDR in online settings

Benjamini-Hochberg '95

The following method is **not** a valid online FDR algorithm:

At the end of experiment t, run BH on $P_1, \ldots, P_t$.

The following method is **not** a valid online FDR algorithm:

At the end of experiment t, run BH on $P_1, \ldots, P_t$.

The reason is that the decision about the first hypothesis depends on all future hypotheses. We cannot commit to a decision and stick to it.

The following method is **not** a valid online FDR algorithm:

At the end of experiment t, run BH on $P_1, \ldots, P_t$.

The reason is that the decision about the first hypothesis depends on all future hypotheses. We cannot commit to a decision and stick to it.

We need the error level α_t for experiment t to be specified when it starts, and we need to make a final decision when experiment t ends.

This multiple testing issue
is not particular to p-values.
It also exists when selectively
reporting treatment effects
with confidence intervals.

Benjamini, Yekutieli '05
Weinstein, Ramdas '19

Multiplicity in reported CIs

One rarely cares about all CIs or follows-up on them, one usually reports only the most "promising" CIs.

Multiplicity in reported CIs

One rarely cares about all CIs or follows-up on them, one usually reports only the most "promising" CIs.

False coverage proportion

$$\text{FCP} = \frac{\#\ \text{incorrectly reported CIs}}{\#\ \text{reported CIs}}$$

Multiplicity in reported CIs

One rarely cares about all CIs or follows-up on them, one usually reports only the most "promising" CIs.

False coverage proportion

$$\mathrm{FCP} = \frac{\#\text{ incorrectly reported CIs}}{\#\text{ reported CIs}}$$

False coverage rate

$$\mathrm{FCR} = \mathbb{E}[\mathrm{FCP}]$$

Benjamini-Yekutieli '06
Weinstein-Yekutieli '14
Fithian et al. '14

Controlling FCR is nontrivial

Constructing marginal 95% CIs for all parameters fails to control FCR at 0.05.

Controlling FCR is nontrivial

Constructing marginal 95% CIs for all parameters fails to control FCR at 0.05.

Suppose treatment effect $\theta_j \in \{\pm 0.1\}$ for all j, and experimental observations are normalized to $X_j \sim N(\theta_j, 1)$.

Controlling FCR is nontrivial

Constructing marginal 95% CIs for all parameters fails to control FCR at 0.05.

Suppose treatment effect $\theta_j \in \{\pm 0.1\}$ for all j, and experimental observations are normalized to $X_j \sim N(\theta_j, 1)$.

Suppose we only care about drugs with large effects. So we only pursue phase II of the trial if $X_j > 3$.

Controlling FCR is nontrivial

Constructing marginal 95% CIs for all parameters fails to control FCR at 0.05.

Suppose treatment effect $\theta_j \in \{\pm 0.1\}$ for all j, and experimental observations are normalized to $X_j \sim N(\theta_j, 1)$.

Suppose we only care about drugs with large effects. So we only pursue phase II of the trial if $X_j > 3$.

For these drugs, the standard marginal 95% CI does not cover θ_j. So FCR=1.

Can we control FCR *online*?

When experiment j starts, we must assign a target confidence level α_j.
When experiment j ends, we must decide if we wish to report θ_j.
This must be done such that the FCR is controlled at *any* time.

Can we control FCR *online*?

When experiment j starts, we must assign a target confidence level α_j.
When experiment j ends, we must decide if we wish to report θ_j.
This must be done such that the FCR is controlled at *any* time.

Weinstein, Ramdas '19

A simple solution for both online FDR control, and online FCR control

Online FCR control: the main idea

Let $S_i \in \{0,1\}$ denote the selection decision made after experiment i.

$$\text{Maintain } \widehat{\text{FCP}}(T) := \frac{\sum_{i=1}^{T} \alpha_i}{1 \vee \sum_{i=1}^{T} S_i} \leq \alpha.$$

Online FCR control: the main idea

Let $S_i \in \{0, 1\}$ denote the selection decision made after experiment i.

$$\text{Maintain } \widehat{\text{FCP}}(T) := \frac{\sum_{i=1}^{T} \alpha_i}{1 \vee \sum_{i=1}^{T} S_i} \leq \alpha.$$

Weinstein & Ramdas '19

Online FCR control: the main idea

Let $S_i \in \{0, 1\}$ denote the selection decision made after experiment i.

$$\text{Maintain } \widehat{\text{FCP}}(T) := \frac{\sum_{i=1}^{T} \alpha_i}{1 \vee \sum_{i=1}^{T} S_i} \leq \alpha.$$

For testing, let $R_i \in \{0,1\}$ represent a rejection.

$$\text{Then maintain } \widehat{\text{FDP}}(T) := \frac{\sum_{i=1}^{T} \alpha_i}{1 \vee \sum_{i=1}^{T} R_i} \leq \alpha\,.$$

Weinstein & Ramdas '19

Online FCR control: the main idea

Let $S_i \in \{0,1\}$ denote the selection decision made after experiment i.

Maintain $$\widehat{\text{FCP}}(T) := \frac{\sum_{i=1}^{T} \alpha_i}{1 \vee \sum_{i=1}^{T} S_i} \leq \alpha.$$

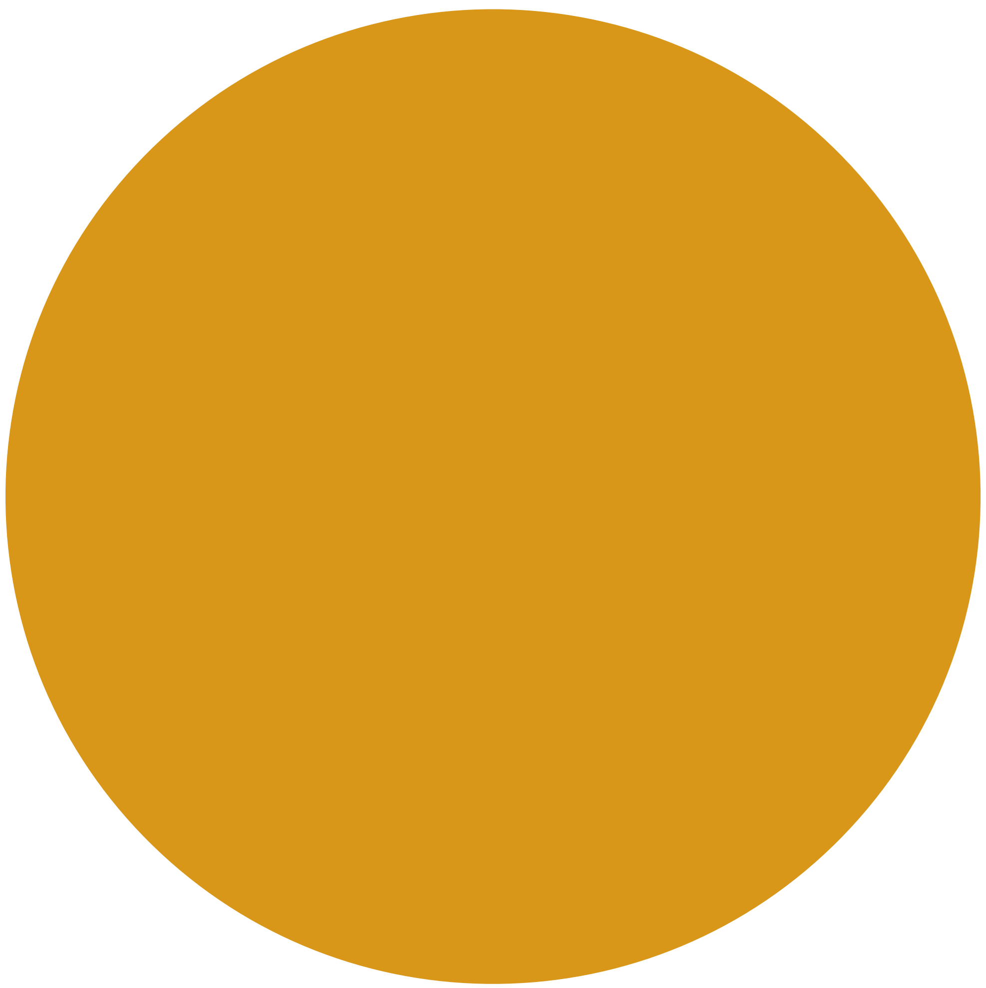
For testing, let $R_i \in \{0,1\}$ represent a rejection.

Then maintain $$\widehat{\text{FDP}}(T) := \frac{\sum_{i=1}^{T} \alpha_i}{1 \vee \sum_{i=1}^{T} R_i} \leq \alpha.$$

This provably controls FCR/FDR at level α.

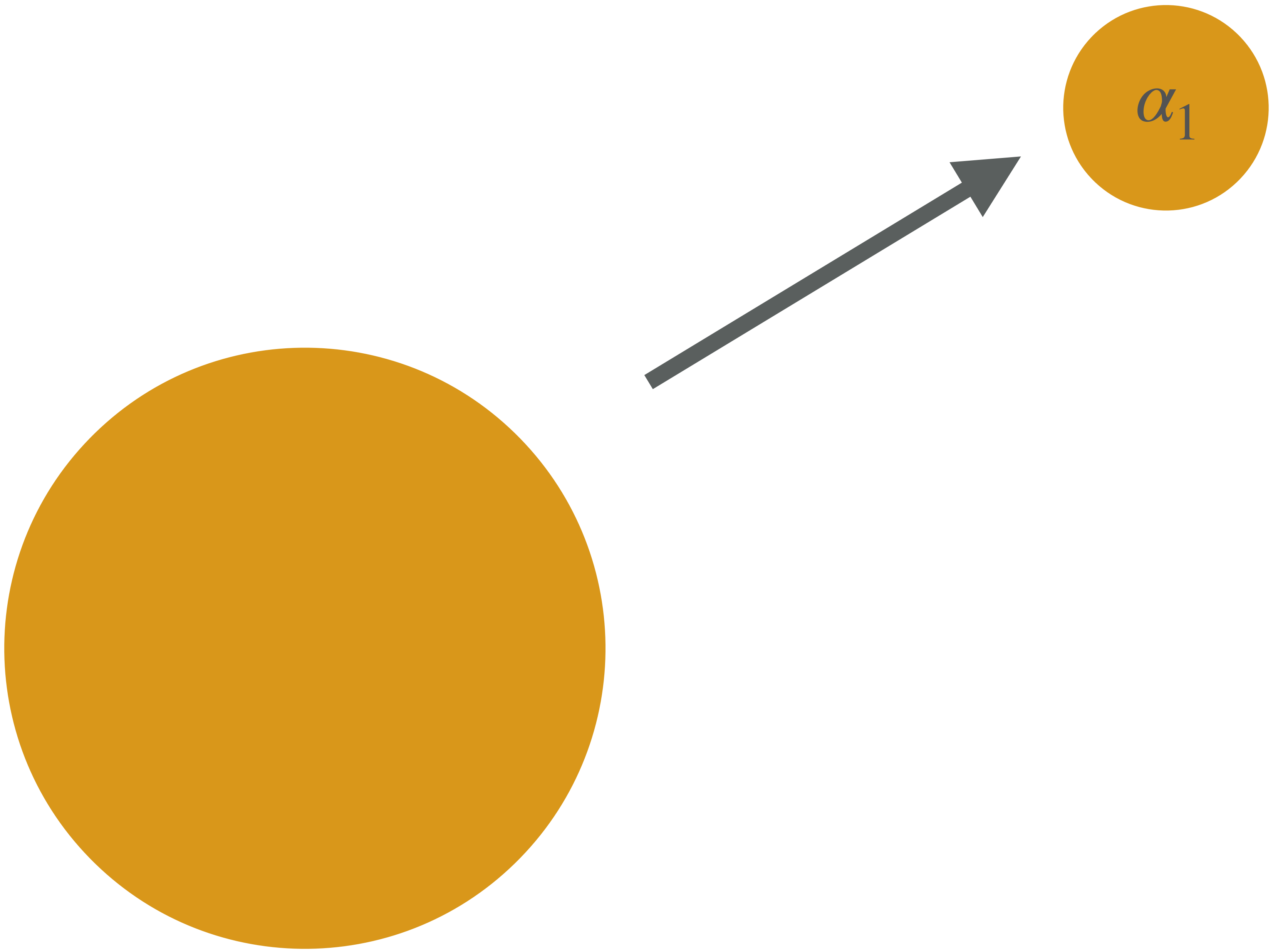
Weinstein & Ramdas '19

Online FCR control : high-level picture



Remaining error budget
or "alpha-wealth"

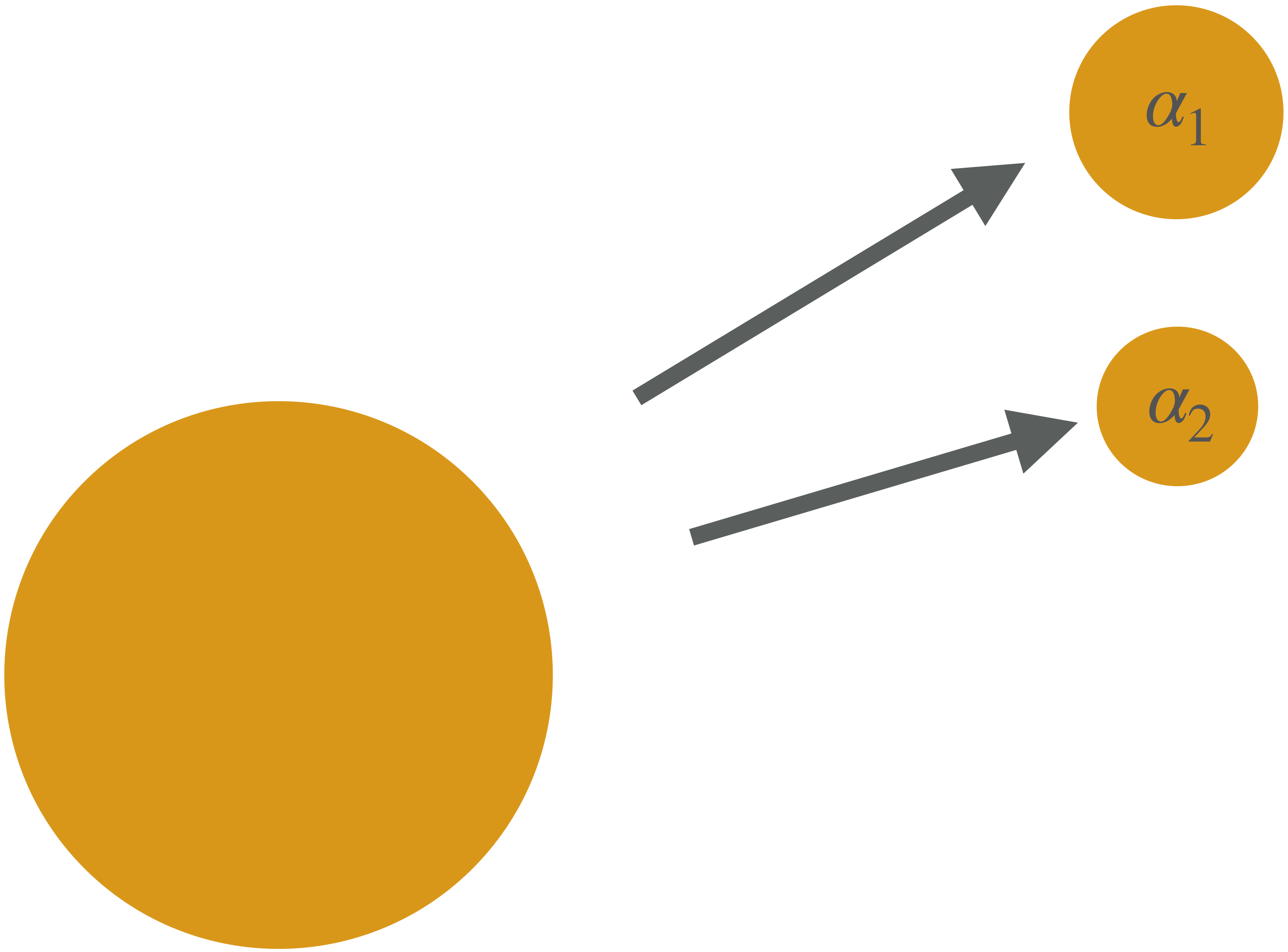
Online FCR control : high-level picture



Error budget
for first expt.

Remaining error budget
or "alpha-wealth"

Online FCR control : high-level picture

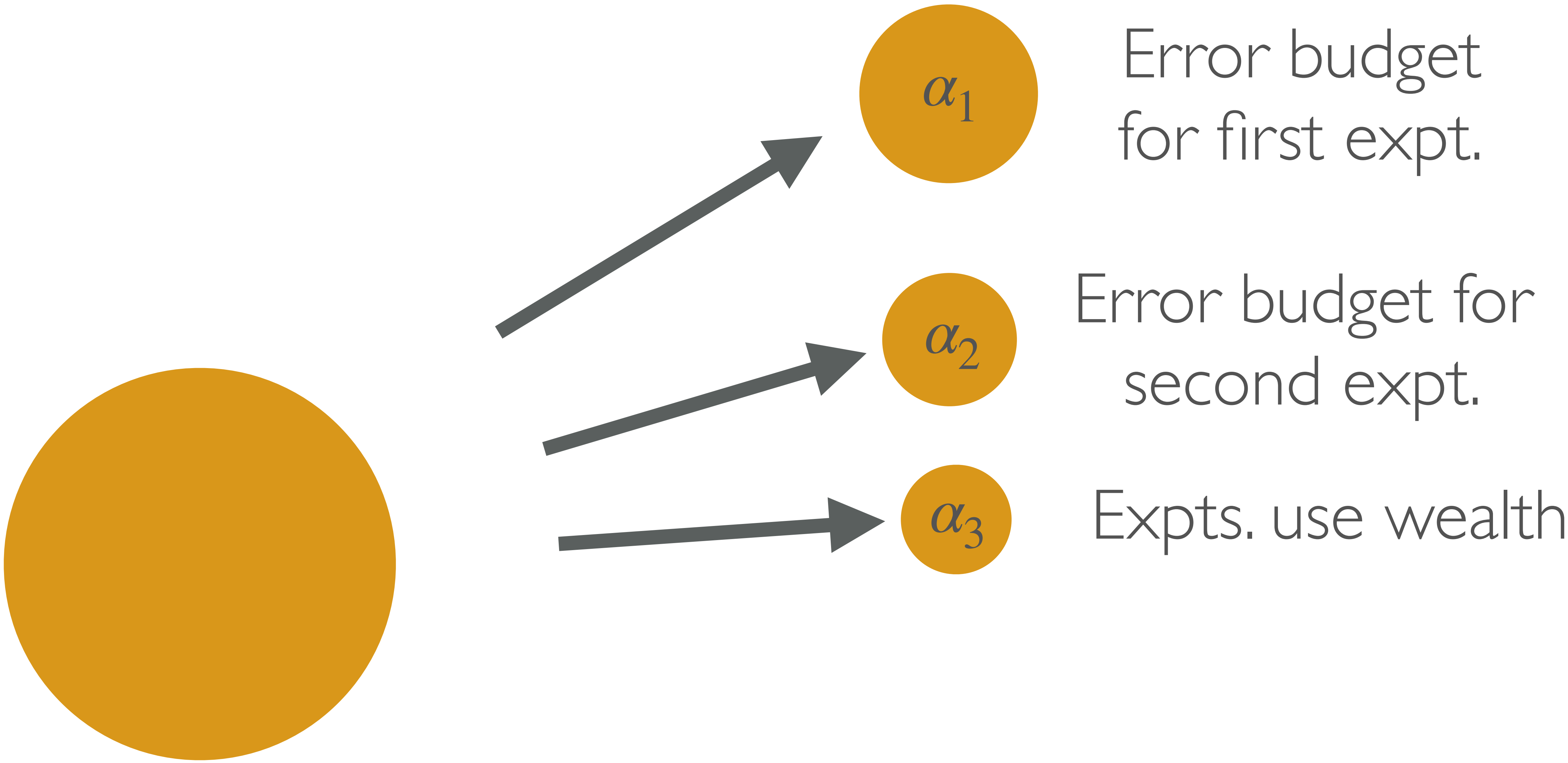


Error budget for first expt.

