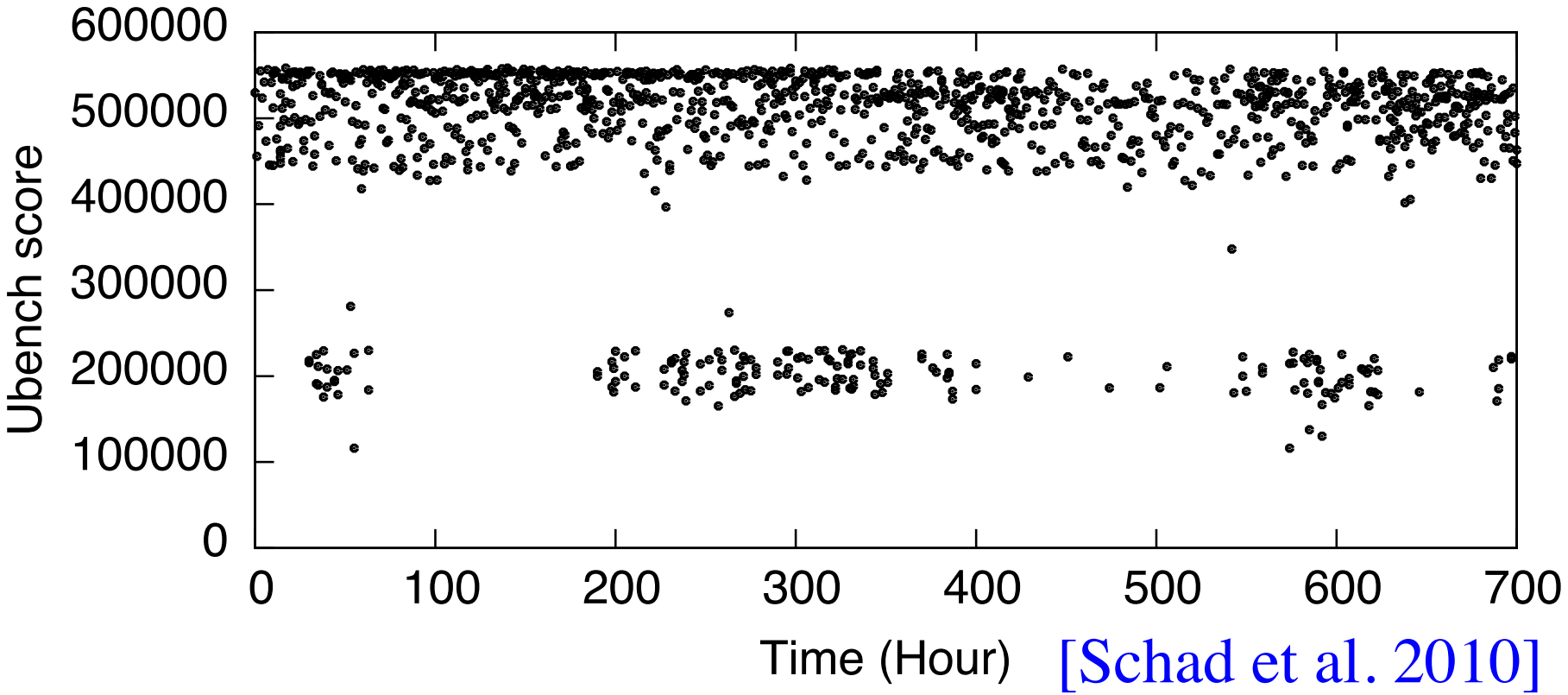
Conducting Repeatable Experiments in Highly Variable Cloud Computing Environments