Error budget for second expt.

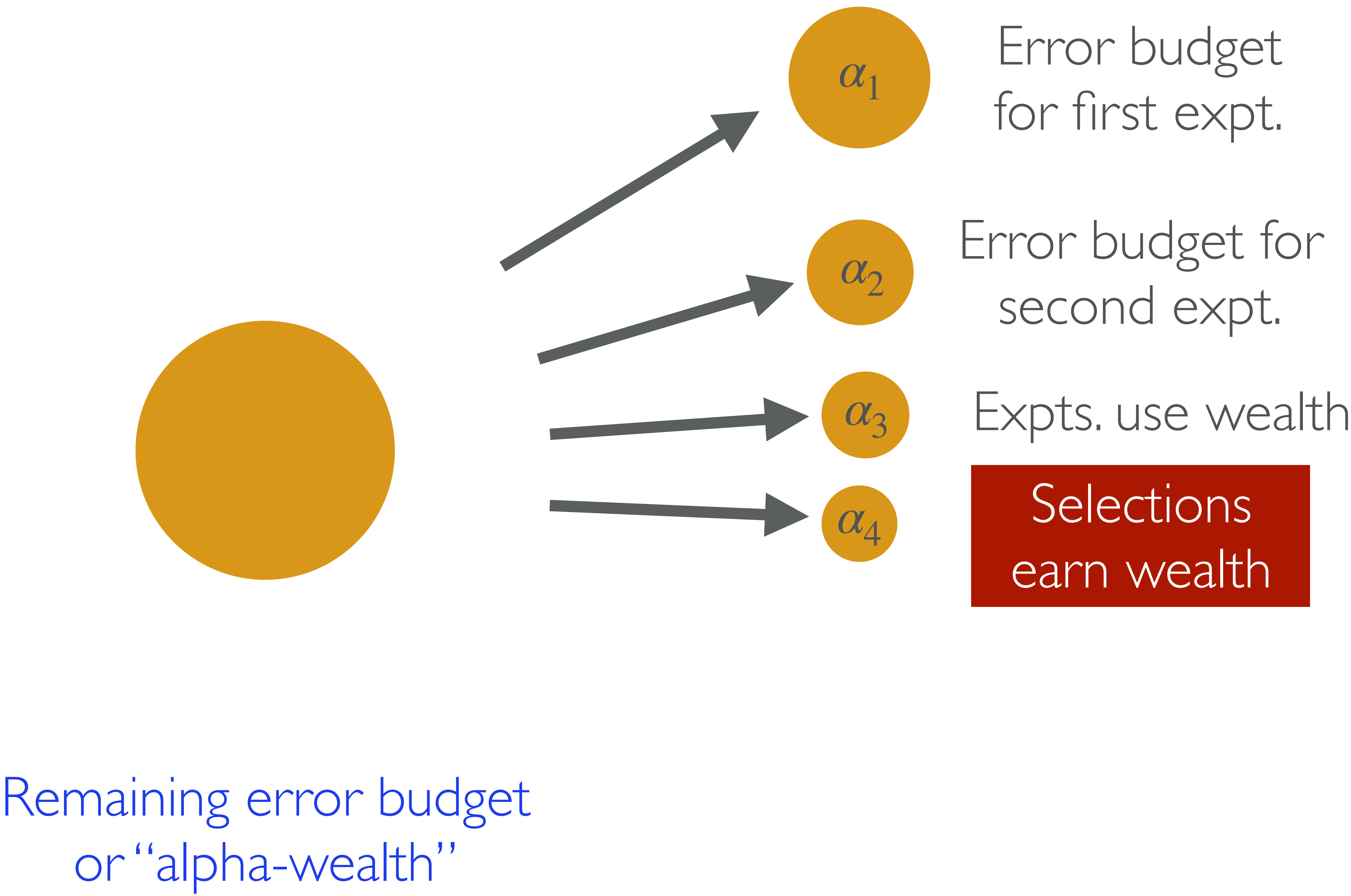
Remaining error budget or "alpha-wealth"

Online FCR control : high-level picture

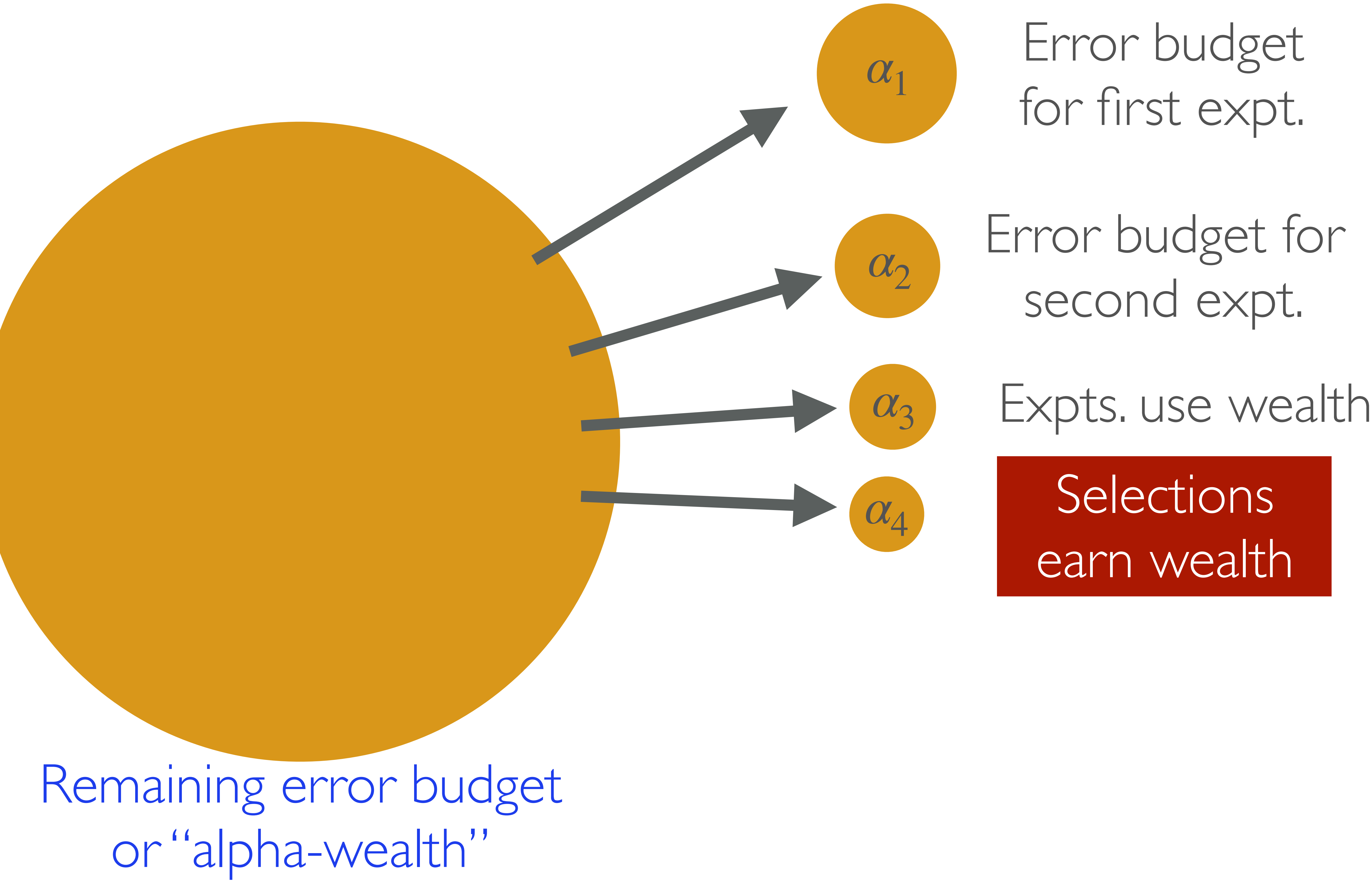


Remaining error budget
or "alpha-wealth"

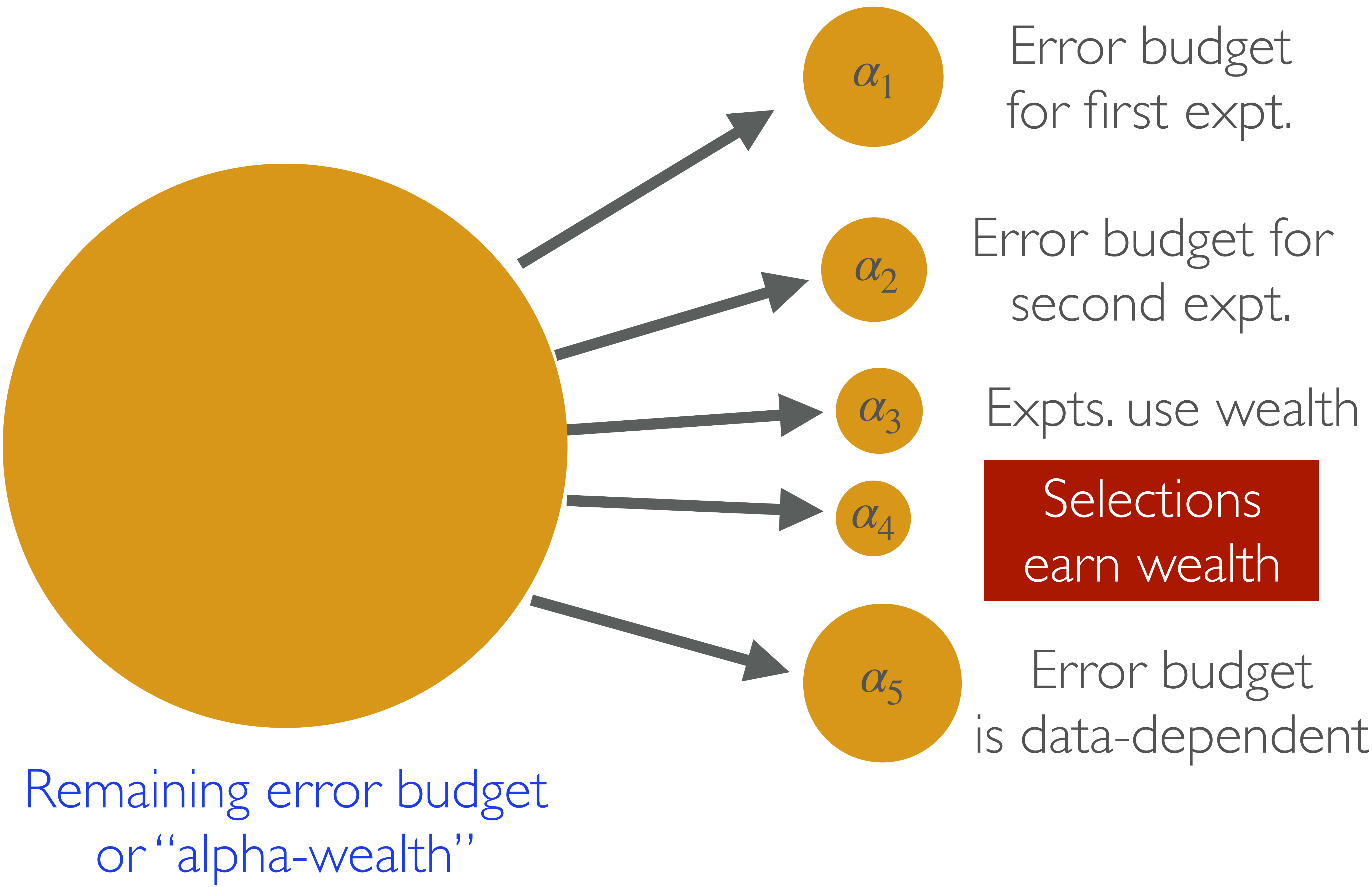
Online FCR control : high-level picture



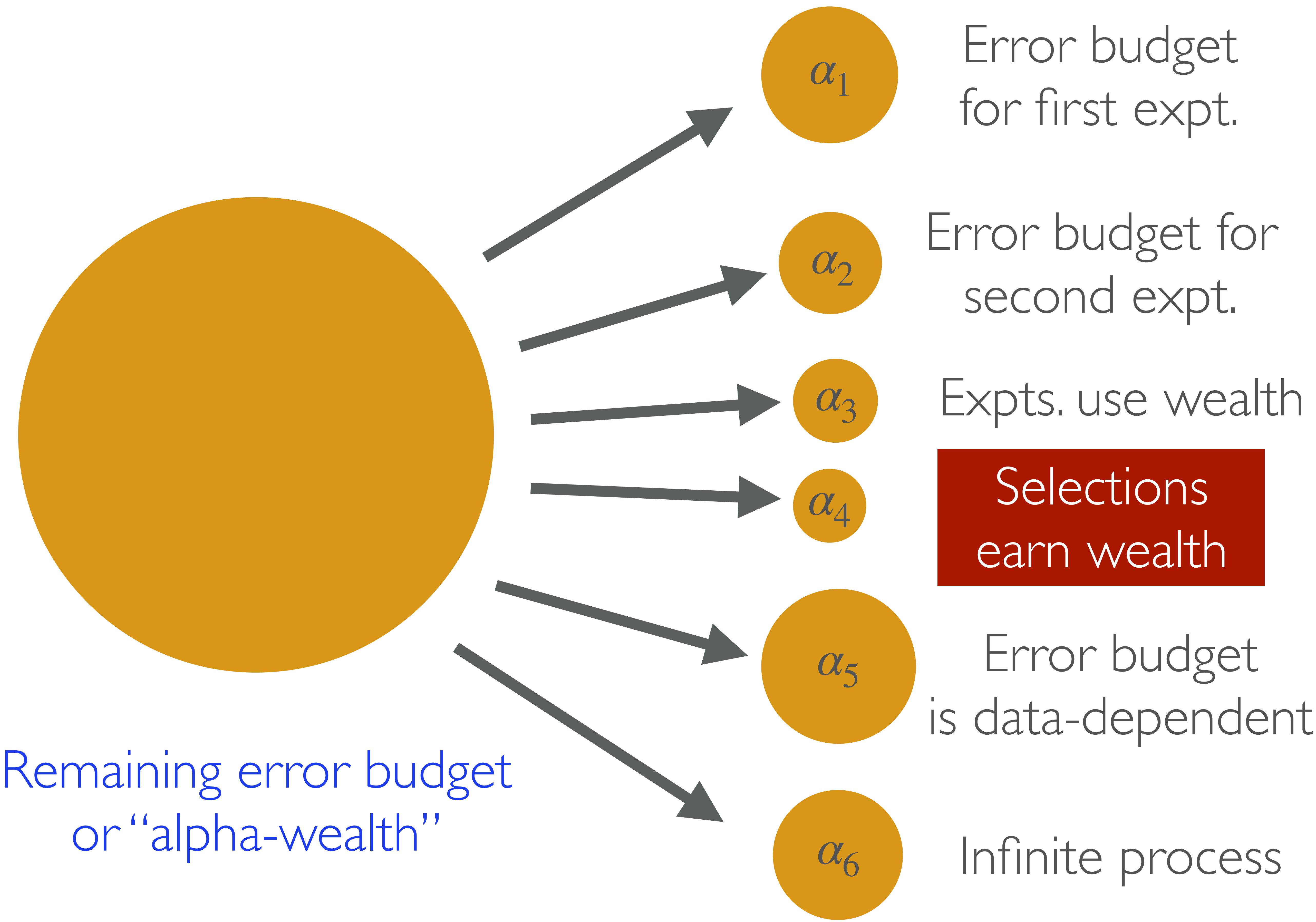
Online FCR control : high-level picture



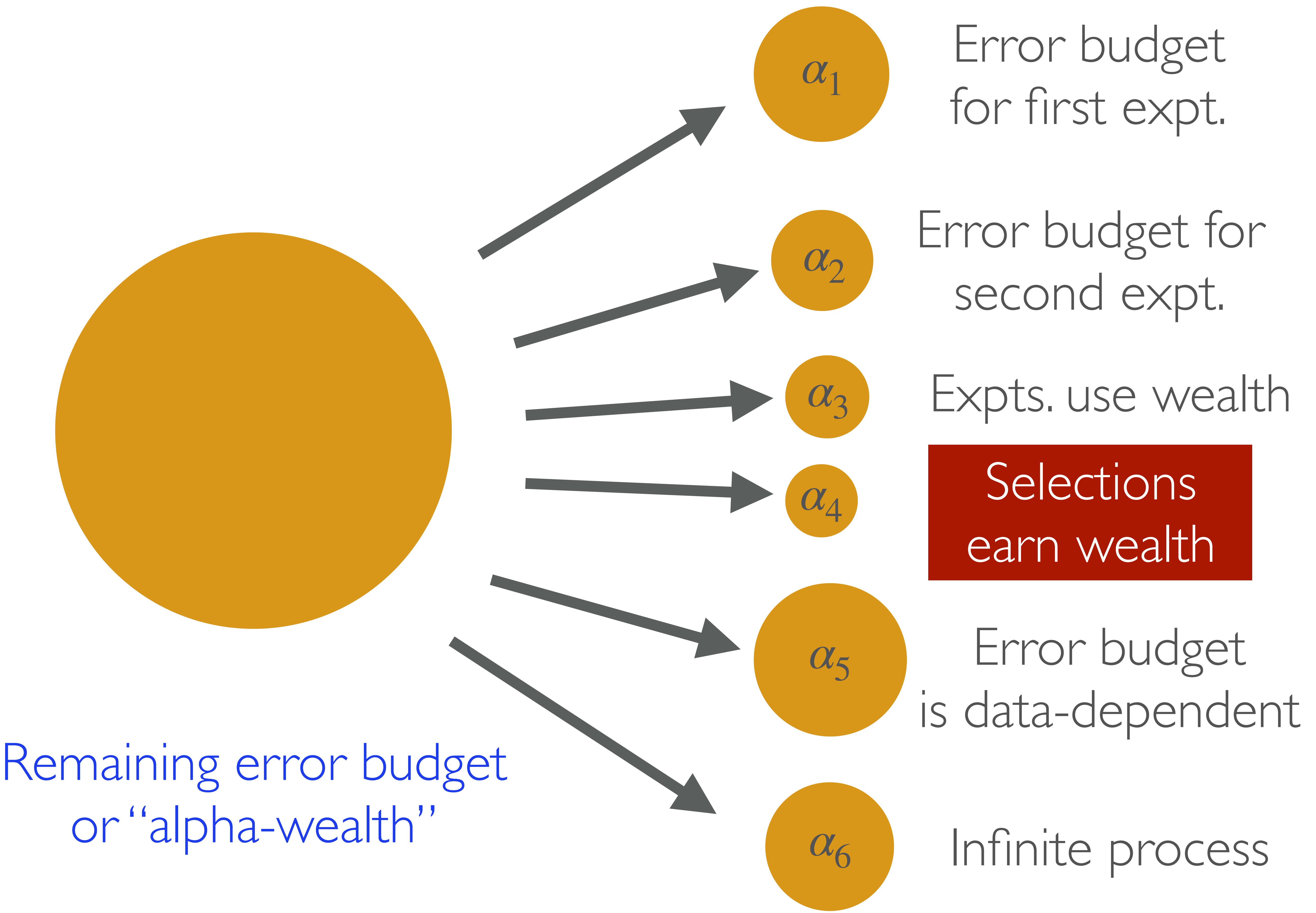
Online FCR control : high-level picture



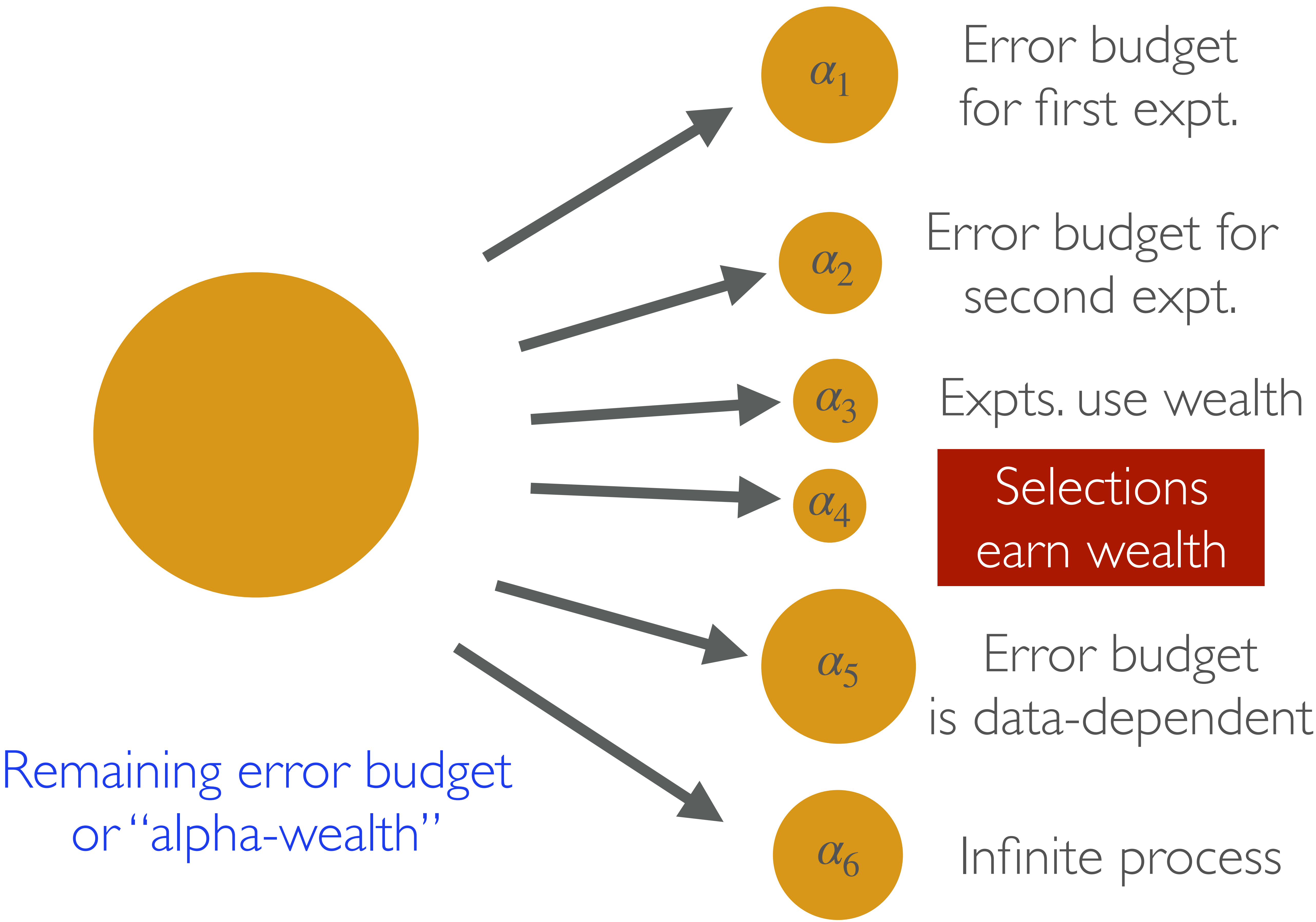
Online FCR control : high-level picture



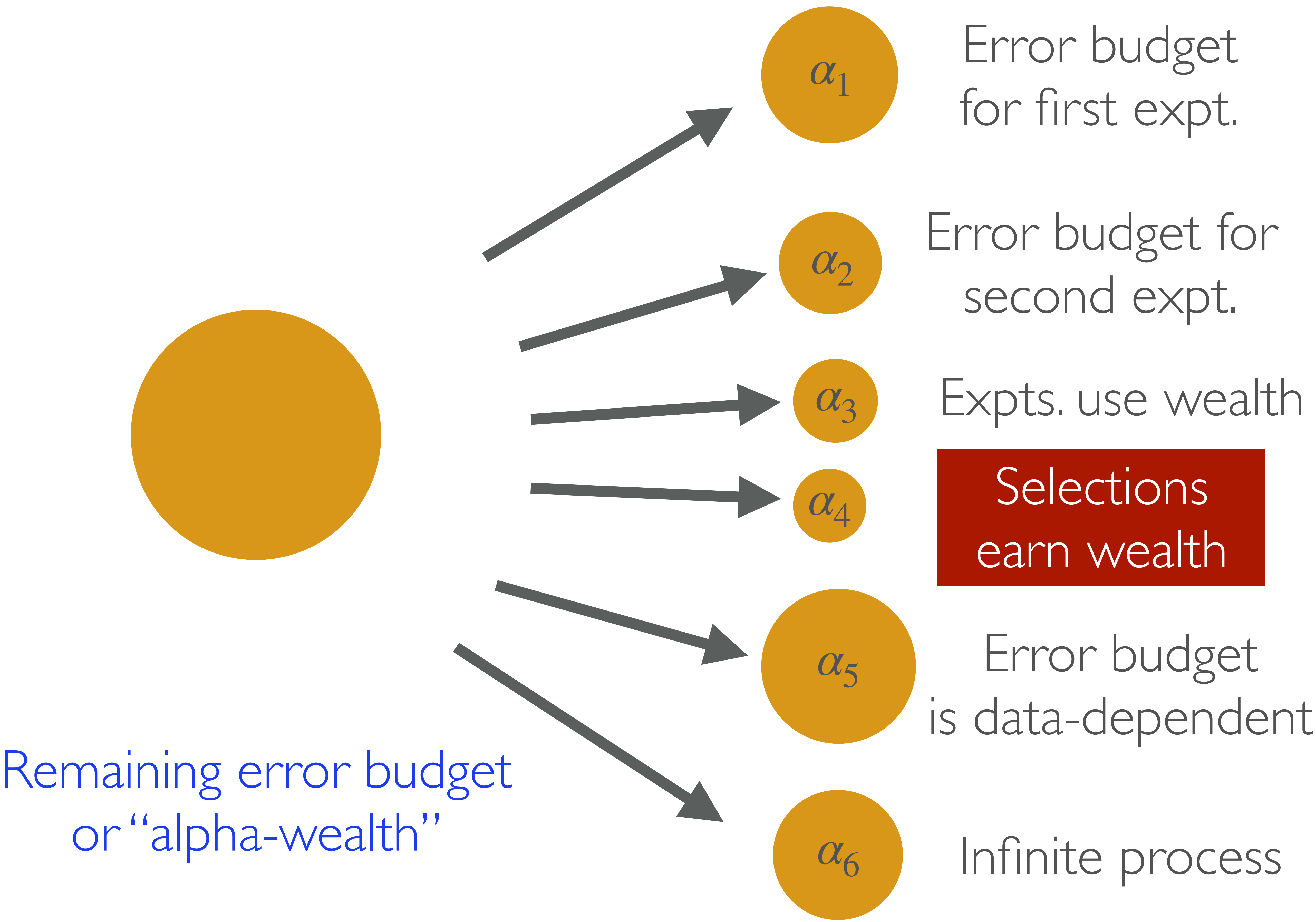
Online FCR control : high-level picture



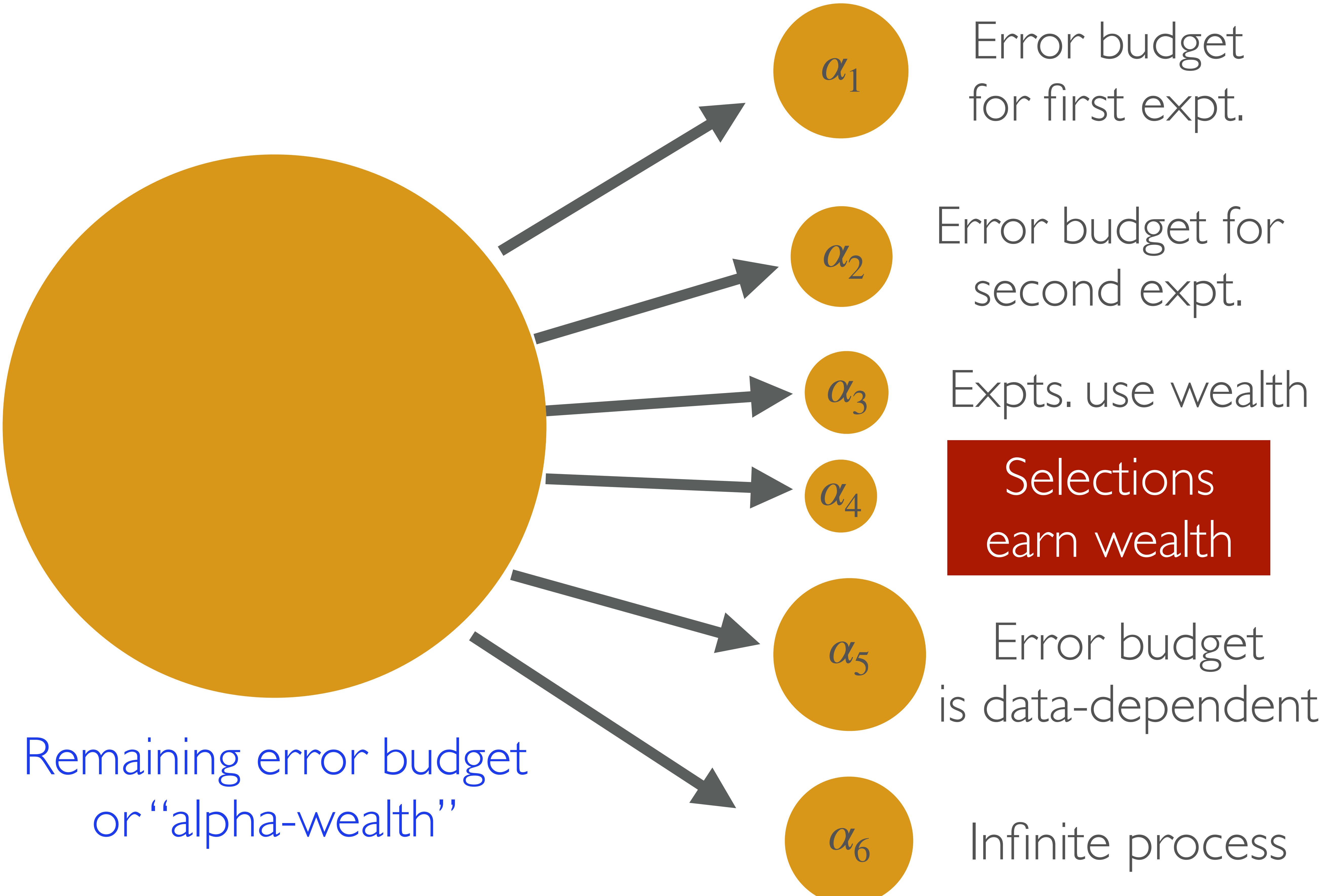
Online FCR control : high-level picture



Online FCR control : high-level picture



Online FCR control : high-level picture



Summary of this section

Summary of this section

- $\forall t,\ \text{error}_t \le \alpha \quad \text{does not imply} \quad \forall t,\ \text{FDR}(t) \le \alpha,$
  even if hypotheses, data, p-values are independent.

Summary of this section

- $\forall t,\ \text{error}_t \leq \alpha \quad \text{does not imply} \quad \forall t,\ \text{FDR}(t) \leq \alpha,$
  even if hypotheses, data, p-values are independent.

- Can track a **running estimate** of the FDP (or FCP):
  a simple update rule to keep this estimate bounded
  also results in the FDR (or FCR) being controlled.

Handling local dependence

Most online FDR algorithms assume independent p-values (but hypotheses can be dependent).

Handling local dependence

Most online FDR algorithms assume independent p-values (but hypotheses can be dependent).

However, assuming arbitrary dependence between *all* p-values is also extremely pessimistic and unrealistic.

Handling local dependence

Most online FDR algorithms assume independent p-values (but hypotheses can be dependent).

However, assuming arbitrary dependence between *all* p-values is also extremely pessimistic and unrealistic.

A middle ground is a flexible notion of local dependence:
P_t arbitrarily depends on the previous L_t p-values,
where L_t is a user-chosen lag-parameter.

Handling local dependence

Most online FDR algorithms assume independent p-values (but hypotheses can be dependent).

However, assuming arbitrary dependence between *all* p-values is also extremely pessimistic and unrealistic.

A middle ground is a flexible notion of local dependence:
P_t arbitrarily depends on the previous L_t p-values,
where L_t is a user-chosen lag-parameter.

The online FDR and FCR algorithms can be easily modified to handle local dependence.

Zrnic, Ramdas, Jordan '18

Solutions for these issues

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Outer sequential process:

Part II

"false coverage rate" for estimation

(correspondingly, "false discovery rate" for testing)

Solutions for these issues

Inner sequential process:

Part I

"confidence sequence" for estimation

also called "anytime confidence intervals"

(correspondingly, "always valid p-values" for testing)

Outer sequential process:

Part II

"false coverage rate" for estimation

(correspondingly, "false discovery rate" for testing)

Modular solutions: fit well together

Many extensions to each piece

Part III

Putting the modular pieces together:
the doubly-sequential process

[Next 10 mins]

Part III

Putting the modular pieces together: the doubly-sequential process

[Next 10 mins]

Combining inner and outer solutions (FCR):

Combining inner and outer solutions (FCR):

(a) Online FCR method assigns α_i when expt. starts

Combining inner and outer solutions (FCR):

(a) Online FCR method assigns α_i when expt. starts

(b) We keep track of $(1 - \alpha_i)$ confidence sequence

Combining inner and outer solutions (FCR):

(a) Online FCR method assigns α_i when expt. starts

(b) We keep track of $(1 - \alpha_i)$ confidence sequence

(c) Adaptively decide to stop, to report final CI or not

Combining inner and outer solutions (FCR):

(a) Online FCR method assigns α_i when expt. starts

(b) We keep track of $(1 - \alpha_i)$ confidence sequence

(c) Adaptively decide to stop, to report final CI or not

(d) Guarantee FCR$(T) \leq \alpha$ at any time positive time T

Combining inner and outer solutions (FCR):

(a) Online FCR method assigns α_i when expt. starts

(b) We keep track of $(1-\alpha_i)$ confidence sequence

(c) Adaptively decide to stop, to report final CI or not

(d) Guarantee FCR$(T) \leq \alpha$ at any time positive time T

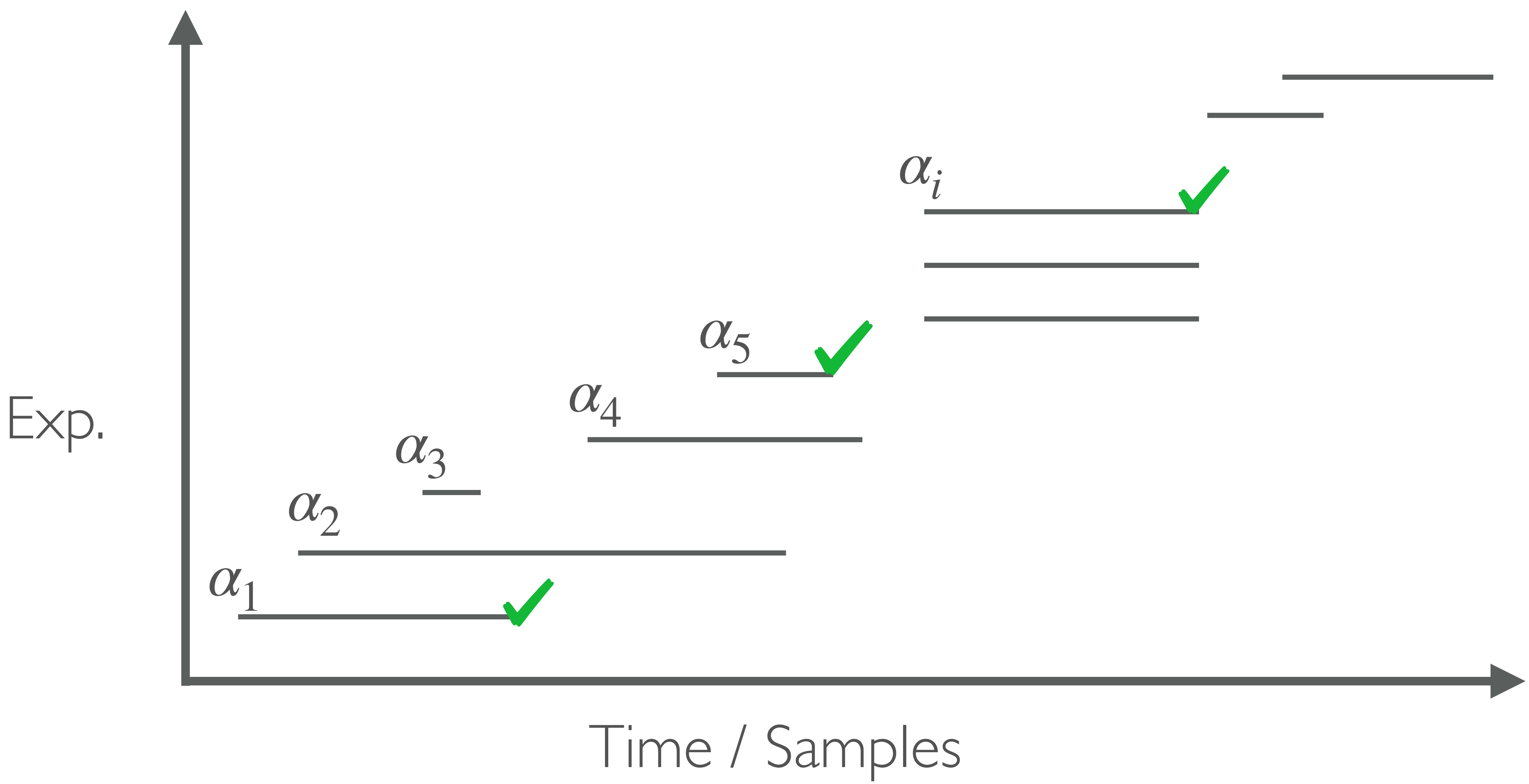
Combining inner and outer solutions (FCR):