Ali Abedi and Tim Brecht

UNIVERSITY OF WATERLOO

Motivation

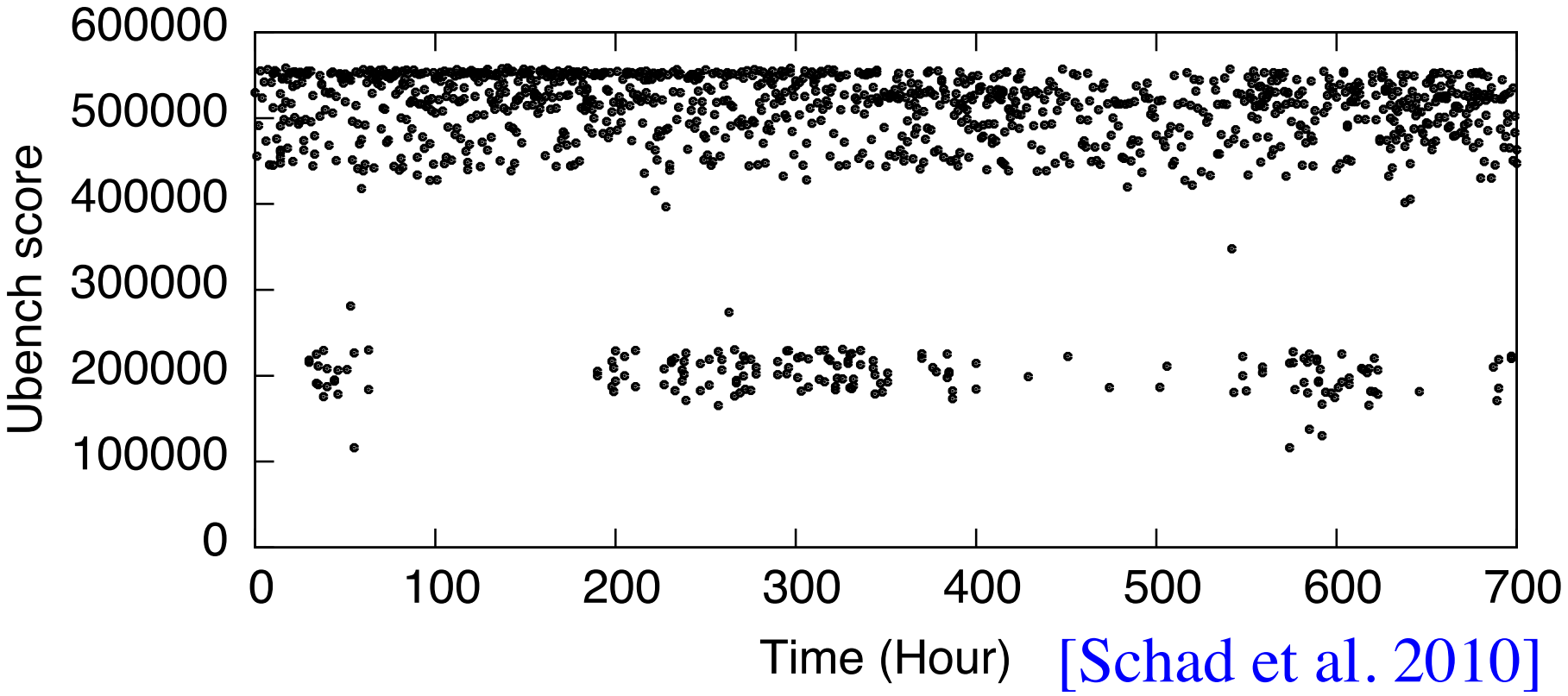
- Performance is highly variable in cloud environments



[Schad et al. 2010]