(a) Online FCR method assigns α_i when expt. starts

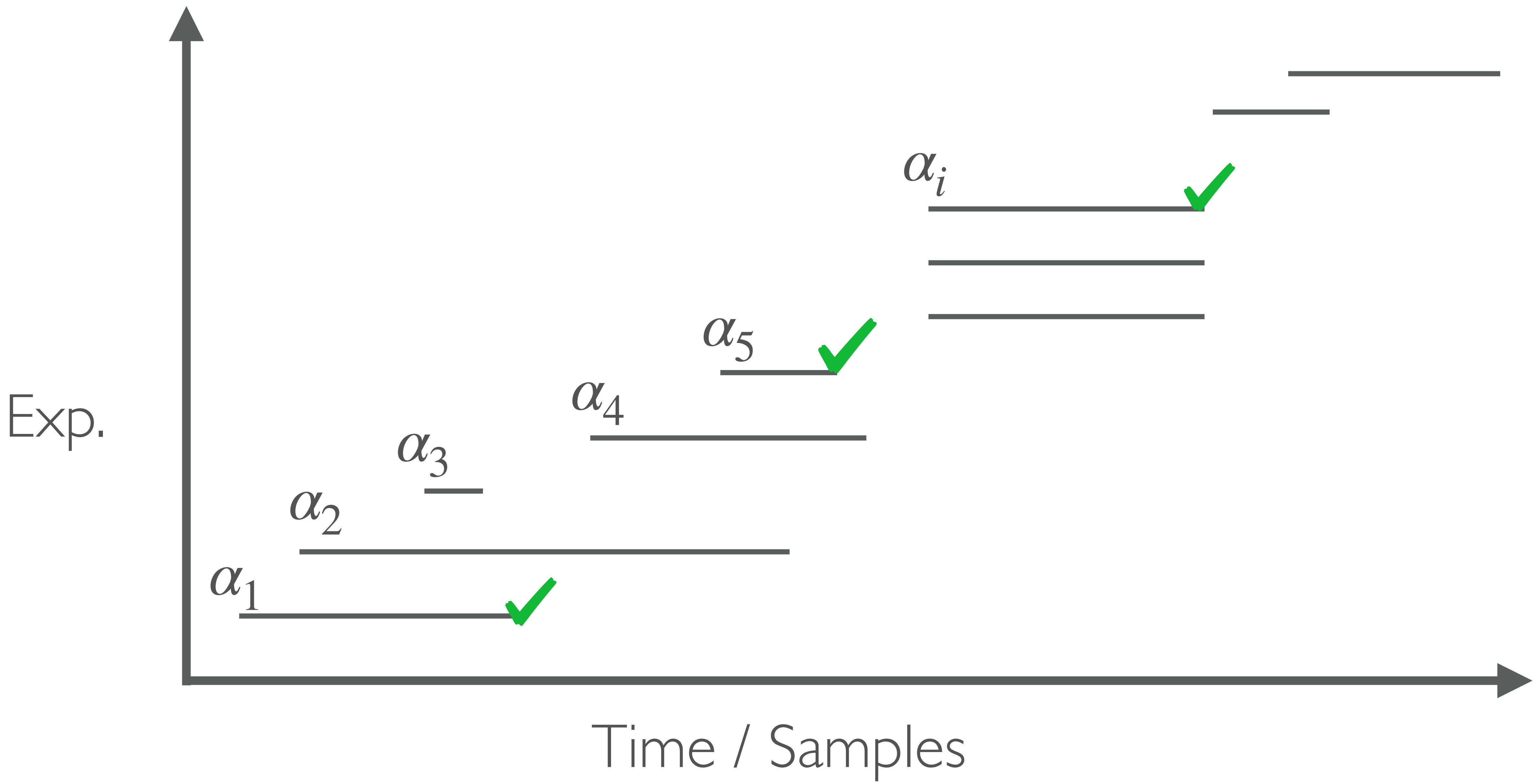
(b) We keep track of $(1-\alpha_i)$ confidence sequence

(c) Adaptively decide to stop, to report final CI or not

(d) Guarantee FCR$(T) \leq \alpha$ at any time positive time T

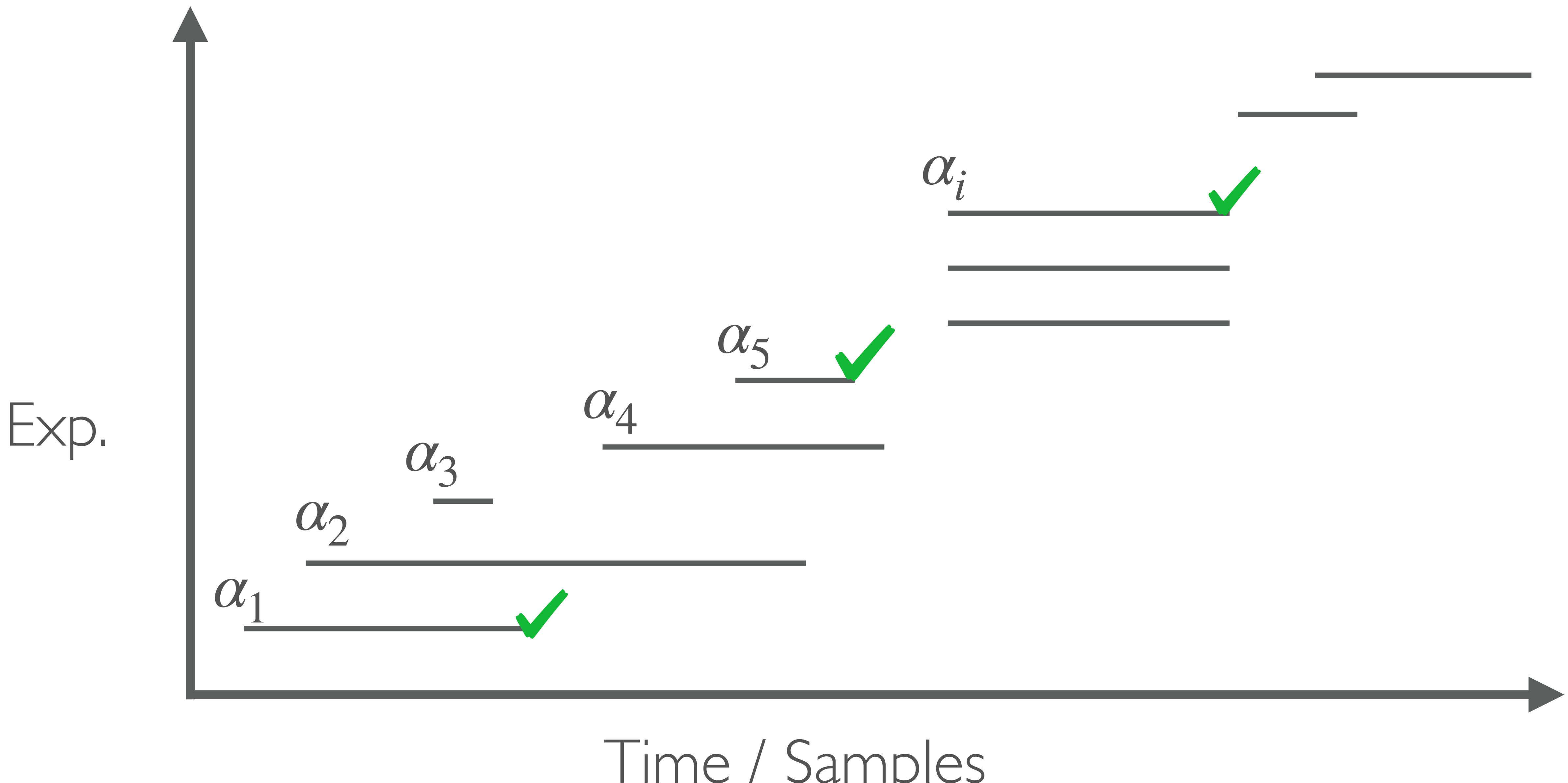


Combining inner and outer solutions (FDR):



Combining inner and outer solutions (FDR):

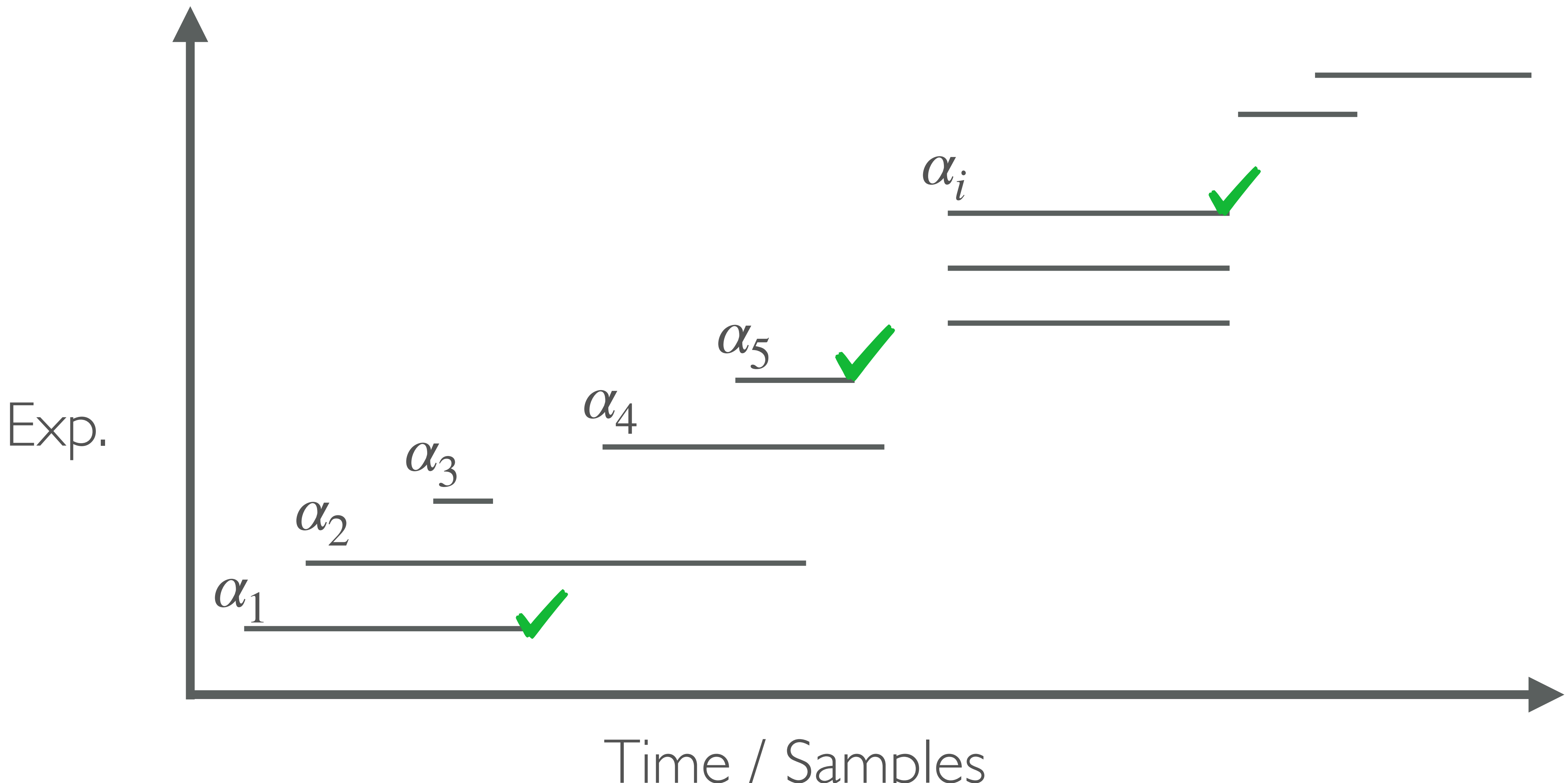
(a) Online FDR method assigns α_i when expt. starts



Combining inner and outer solutions (FDR):

(a) Online FDR method assigns α_i when expt. starts

(b) We keep track of anytime p-value $P_i^{(n)}$

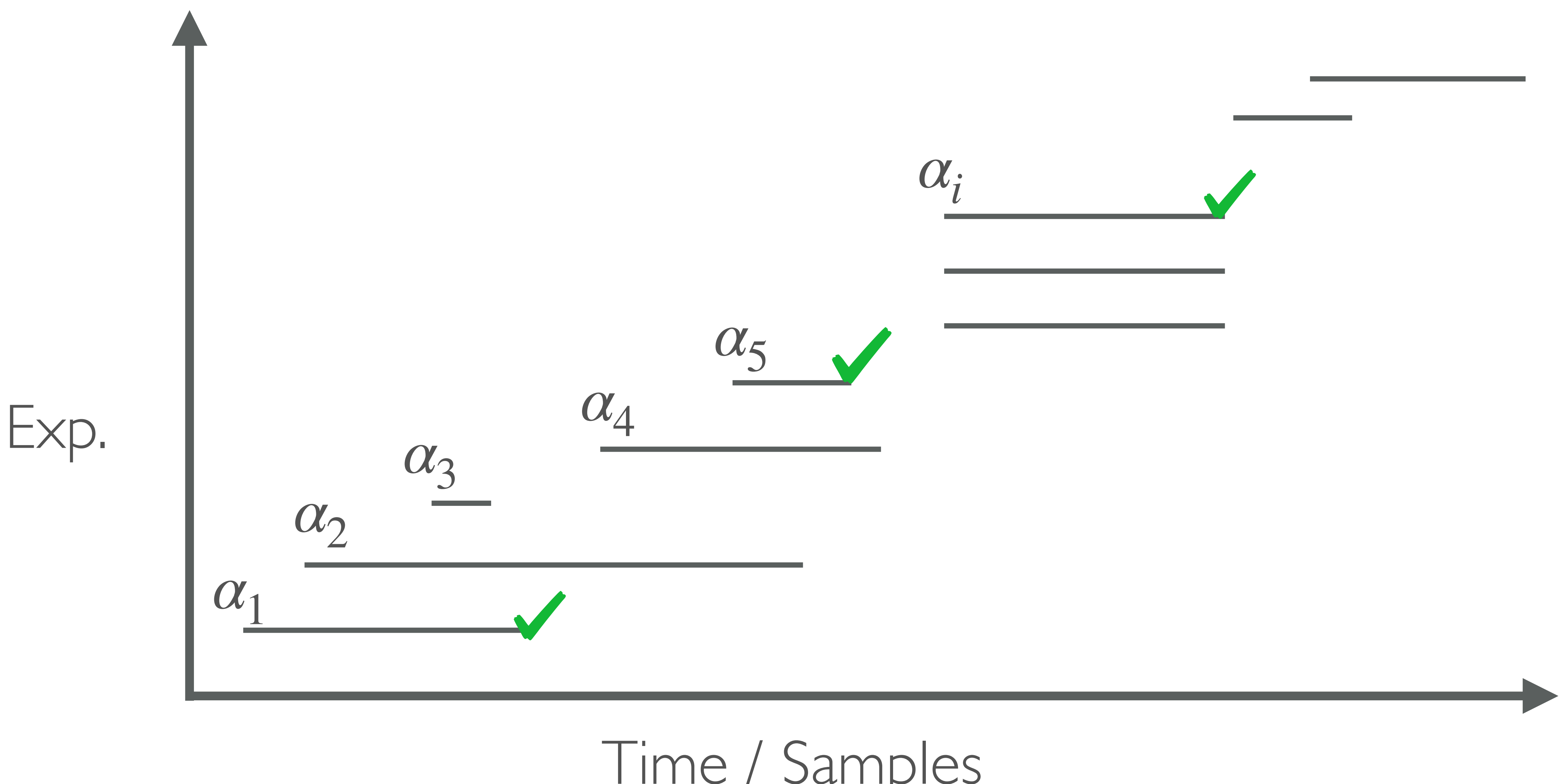


Combining inner and outer solutions (FDR):

(a) Online FDR method assigns α_i when expt. starts

(b) We keep track of anytime p-value $P_i^{(n)}$

(c) Adaptively stop at time τ, report discovery if $P_i^{(\tau)} \leq \alpha_i$



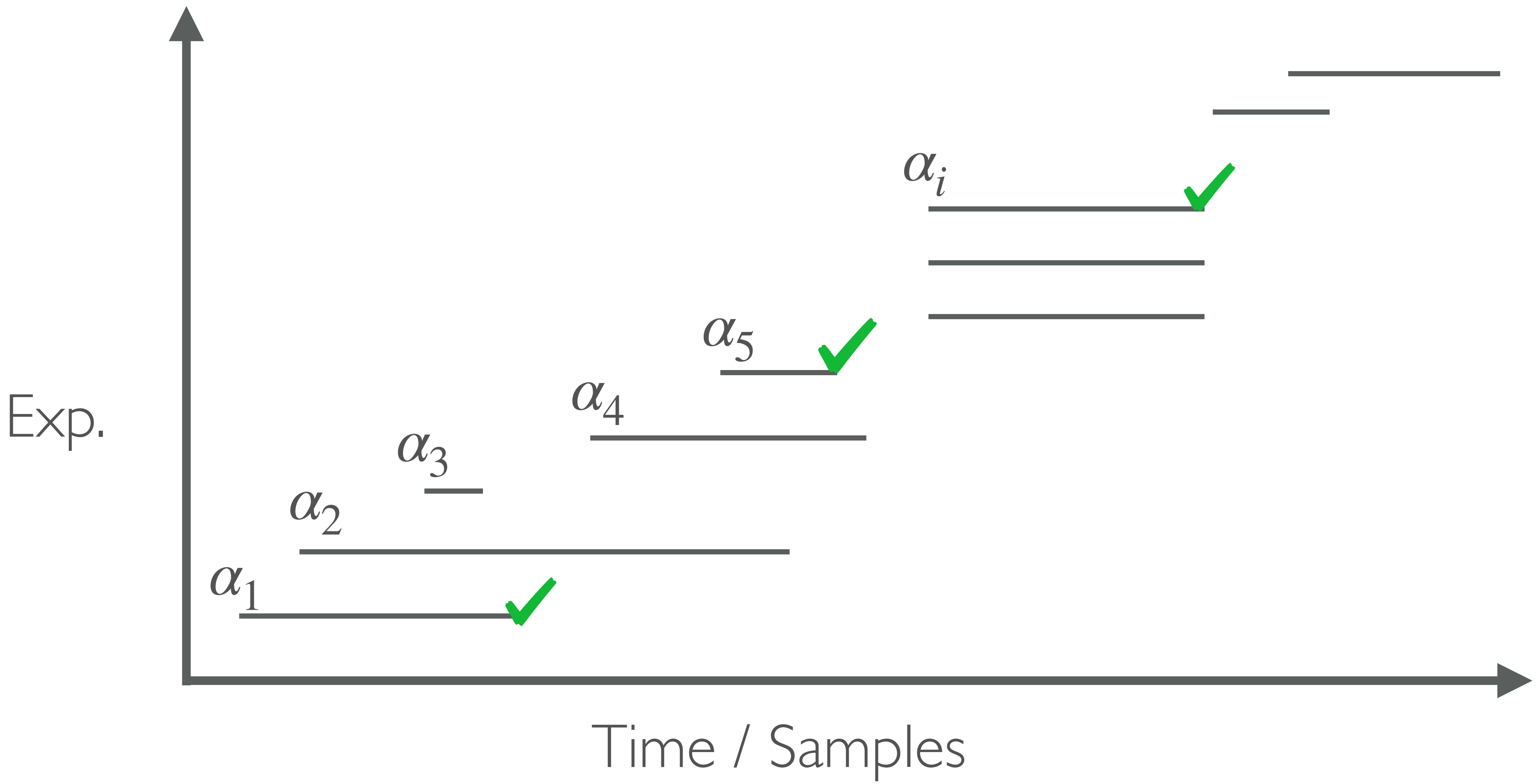
Combining inner and outer solutions (FDR):

(a) Online FDR method assigns α_i when expt. starts

(b) We keep track of anytime p-value $P_i^{(n)}$

(c) Adaptively stop at time τ, report discovery if $P_i^{(\tau)} \leq \alpha_i$

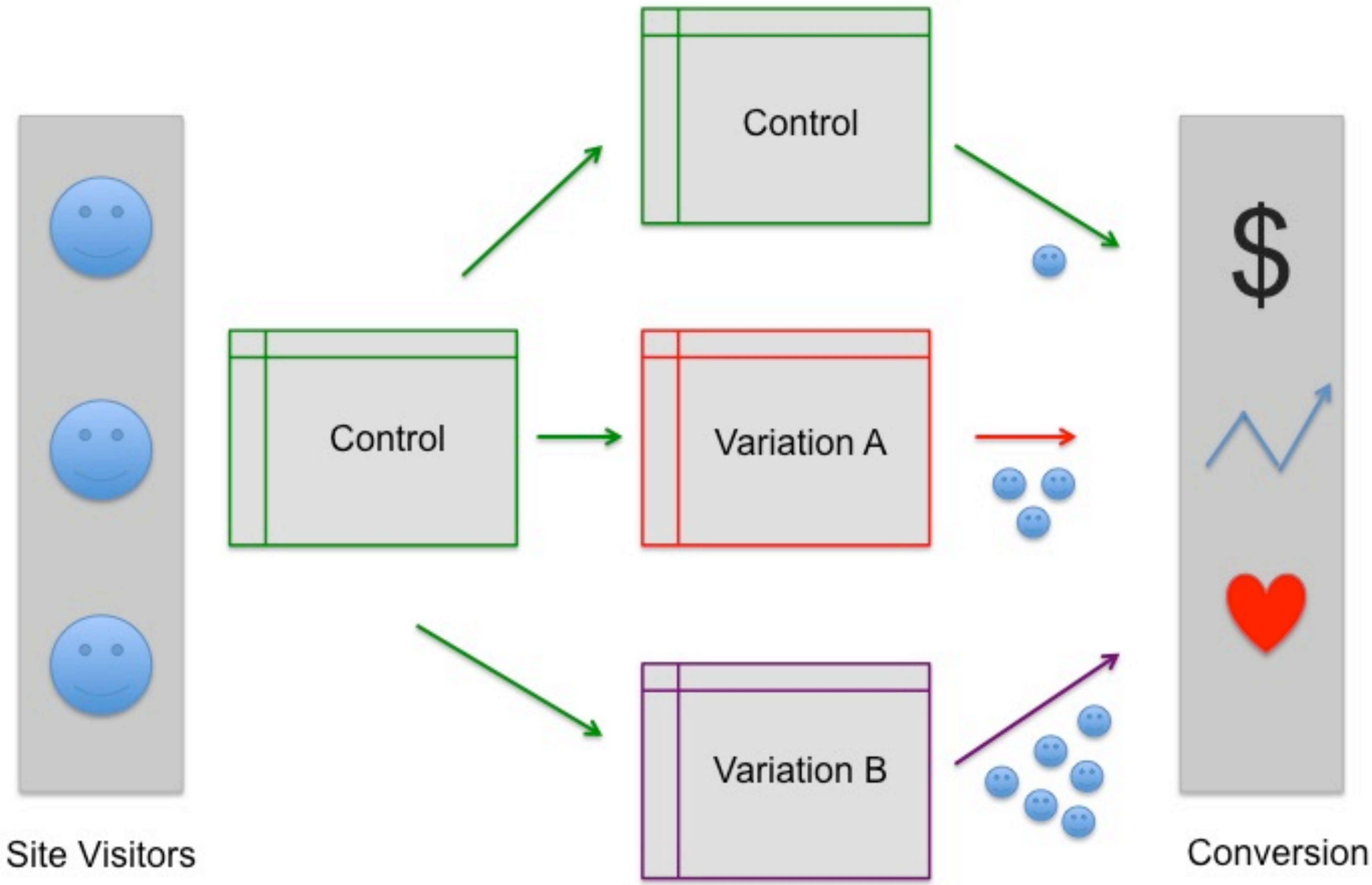
(d) Guarantee FDR$(T) \leq \alpha$ at any time positive time T



PART IV: Advanced topics (inner sequential process)

[Next 25 mins]

1. What if we are testing more than one alternative?



Much more traffic needed by an A/B/n test

1. Multi-armed bandits for hypothesis testing

What would **you** do?

1. Multi-armed bandits for hypothesis testing

What would **you** do?

Depends on the aim: minimize regret OR identify best arm?
We would like to test null hypothesis

$$H_0 : \mu_A \geq \max\{\mu_B, \mu_C\} \ .$$

1. Multi-armed bandits for hypothesis testing

What would **you** do?

Depends on the aim: minimize regret OR identify best arm?
We would like to test null hypothesis

$$H_0 : \mu_A \geq \max\{\mu_B, \mu_C\} .$$

Can design variant of UCB algorithms to define anytime p-value, with optimal sample complexity for high power.

1. Multi-armed bandits for hypothesis testing

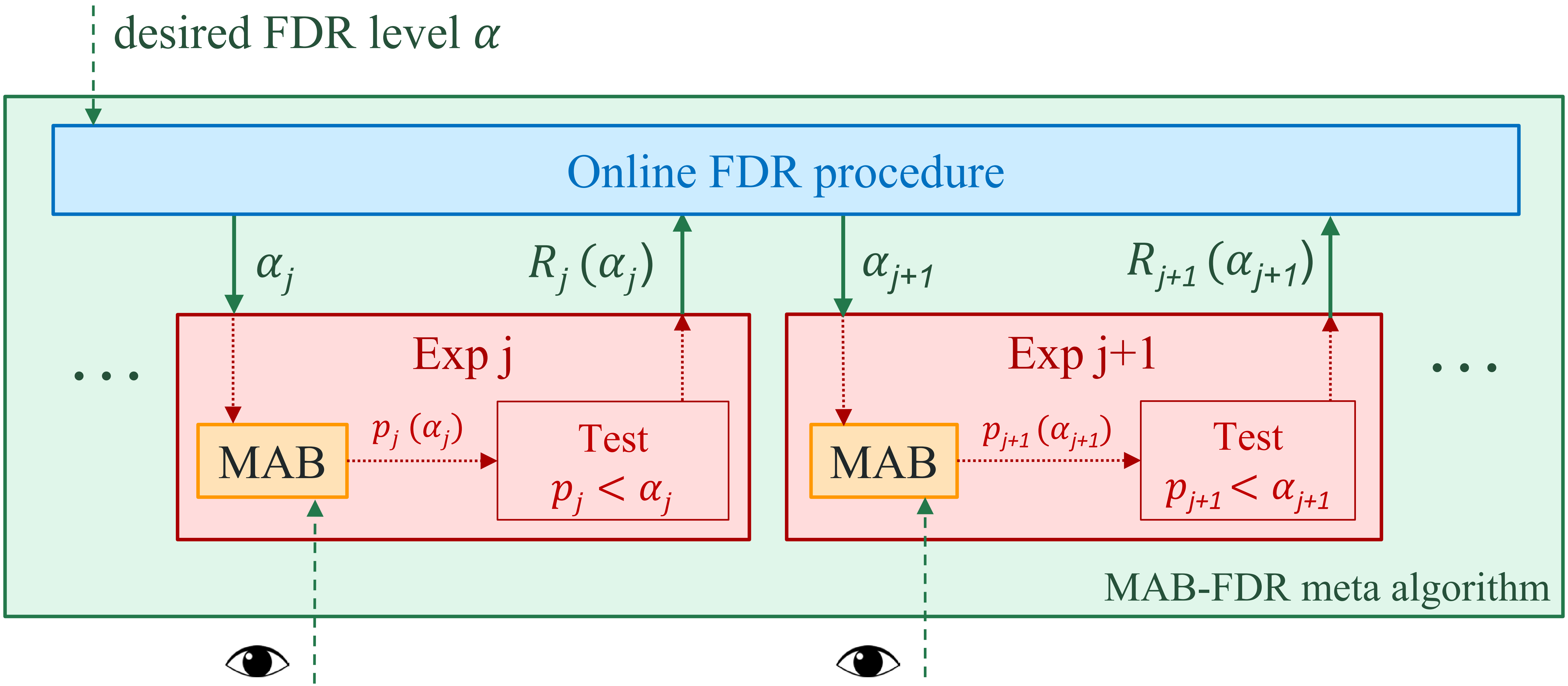
What would **you** do?

Depends on the aim: minimize regret OR identify best arm?
We would like to test null hypothesis

$$H_0 : \mu_A \geq \max\{\mu_B, \mu_C\} \,.$$

Can design variant of UCB algorithms to define anytime p-value, with optimal sample complexity for high power.

Yang, Ramdas, Jamieson, Wainwright '17



desired FDR level α
Online FDR procedure
α_j
$R_j(\alpha_j)$
α_{j+1}
$R_{j+1}(\alpha_{j+1})$
Exp j
Exp j+1
MAB
$p_j(\alpha_j)$
Test
$p_j < \alpha_j$
MAB
$p_{j+1}(\alpha_{j+1})$
Test
$p_{j+1} < \alpha_{j+1}$
MAB-FDR meta algorithm

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

Howard, Ramdas '19

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

Reasons to use quantiles include:

Howard, Ramdas '19

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

Reasons to use quantiles include:

- Quantiles always exist for any distribution,
  while means (moments) do not always exist (eg: Cauchy).

Howard, Ramdas '19

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

Reasons to use quantiles include:

- Quantiles always exist for any distribution, while means (moments) do not always exist (eg: Cauchy).
- Quantiles can be defined for any totally ordered space, eg: ratings A-F, where "distance between ratings" undefined.

Howard, Ramdas '19

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

Reasons to use quantiles include:

- Quantiles always exist for any distribution, while means (moments) do not always exist (eg: Cauchy).
- Quantiles can be defined for any totally ordered space, eg: ratings A-F, where "distance between ratings" undefined.
- Estimating quantiles can be done sequentially, without any tail assumptions, unlike estimating means.

Howard, Ramdas '19

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

Reasons to use quantiles include:

- Quantiles always exist for any distribution, while means (moments) do not always exist (eg: Cauchy).
- Quantiles can be defined for any totally ordered space, eg: ratings A-F, where "distance between ratings" undefined.
- Estimating quantiles can be done sequentially, without any tail assumptions, unlike estimating means.
- Can run A/B tests and get always valid p-values for testing the difference in quantiles.

Howard, Ramdas '19

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

Reasons to use quantiles include:

- Quantiles always exist for any distribution, while means (moments) do not always exist (eg: Cauchy).
- Quantiles can be defined for any totally ordered space, eg: ratings A-F, where "distance between ratings" undefined.
- Estimating quantiles can be done sequentially, without any tail assumptions, unlike estimating means.
- Can run A/B tests and get always valid p-values for testing the difference in quantiles.
- Can run bandit experiments, including best-arm identification.

Howard, Ramdas '19

2. Switch from estimating means to quantiles?

Let $X \sim F$. Define the α-quantile as $q_\alpha := \sup\{x : F(x) \leq \alpha\}$.

(hence $q_{1/2}$ is the median)

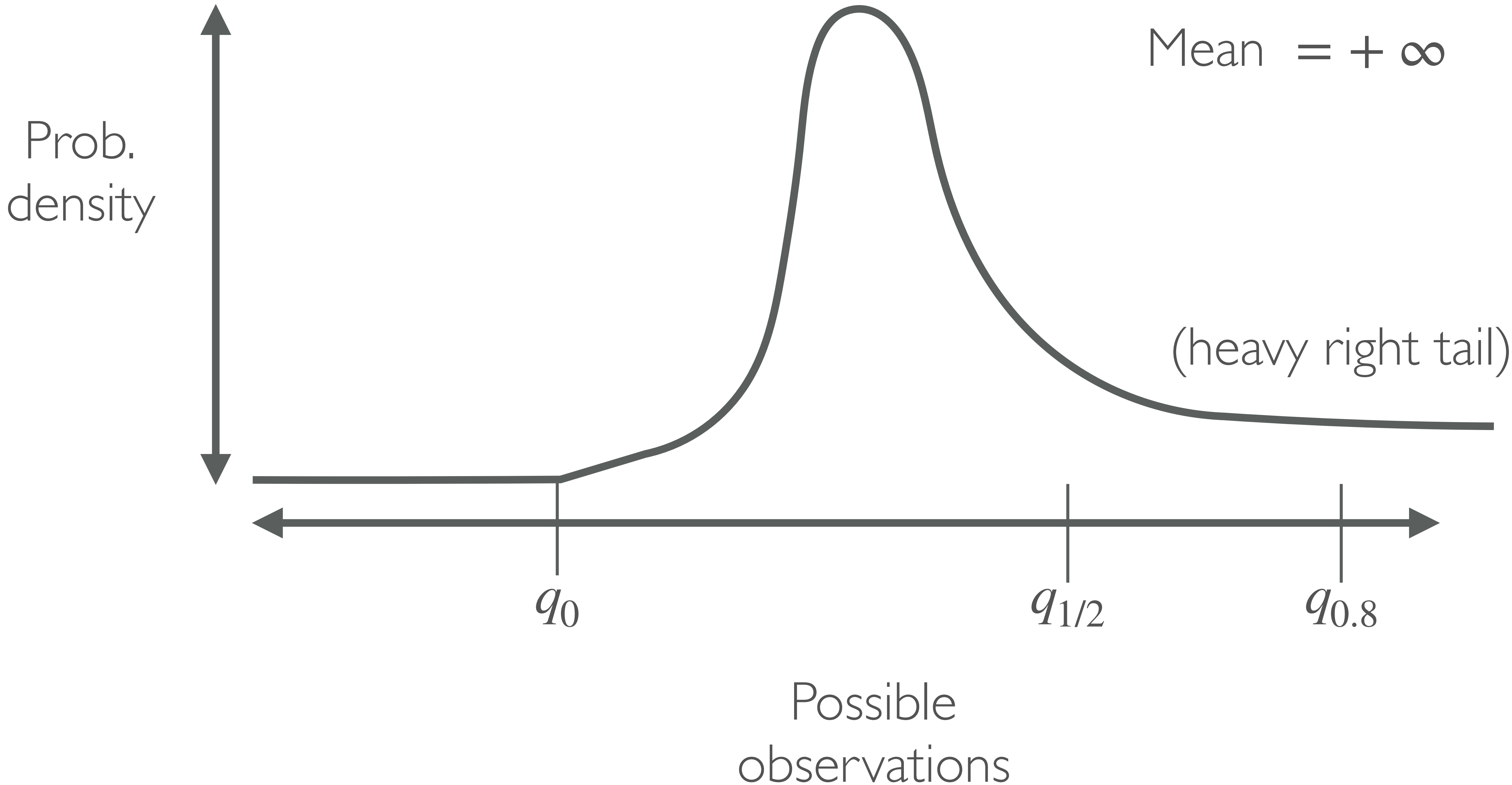
Reasons to use quantiles include:

- Quantiles always exist for any distribution, while means (moments) do not always exist (eg: Cauchy).
- Quantiles can be defined for any totally ordered space, eg: ratings A-F, where "distance between ratings" undefined.
- Estimating quantiles can be done sequentially, without any tail assumptions, unlike estimating means.
- Can run A/B tests and get always valid p-values for testing the difference in quantiles.
- Can run bandit experiments, including best-arm identification.

Can also estimate all quantiles simultaneously!