Motivation

- Performance is highly variable in cloud environments

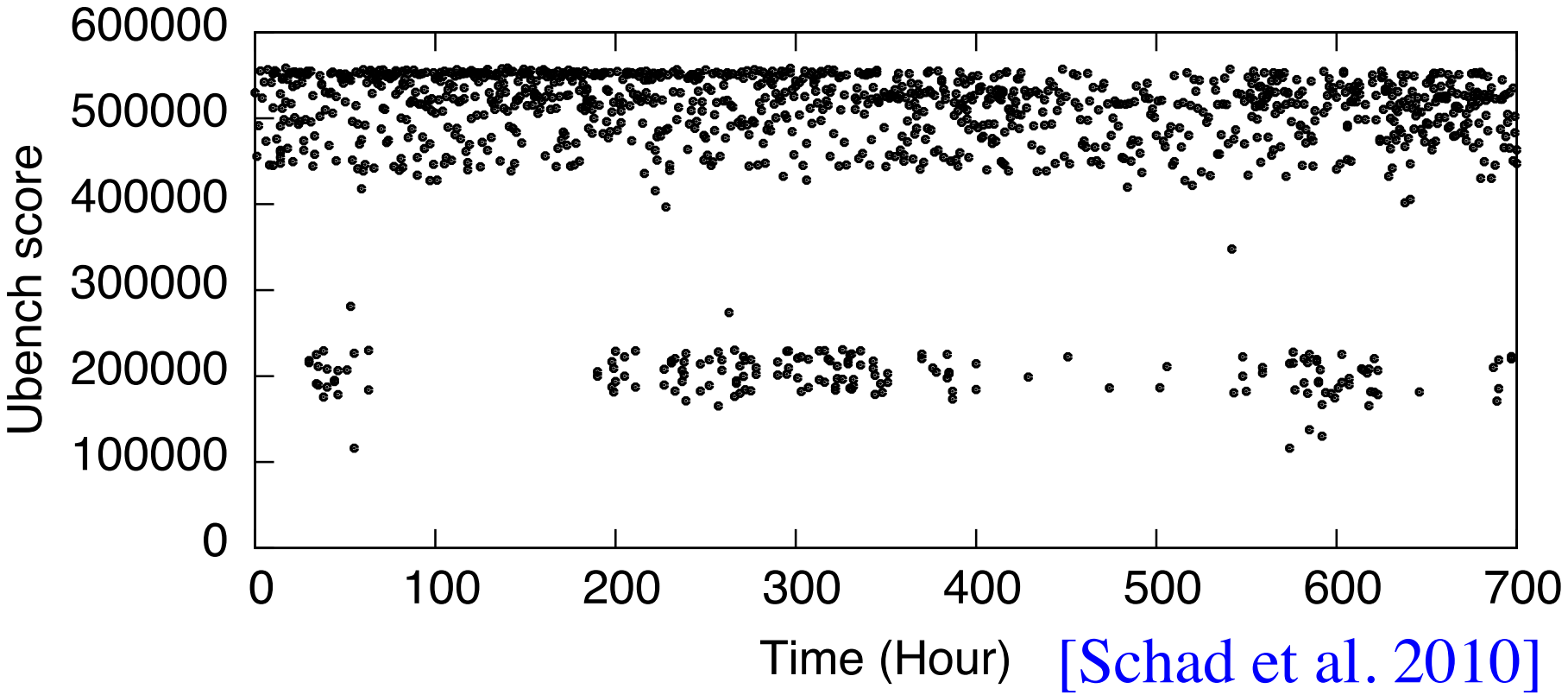


[Schad et al. 2010]

Repeating exactly the same experiment/benchmark !