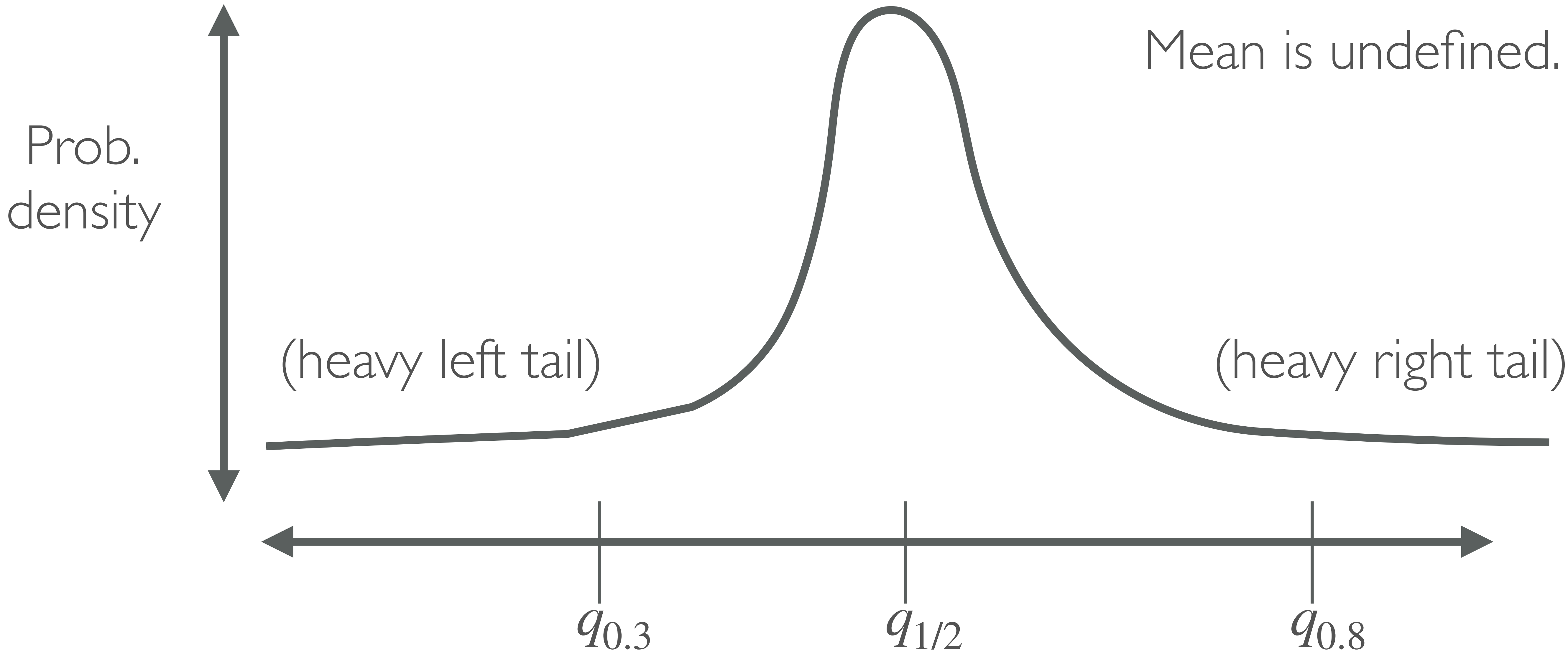
Howard, Ramdas '19

2. Quantiles are informative for heavy tails



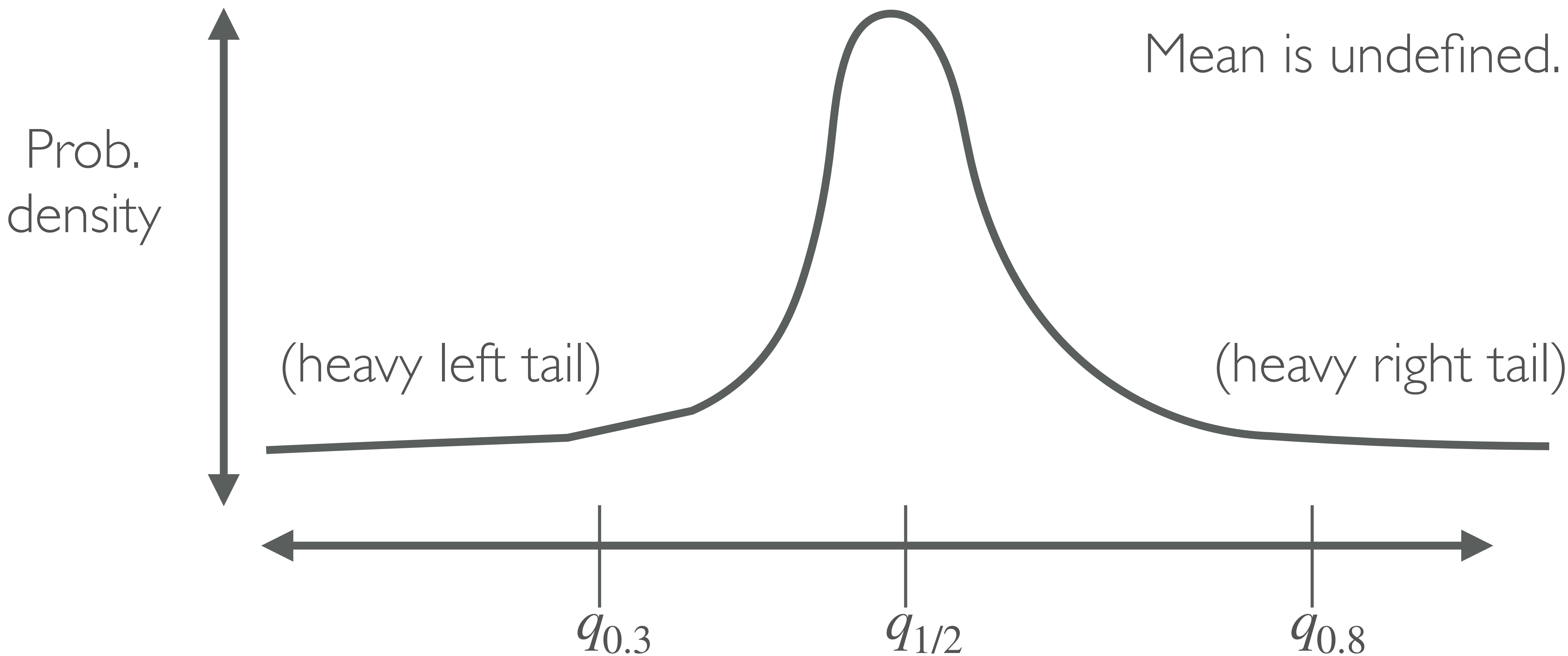
Eg: amount of time spent on Reddit

2. The mean need not even exist (eg: Cauchy)



Eg: amount of money won/lost in a casino

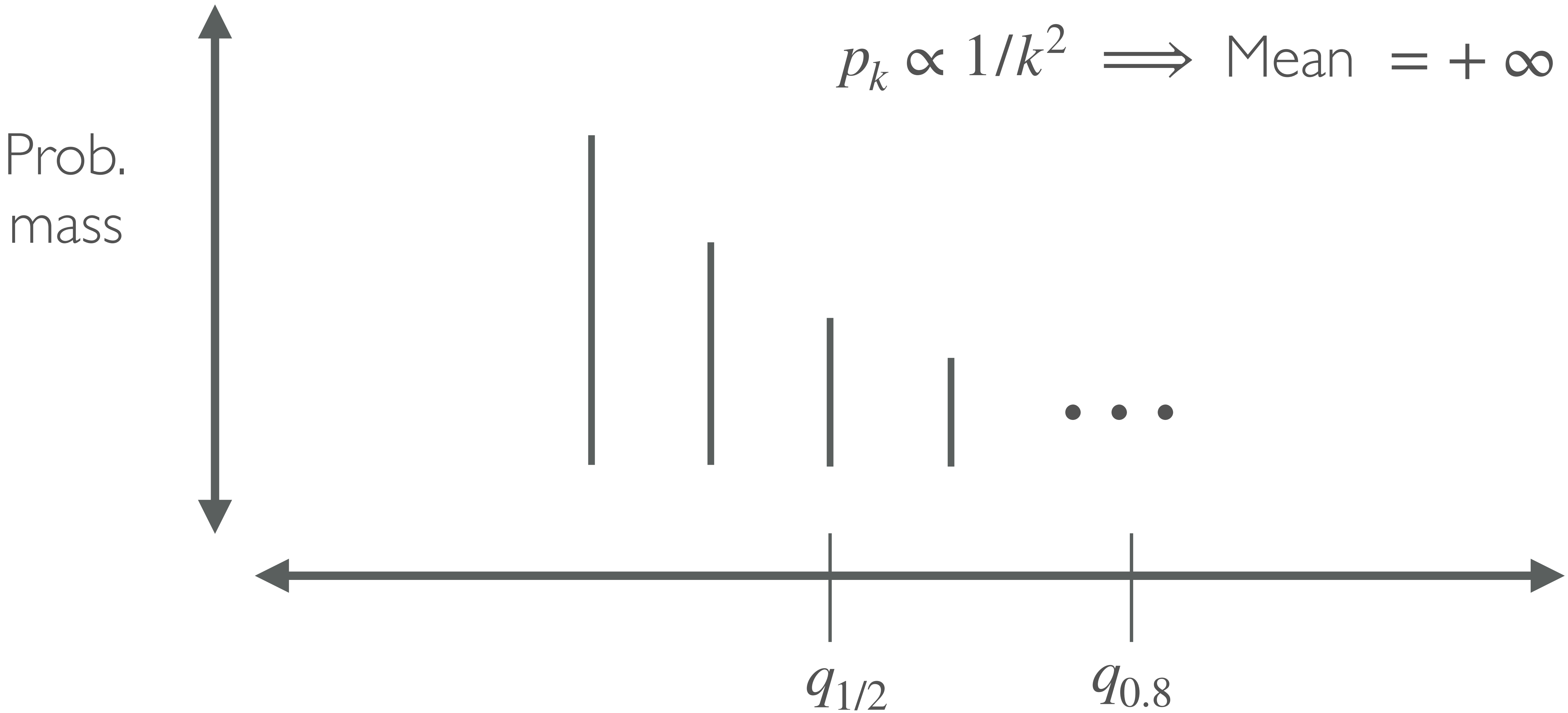
2. The mean need not even exist (eg: Cauchy)



Eg: amount of money won/lost in a casino

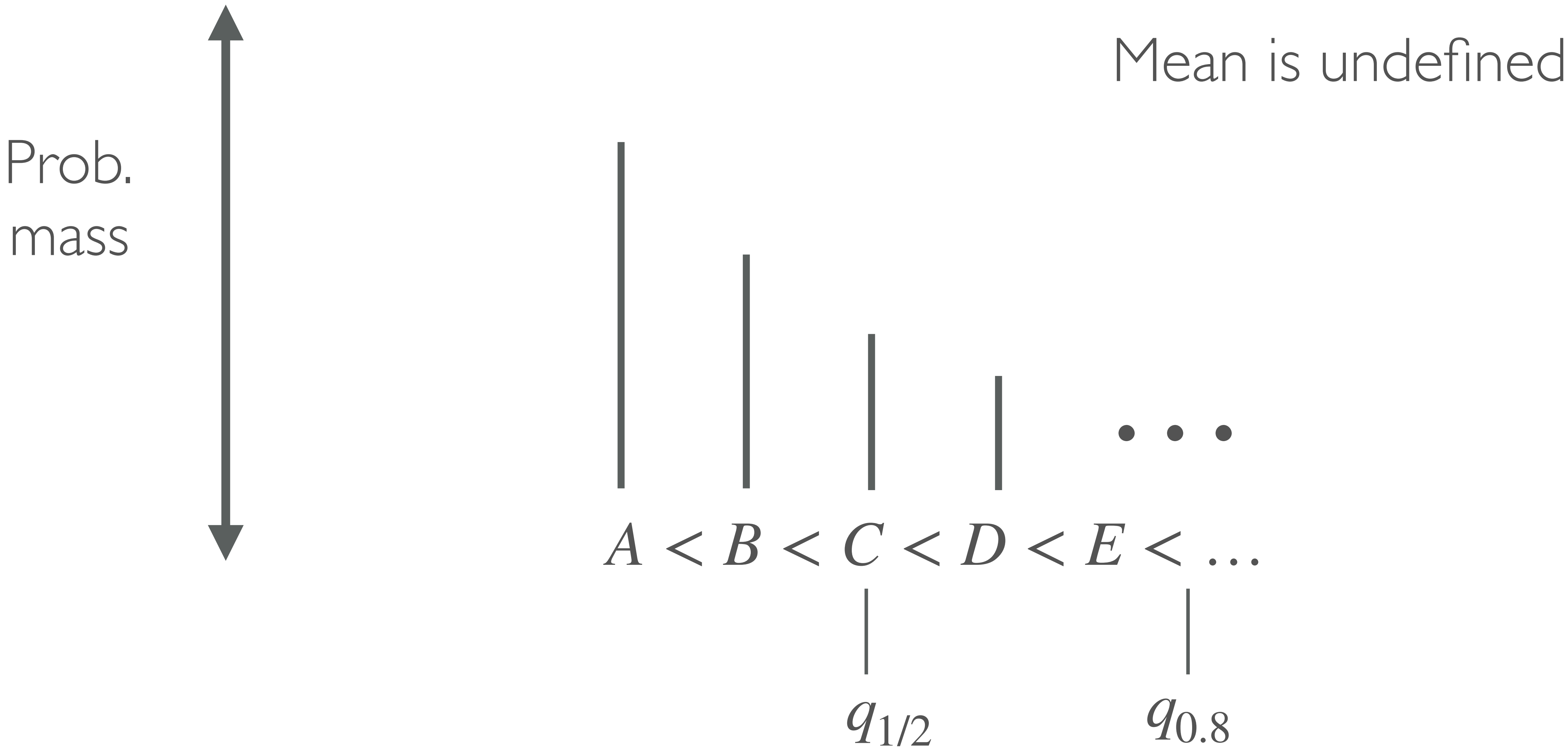
Do not need to resort to trimming "outliers".
(How to pick threshold? Throw away or cap?)

2. The same could arise in discrete settings



Eg: number of links clicked

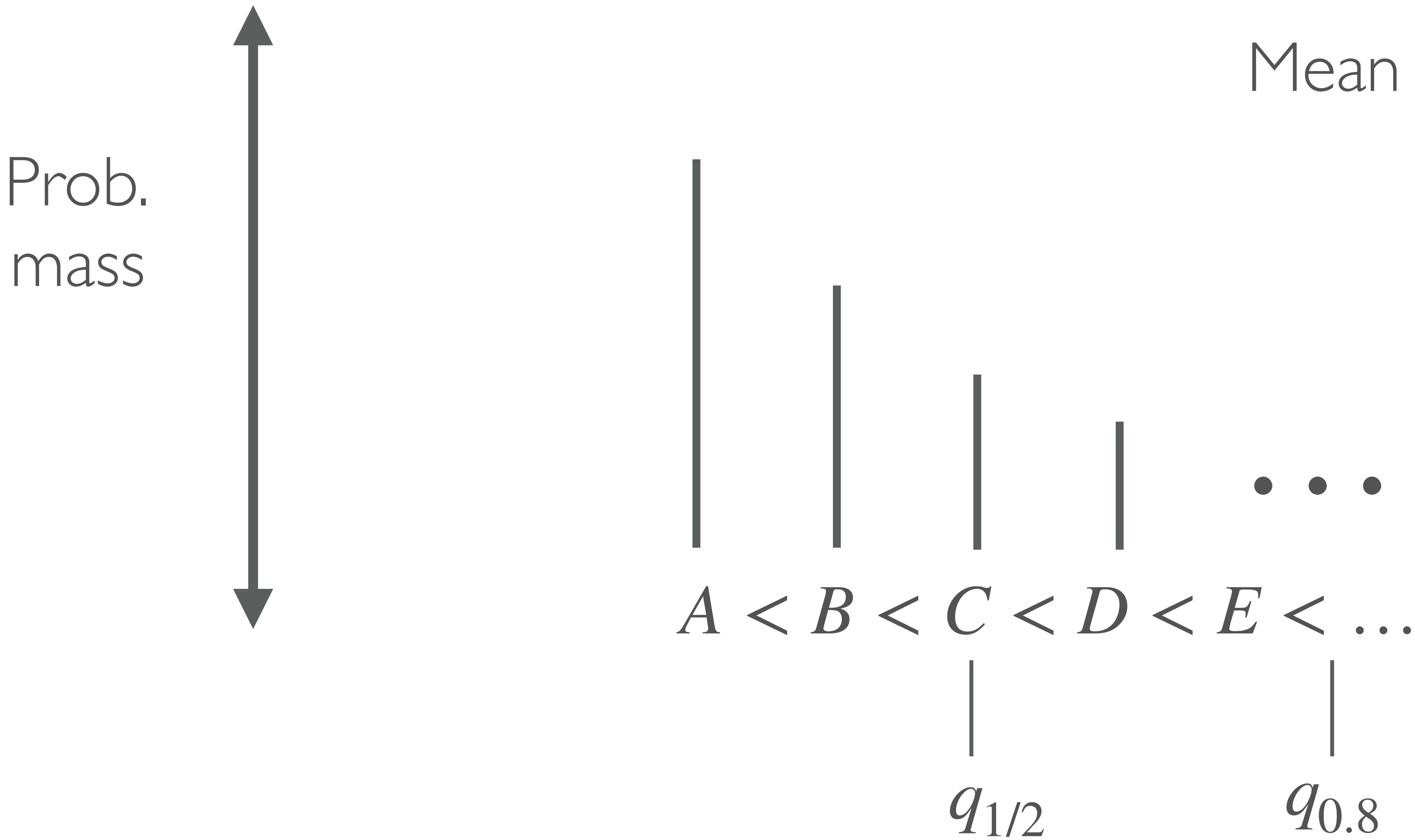
2. Quantile sensible in totally ordered settings



Eg: grades or non-numerical ratings

2. Quantile sensible in totally ordered settings

Mean is undefined.



Eg: grades or non-numerical ratings

Do not need to artificially assign numerical values.
(Are they equally spaced? Spacing and start point matter.)

2. A/B testing with quantiles

First pick target quantile α (say 0.9).

$$H_0 : q_{0.9}(A) = q_{0.9}(B)$$

$$H_1 : q_{0.9}(A) < q_{0.9}(B)$$

2. A/B testing with quantiles

First pick target quantile α (say 0.9).

$$H_0 : q_{0.9}(A) = q_{0.9}(B)$$

$$H_1 : q_{0.9}(A) < q_{0.9}(B)$$

Howard, Ramdas '19

2. A/B testing with quantiles

First pick target quantile α (say 0.9).

$$H_0 : q_{0.9}(A) = q_{0.9}(B)$$

$$H_1 : q_{0.9}(A) < q_{0.9}(B)$$

Can construct always valid p-value.

Howard, Ramdas '19

2. A/B testing with quantiles

First pick target quantile α (say 0.9).

$$H_0 : q_{0.9}(A) = q_{0.9}(B)$$

$$H_1 : q_{0.9}(A) < q_{0.9}(B)$$

Can construct always valid p-value.

If numerical, can construct confidence sequence for $q_{0.9}(B) - q_{0.9}(A)$.

Howard, Ramdas '19

2. A/B testing with quantiles

First pick target quantile α (say 0.9).

$$H_0 : q_{0.9}(A) = q_{0.9}(B)$$

$$H_1 : q_{0.9}(A) < q_{0.9}(B)$$

Can construct always valid p-value.

If numerical, can construct confidence sequence for $q_{0.9}(B) - q_{0.9}(A)$.

(In that case, one way to define sequential p-value is the smallest δ such that the $(1-\delta)$ CS overlaps with $\mathbb{R}_0^-$.)

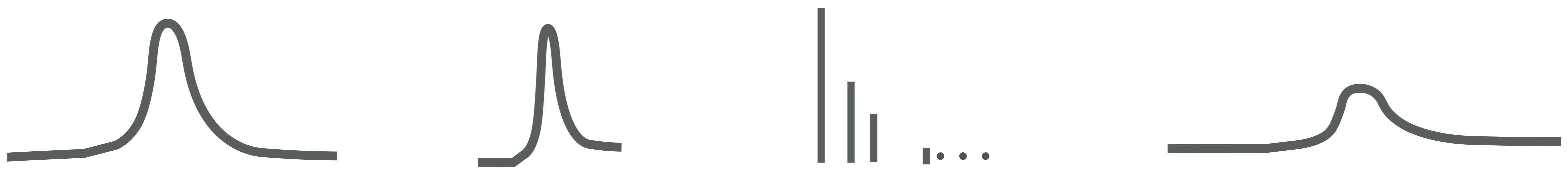
Howard, Ramdas '19

2. Best-arm identification with quantiles

Which arm has the highest 80% quantile?

Howard, Ramdas '19

2. Best-arm identification with quantiles



Which arm has the highest 80% quantile?

Can design MAB algorithms to adaptively determine the "best" arm with a prescribed failure probability.

Howard, Ramdas '19

2. Best-arm identification with quantiles

Which arm has the highest 80% quantile?

Can design MAB algorithms to adaptively determine the "best" arm with a prescribed failure probability.

If the first arm is "special", can design MAB algorithms to adaptively test the null hypothesis that A is best, and get a sequential p-value.

Howard, Ramdas '19

3. Running intersections or minimums: pros/cons

Fact 1: if $P^{(n)}$ is an anytime p-value, so is $\min_{m \le n} P^{(m)}$.

Fact 2: if $(L^{(n)}, U^{(n)})$ is a confidence sequence, so is $\bigcap_{m \le n} (L^{(n)}, U^{(n)})$.

3. Running intersections or minimums: pros/cons

Fact 1: if $P^{(n)}$ is an anytime p-value, so is $\min_{m \le n} P^{(m)}$.

Fact 2: if $(L^{(n)}, U^{(n)})$ is a confidence sequence, so is $\bigcap_{m \le n} (L^{(n)}, U^{(n)})$.

Howard, Ramdas, McAuliffe, Sekhon '19

3. Running intersections or minimums: pros/cons

Fact 1: if $P^{(n)}$ is an anytime p-value, so is $\min_{m \le n} P^{(m)}$.

Fact 2: if $(L^{(n)}, U^{(n)})$ is a confidence sequence, so is $\bigcap_{m \le n} (L^{(n)}, U^{(n)})$.

Pro of taking running intersections of CIs :

- Smaller width, hence tighter inference, without inflating error.

Howard, Ramdas, McAuliffe, Sekhon '19

3. Running intersections or minimums: pros/cons

Fact 1: if $P^{(n)}$ is an anytime p-value, so is $\min_{m \le n} P^{(m)}$.

Fact 2: if $(L^{(n)}, U^{(n)})$ is a confidence sequence, so is $\bigcap_{m \le n} (L^{(n)}, U^{(n)})$.

Pro of taking running intersections of CIs :

- Smaller width, hence tighter inference, without inflating error.

Con of taking running intersections of CIs :

- Can have intervals of decreasing width (great!) and then in the next step, end up with an empty interval (disconcerting).

Howard, Ramdas, McAuliffe, Sekhon '19

3. Running intersections or minimums: pros/cons

Fact 1: if $P^{(n)}$ is an anytime p-value, so is $\min_{m \leq n} P^{(m)}$.

Fact 2: if $(L^{(n)}, U^{(n)})$ is a confidence sequence, so is $\bigcap_{m \leq n} (L^{(n)}, U^{(n)})$.

Pro of taking running intersections of CIs :

- Smaller width, hence tighter inference, without inflating error.