Motivation

- Performance is highly variable in cloud environments

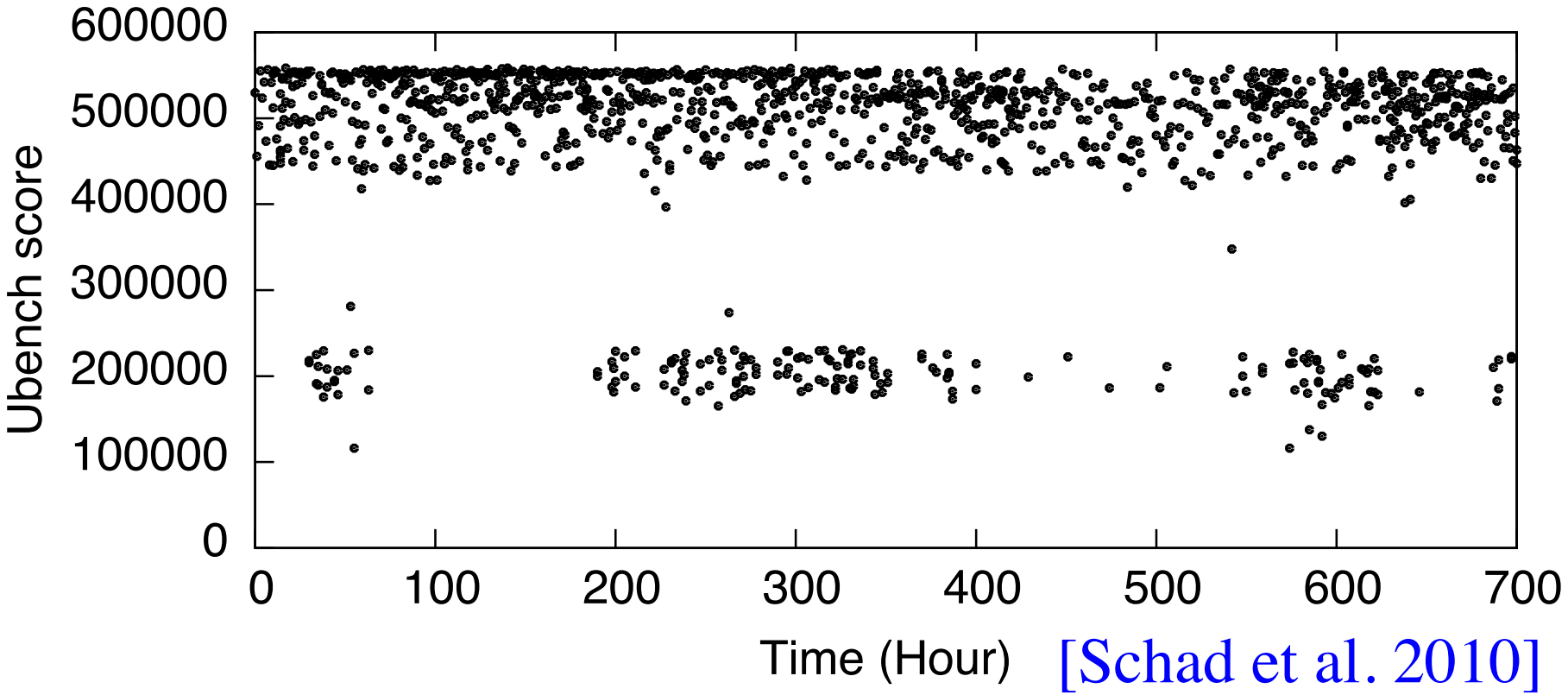


[Schad et al. 2010]

[Iosup et al. 2011] [Ballani et al. 2013] [Leitner et al. 2016]

Motivation

- Performance is highly variable in cloud environments



[Schad et al. 2010]

[Iosup et al. 2011] [Ballani et al. 2013] [Leitner et al. 2016]

High variability for CPU, Memory, Disk and Network

University of Waterloo

Problem

- Given high variability (lack of repeatability)
  - How to compare competing alternatives (FAIRLY) (e.g., A, B, and C)?

Existing Approaches

Single Trial

Existing Approaches

Single Trial

Potential issue: Variation in performance

Existing Approaches

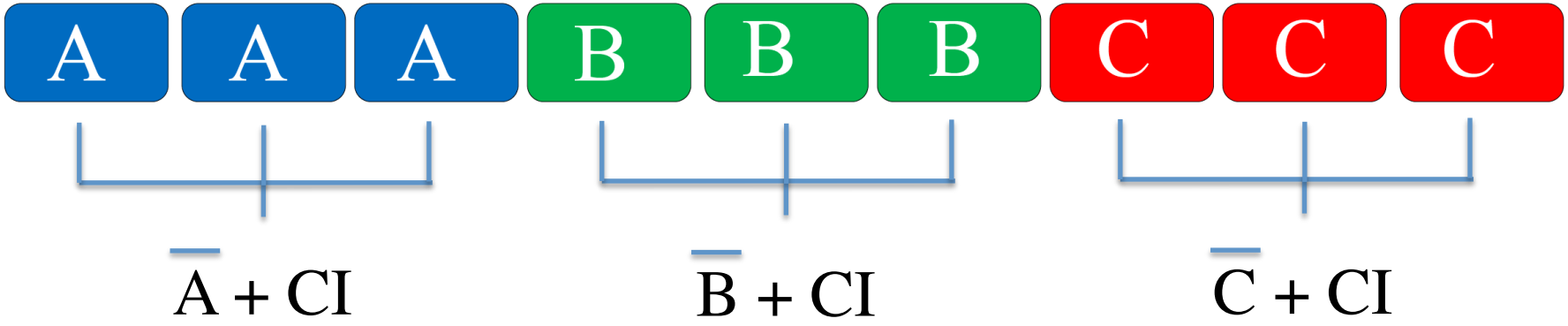
Single Trial

Multiple Consecutive Trials (MCT)