Con of taking running intersections of CIs :

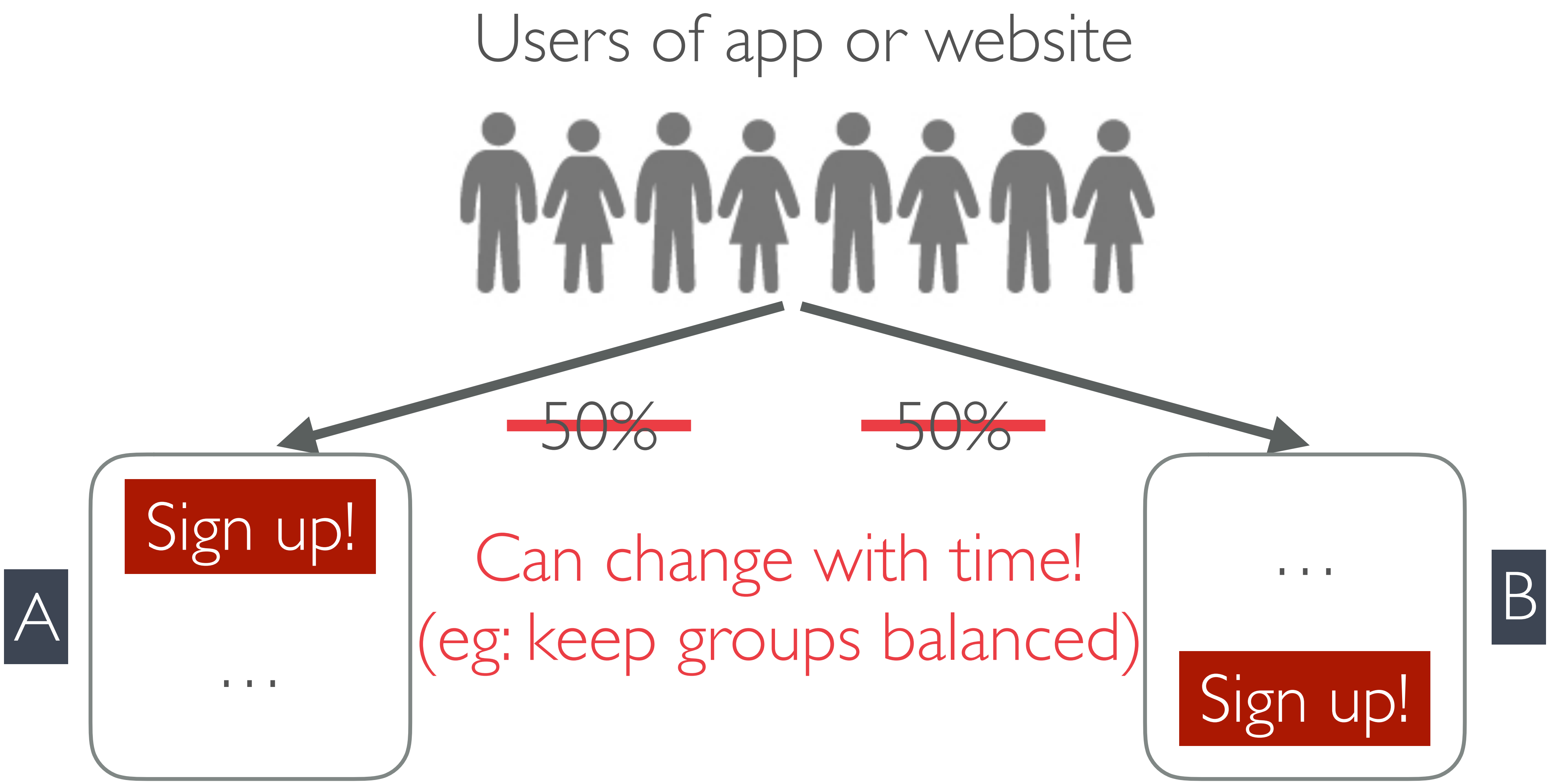
- Can have intervals of decreasing width (great!) and then in the next step, end up with an empty interval (disconcerting).

Pro of ending up with zero width :

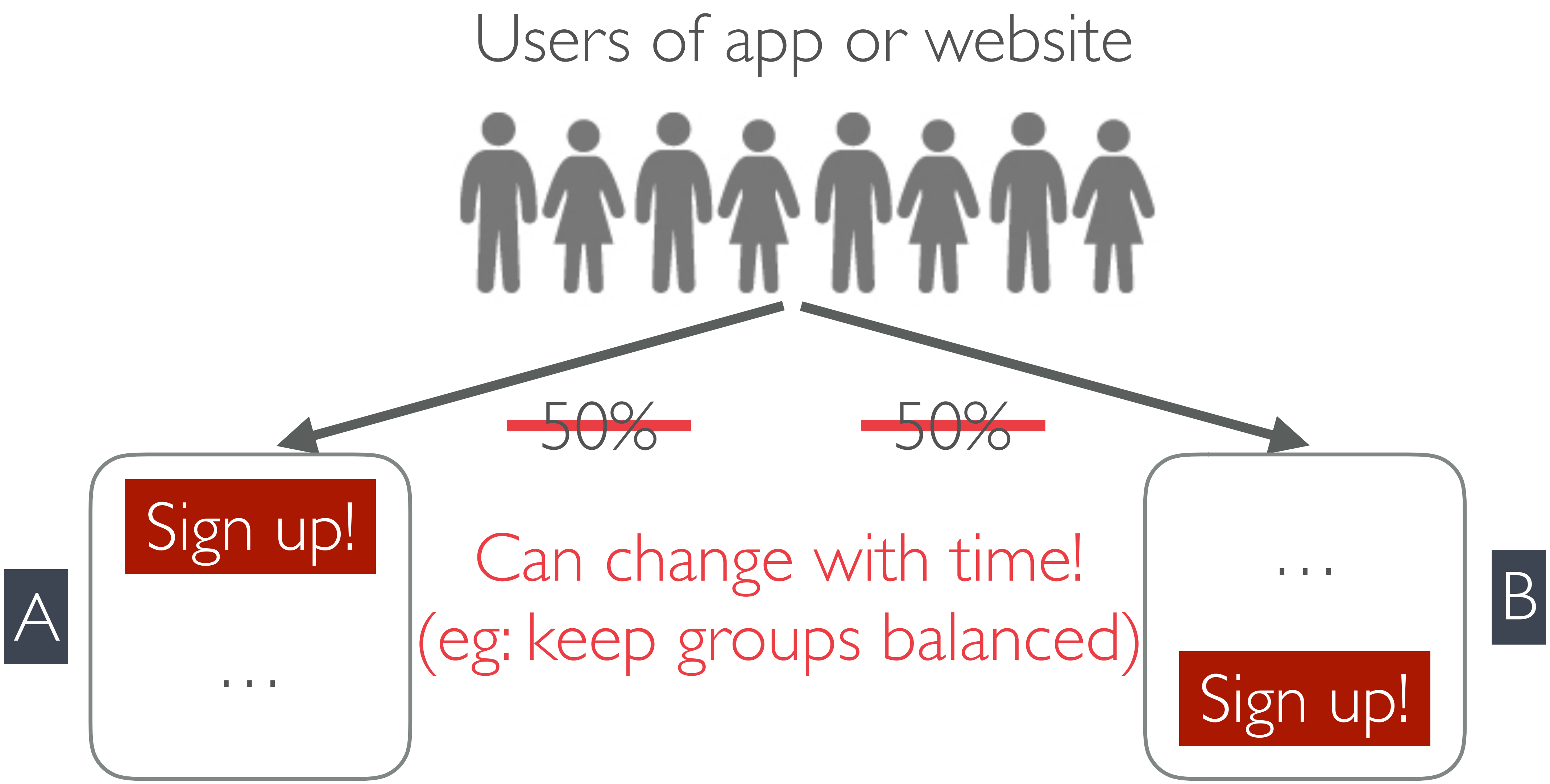
- "Failing loudly": you know you're in the low-probability error event, or assumptions have been violated.

Howard, Ramdas, McAuliffe, Sekhon '19

4. Sequential Average Treatment Effect estimation with adaptive randomization



4. Sequential Average Treatment Effect estimation with adaptive randomization



Can infer the treatment effect sequentially (Neyman-Rubin potential outcomes model) using anytime p-value or CI.

Howard, Ramdas, McAuliffe, Sekhon '19

PART V: Advanced topics (outer sequential process)

[Next 15 mins]

1. Smoothly forgetting the past

Recent tests may be more relevant than older ones, and hence we may wish to **smoothly forget the past.**

1. Smoothly forgetting the past

Recent tests may be more relevant than older ones, and hence we may wish to **smoothly forget the past.**

With this motivation, we may wish to control the

decaying memory FDR: (user-chosen decay $d < 1$)

1. Smoothly forgetting the past

Recent tests may be more relevant than older ones, and hence we may wish to **smoothly forget the past.**

With this motivation, we may wish to control the

decaying memory FDR: (user-chosen decay d < 1)

$$\text{mem-FDR}(T) = \mathbb{E}\left[\frac{\sum_t d^{T-t}\mathbf{1}(\text{false discovery}_t)}{\sum_t d^{T-t}\mathbf{1}(\text{discovery}_t)}\right]$$

Ramdas et al. '17

1. Smoothly forgetting the past

Recent tests may be more relevant than older ones, and hence we may wish to **smoothly forget the past.**

With this motivation, we may wish to control the

decaying memory FDR: (user-chosen decay d < 1)

$$\text{mem-FDR}(T) = \mathbb{E}\left[\frac{\sum_t d^{T-t}\mathbf{1}(\text{false discovery}_t)}{\sum_t d^{T-t}\mathbf{1}(\text{discovery}_t)}\right]$$

(similarly mem-FCR)

Ramdas et al. '17

2. Post-hoc analysis

2. Post-hoc analysis

Katsevich, Ramdas '18

2. Post-hoc analysis

What if you did *not* use an online FCR or FDR algorithm, but at the end of the year, you would like to answer *"based on the decisions made and error levels used, how large could my FCR or FDR be?"*

Katsevich, Ramdas '18

2. Post-hoc analysis

What if you did *not* use an online FCR or FDR algorithm, but at the end of the year, you would like to answer *"based on the decisions made and error levels used, how large could my FCR or FDR be?"*

With probability at least $1 - \delta$ we have

$$FDP_t \leq \frac{1 + \sum_{i \leq t} \alpha_i}{\sum_{i \leq t} R_i} \cdot \frac{\log(1/\delta)}{\log(1 + \log(1/\delta))}$$ simultaneously for all t.

Katsevich, Ramdas '18

2. Post-hoc analysis

What if you did *not* use an online FCR or FDR algorithm, but at the end of the year, you would like to answer *"based on the decisions made and error levels used, how large could my FCR or FDR be?"*

With probability at least $1 - \delta$ we have

$$FDP_t \leq \frac{1 + \sum_{i \leq t} \alpha_i}{\sum_{i \leq t} R_i} \cdot \frac{\log(1/\delta)}{\log(1 + \log(1/\delta))} \text{ simultaneously for all } t.$$

$$FCP_t \leq \frac{1 + \sum_{i \leq t} \alpha_i}{\sum_{i \leq t} S_i} \cdot \frac{\log(1/\delta)}{\log(1 + \log(1/\delta))} \text{ simultaneously for all } t.$$

Katsevich, Ramdas '18

3. Weighted error metrics and algorithms

3. Weighted error metrics and algorithms

The usual error metrics count all mistakes equally.

But, different experiments may have differing importances.

3. Weighted error metrics and algorithms

The usual error metrics count all mistakes equally.

But, different experiments may have differing importances.

Can define "weighted" variants of FDR and FCR
in the natural way: weighted sums in numerator/denominator.

Benjamini, Hochberg '97
Ramdas, Yang, Jordan, Wainwright '17

3. Weighted error metrics and algorithms

The usual error metrics count all mistakes equally.

But, different experiments may have differing importances.

Can define "weighted" variants of FDR and FCR
in the natural way: weighted sums in numerator/denominator.

Online FDR and FCR algorithms can be extended to
control weighted error metrics.

Benjamini, Hochberg '97
Ramdas, Yang, Jordan, Wainwright '17

4. False-sign rate

4. False-sign rate

Sometimes, all we want is a "sign decision" about parameter:
an output of +1 if treatment effect is +ve,
and an output of -1 if treatment effect is -ve,
or no output at all if it is uncertain.

4. False-sign rate

Sometimes, all we want is a "sign decision" about parameter: an output of +1 if treatment effect is +ve, and an output of -1 if treatment effect is -ve, or no output at all if it is uncertain.

We may correspondingly define the false sign rate as

$$FSR := \mathbb{E}\left[\frac{\text{\# incorrect sign decisions made}}{\text{\# sign decisions made}}\right]$$

4. False-sign rate

Sometimes, all we want is a "sign decision" about parameter:
an output of +1 if treatment effect is +ve,
and an output of -1 if treatment effect is -ve,
or no output at all if it is uncertain.

We may correspondingly define the false sign rate as

$$FSR := \mathbb{E}\left[\frac{\text{\# incorrect sign decisions made}}{\text{\# sign decisions made}}\right]$$

To control the FSR, just using the online FCR algorithm,
and report the sign iff the CI does not contain zero.

Weinstein, Ramdas '19

Open Problems [5 mins]

1. Errors and incentives in large organizations

A large number of different teams run such A/B tests or randomized experiments

From the larger organization's perspective, coordination is necessary to control FDR or FCR, since that might affect the bottom line of the company.

1. Errors and incentives in large organizations

A large number of different teams run
such A/B tests or randomized experiments

From the larger organization's perspective,
coordination is necessary to control FDR or FCR,
since that might affect the bottom line of the company.

But each individual group or team might feel
"why do we have to pay if some other group
is running lots of random tests/experiments"?

1. Errors and incentives in large organizations

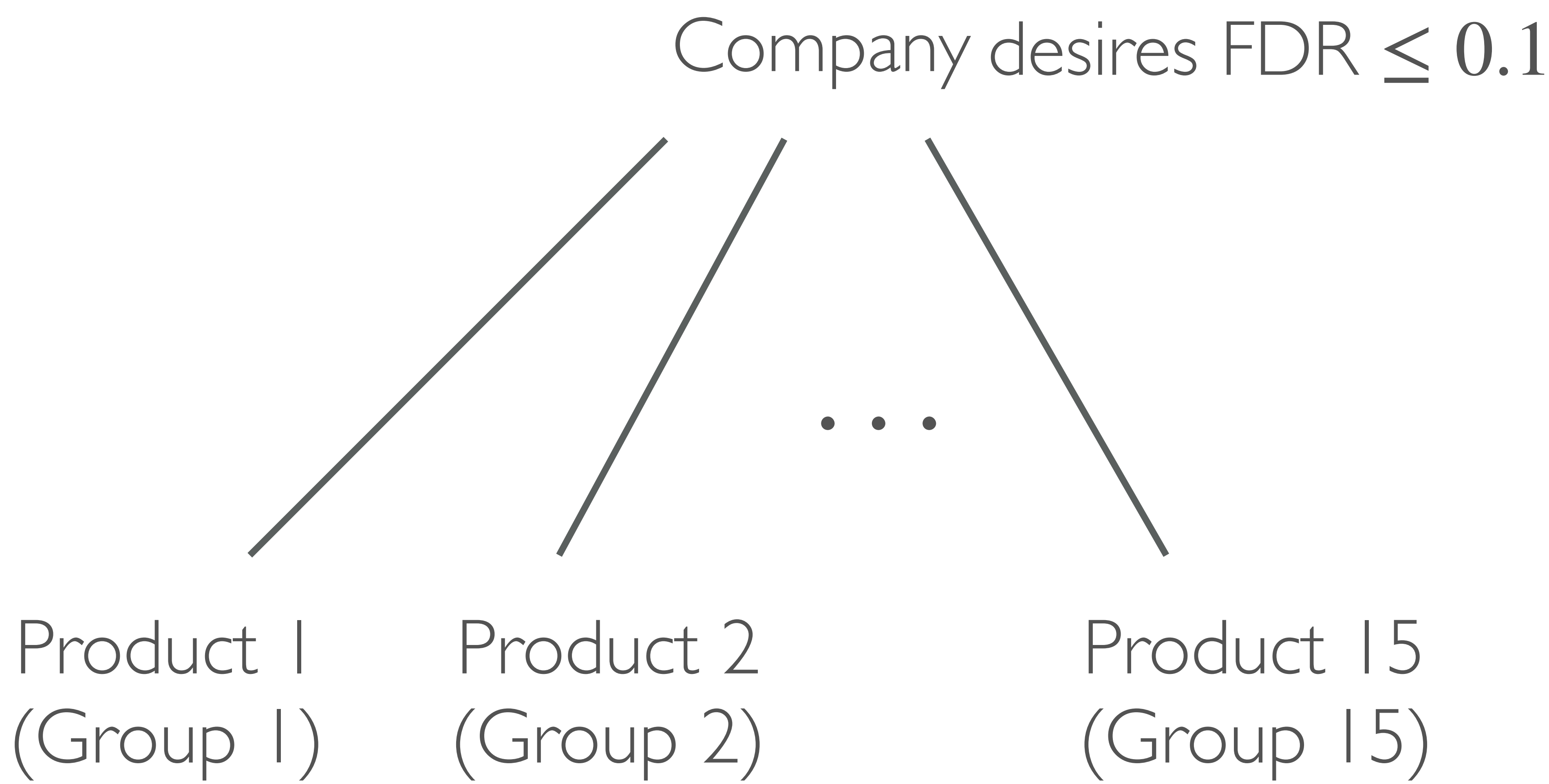
A large number of different teams run
such A/B tests or randomized experiments

From the larger organization's perspective,
coordination is necessary to control FDR or FCR,
since that might affect the bottom line of the company.

But each individual group or team might feel
"why do we have to pay if some other group
is running lots of random tests/experiments"?

How do we align incentives?
Should our notion of error be hierarchical?

1. A hierarchical FDR or FCR control?



The average of group FDRs does not give company FDR.

FDR is additive in the worst case: if each group separately controls FDR at 0.1, the company FDR could be trivial.

2. Utilizing contextual information

Often, we have contextual information about each visitor (sample), like age, gender, etc. These have been utilized for contextual bandit algorithms that minimize regret.

2. Utilizing contextual information

Often, we have contextual information about each visitor (sample), like age, gender, etc. These have been utilized for contextual bandit algorithms that minimize regret.

Is such information useful for hypothesis testing?
How do we use contextual bandits for hypothesis testing?

3. Designing systems that fail loudly

When our assumptions are wrong, and the system is not behaving like intended or expected, how can we *automatically* detect and report this?

3. Designing systems that fail loudly

When our assumptions are wrong, and the system is not behaving like intended or expected, how can we *automatically* detect and report this?

Is it possible to design such self-critical systems that "announce" failures?

Summary [15 mins]

A selective history (inner process)

A selective history (inner process)

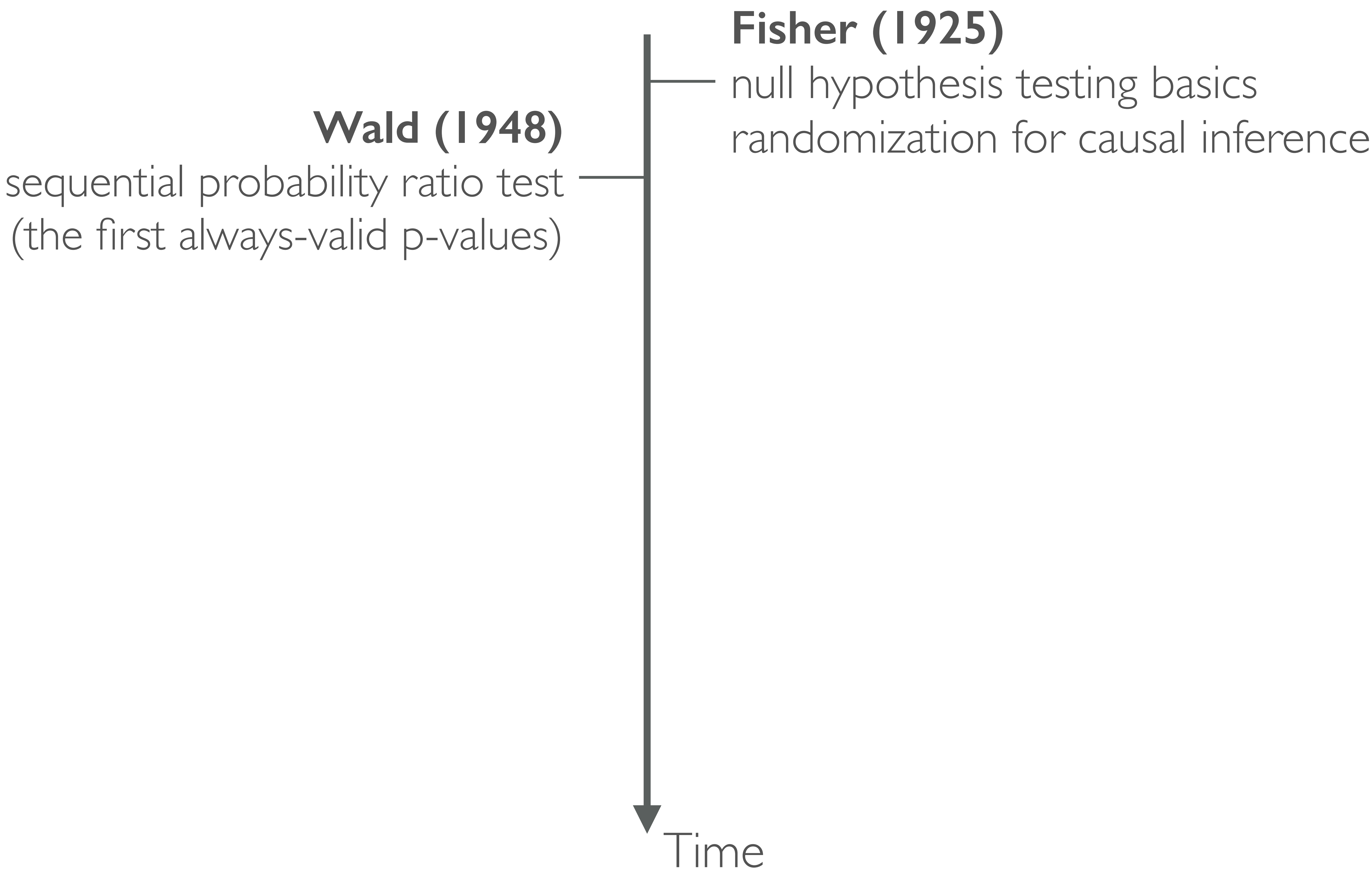
Fisher (1925)

null hypothesis testing basics

randomization for causal inference

Time

A selective history (inner process)



A selective history (inner process)

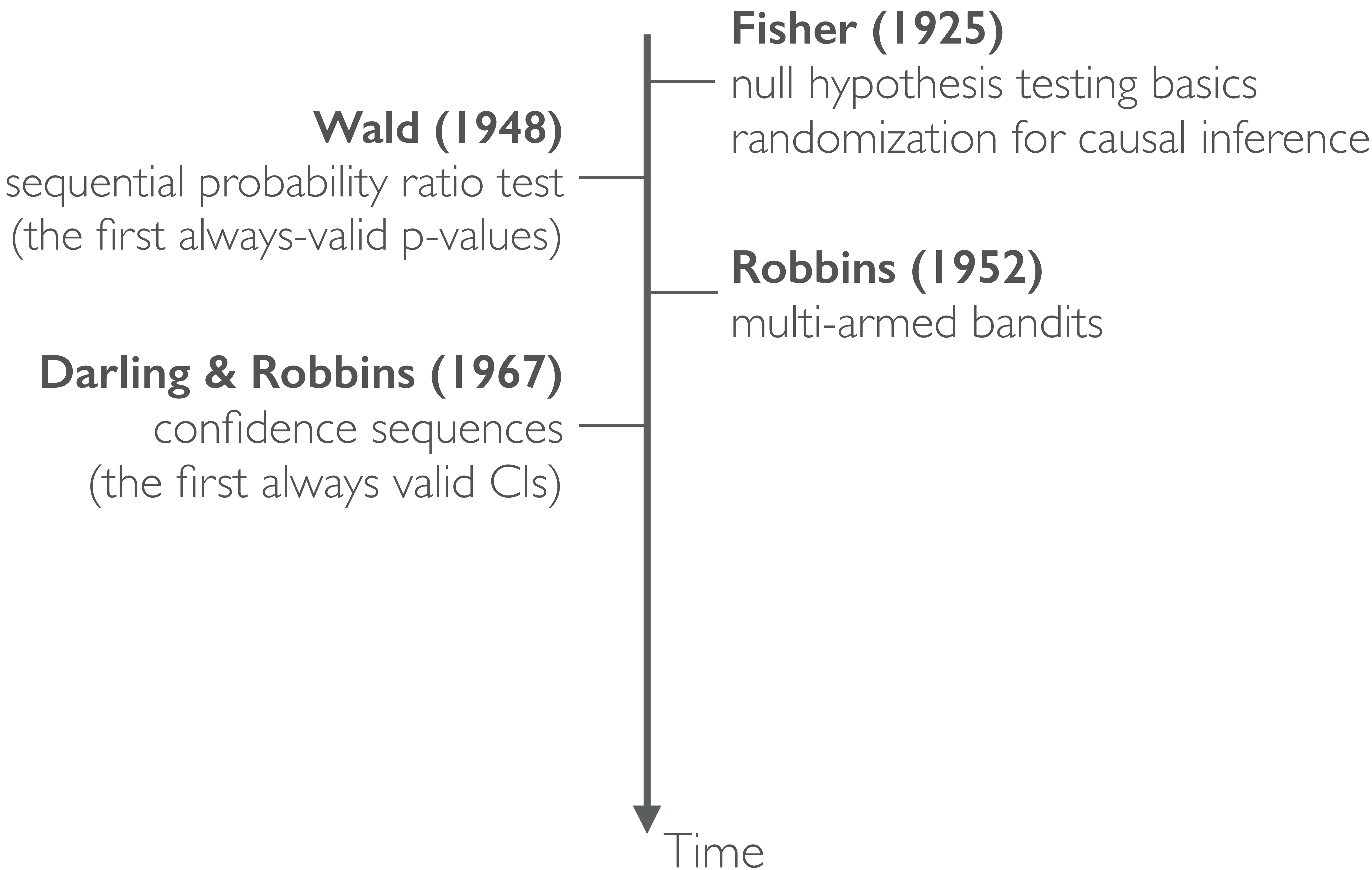
Fisher (1925)
null hypothesis testing basics
randomization for causal inference

Wald (1948)
sequential probability ratio test
(the first always-valid p-values)

Robbins (1952)
multi-armed bandits

Time

A selective history (inner process)



A selective history (inner process)

Fisher (1925)
null hypothesis testing basics
randomization for causal inference

Wald (1948)
sequential probability ratio test
(the first always-valid p-values)

Robbins (1952)
multi-armed bandits

Darling & Robbins (1967)
confidence sequences
(the first always valid CIs)

Lai, Siegmund,… (1970s)
confidence sequences, inference
after stopping experiments

Time

A selective history (inner process)

Fisher (1925)
null hypothesis testing basics
randomization for causal inference

Wald (1948)
sequential probability ratio test
(the first always-valid p-values)

Robbins (1952)
multi-armed bandits

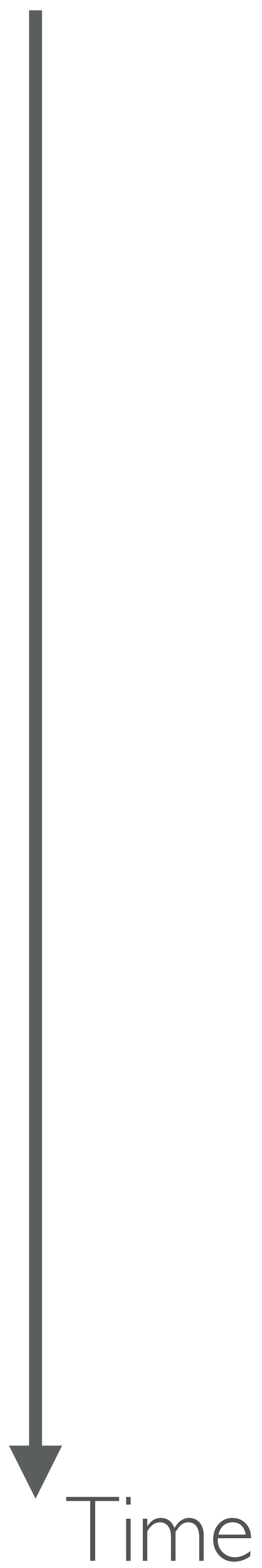
Darling & Robbins (1967)
confidence sequences
(the first always valid CIs)

Lai, Siegmund,… (1970s)
confidence sequences, inference
after stopping experiments

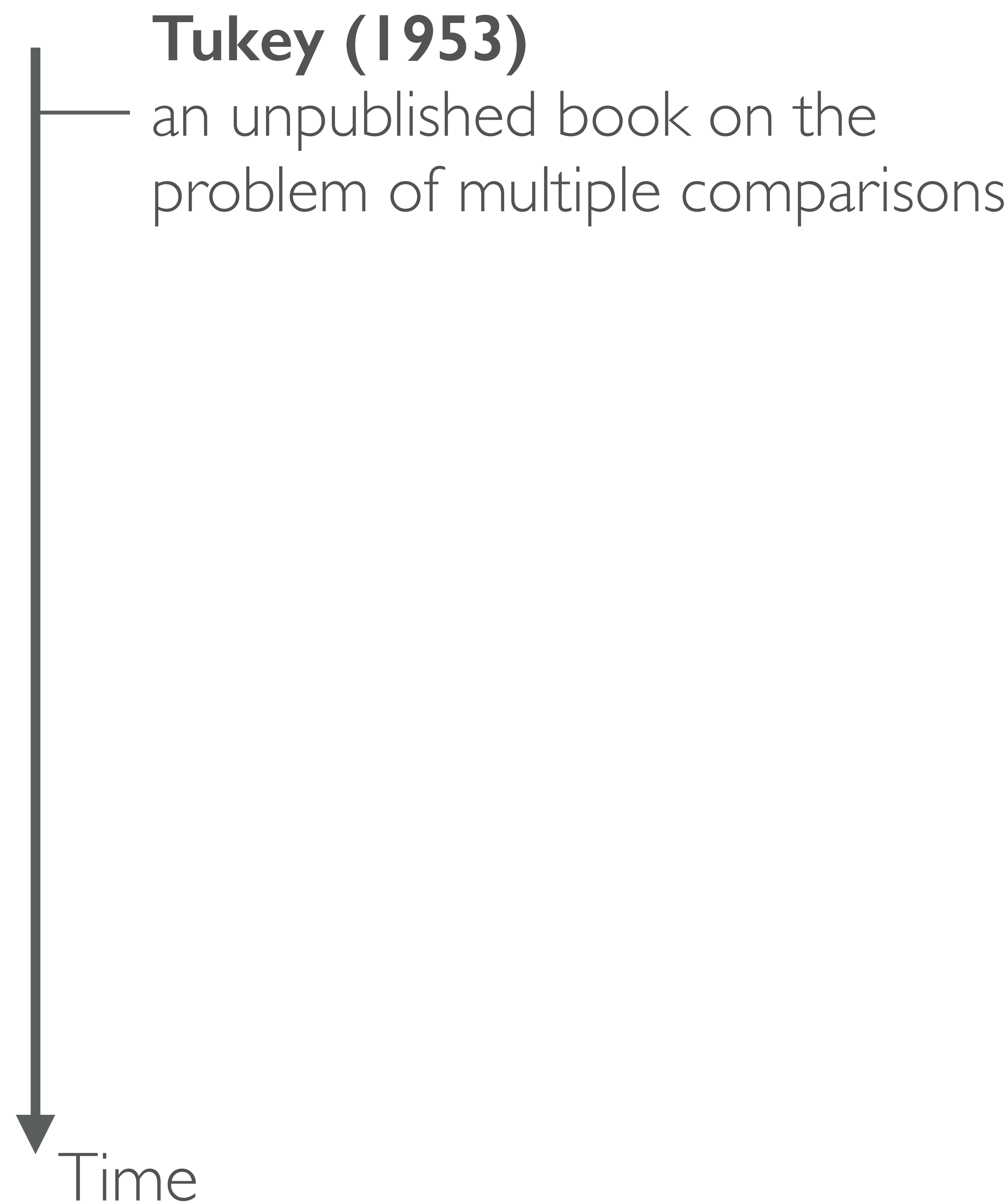
Jennison & Turnbull (1980s)
group sequential methods
(peeking only 2 or 3 times)

Time

A selective history (outer process)



A selective history (outer process)



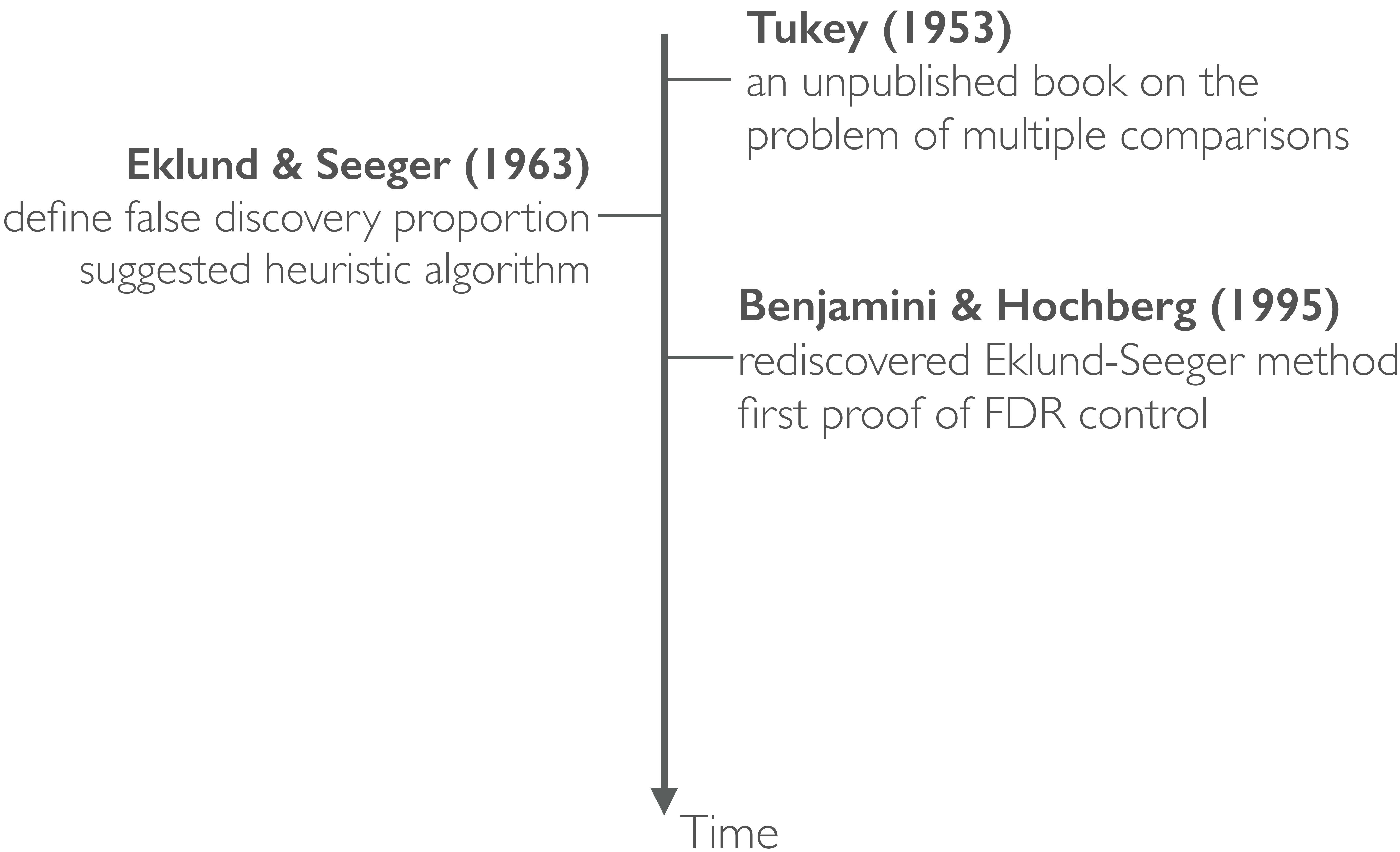
A selective history (outer process)

Tukey (1953)
an unpublished book on the
problem of multiple comparisons

Eklund & Seeger (1963)
define false discovery proportion
suggested heuristic algorithm

Time

A selective history (outer process)



A selective history (outer process)

Tukey (1953)
an unpublished book on the
problem of multiple comparisons

Eklund & Seeger (1963)
define false discovery proportion
suggested heuristic algorithm

Benjamini & Hochberg (1995)
rediscovered Eklund-Seeger method
first proof of FDR control

Benjamini & Yekutieli (2005)
false coverage rate (FCR)
first methods to control it

Time

A selective history (outer process)

Tukey (1953)
an unpublished book on the problem of multiple comparisons

Eklund & Seeger (1963)
define false discovery proportion
suggested heuristic algorithm

Benjamini & Hochberg (1995)
rediscovered Eklund-Seeger method
first proof of FDR control

Benjamini & Yekutieli (2005)
false coverage rate (FCR)
first methods to control it

Foster & Stine (2008)
conceptualized online FDR control
first method to control it

Time

In this tutorial, you learnt the basics of

In this tutorial, you learnt the basics of

- **How to think about a single experiment**
  A. Why peeking is an issue in practice
  B. Why applying a t-test repeatedly inflates errors
  C. Anytime confidence intervals and p-values

In this tutorial, you learnt the basics of

- **How to think about a single experiment**
  A. Why peeking is an issue in practice
  B. Why applying a t-test repeatedly inflates errors
  C. Anytime confidence intervals and p-values

- **How to think about a sequence of experiments**
  A. Why selective reporting is an issue in practice
  B. Why Benjamini-Hochberg fails in the online setting
  C. Online FCR and FDR controlling algorithms

In this tutorial, you learnt the basics of

- **How to think about a single experiment**
  A. Why peeking is an issue in practice
  B. Why applying a t-test repeatedly inflates errors
  C. Anytime confidence intervals and p-values

- **How to think about a sequence of experiments**
  A. Why selective reporting is an issue in practice
  B. Why Benjamini-Hochberg fails in the online setting
  C. Online FCR and FDR controlling algorithms

- **How to think about doubly-sequential experimentation**
  A. Using anytime CIs with online FCR control
  B. Using anytime p-values with online FDR control
  C. Handling asynchronous tests with local dependence

You also learnt some advanced topics:

You also learnt some advanced topics:

- **Within a single experiment:**
    A. Using bandits for hypothesis testing
    B. Quantiles can be estimated sequentially
    C. The pros and cons of running intersections
    D. SATE with adaptive randomization

You also learnt some advanced topics:

- **Within a single experiment:**
  A. Using bandits for hypothesis testing
  B. Quantiles can be estimated sequentially
  C. The pros and cons of running intersections
  D. SATE with adaptive randomization

- **Across experiments:**
  A. Error metrics with decaying-memory
  B. The false sign rate
  C. Weighted error metrics
  D. Post-hoc analysis

You also learnt some advanced topics:

- **Within a single experiment:**
  A. Using bandits for hypothesis testing
  B. Quantiles can be estimated sequentially
  C. The pros and cons of running intersections
  D. SATE with adaptive randomization

- **Across experiments:**
  A. Error metrics with decaying-memory
  B. The false sign rate
  C. Weighted error metrics
  D. Post-hoc analysis

- **Open problems:**
  A. Incentives/errors within hierarchical organizations
  B. Utilizing contextual information for testing
  C. Designing systems that fail loudly

SOFTWARE

- Within a single experiment:
  Python package called "**confseq**"
  Maintained by Steve Howard (Berkeley)
  Frequent updates + wrappers for months to come

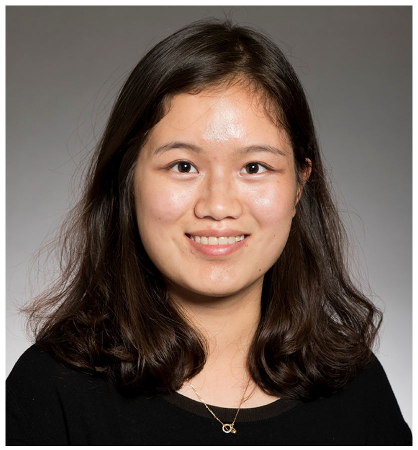
- Across experiments:
  R package called "**onlineFDR**"
  Maintained by David Robertson (Cambridge)
  Frequent updates + wrappers for months to come

References and links at

www.stat.cmu.edu/~aramdas/kdd19/

Collaborators from this talk

Steve Howard



Jinjin Tian

Asaf Weinstein

Eugene Katsevich

Akshay Balsubramani

Tijana Zrnic

David Robertson

Jasjeet Sekhon

Jon McAuliffe

Kevin Jamieson

Fanny Yang

Martin Wainwright

Michael Jordan

Foundations of large-scale "doubly-sequential" experimentation

(KDD tutorial in Anchorage, on 4 Aug 2019)

Aaditya Ramdas

Assistant Professor
Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

Funding welcomed!

www.stat.cmu.edu/~aramdas/kdd19/

Thank you! Questions?