Existing Approaches

Single Trial

Multiple Consecutive Trials (MCT)



Existing Approaches

Single Trial

Multiple Consecutive Trials (MCT)

In 2016 Symposium on Cloud Computing (SoCC):

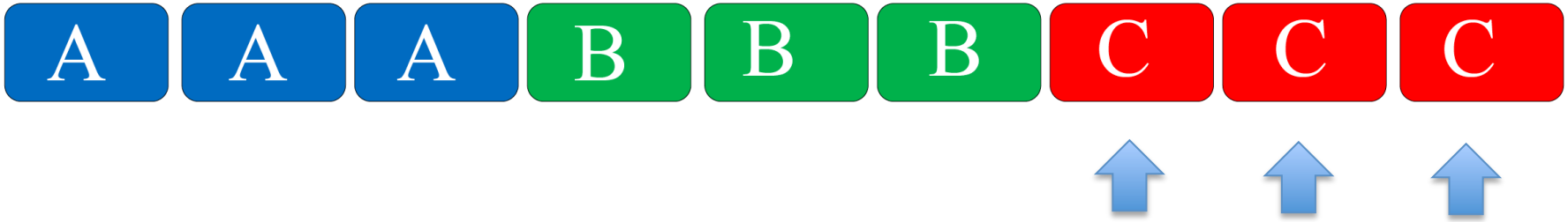
- Use only these two approaches (18 evaluations)
- Mostly Single Trial

  Public Clouds: Single Trial (7), MCT (4)

Existing Approaches

Single Trial

Multiple Consecutive Trials (MCT)



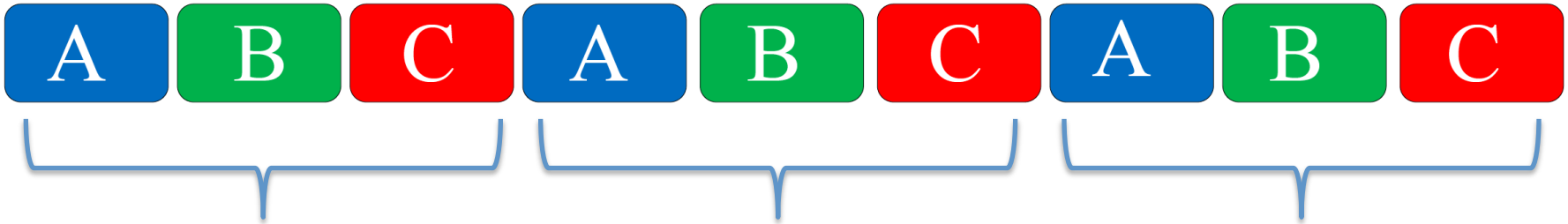
Potential issue: Alternatives are separated by time

Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)



Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

Potential issue: Periodic changes in cloud environment

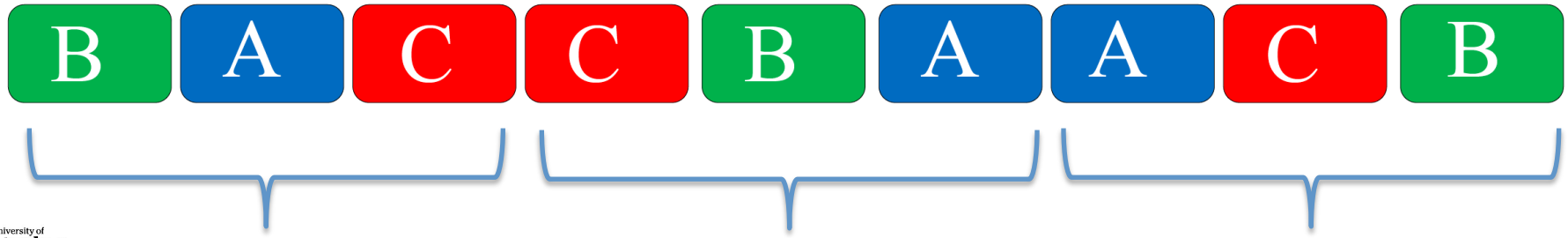
Proposed Approaches

Single Trial

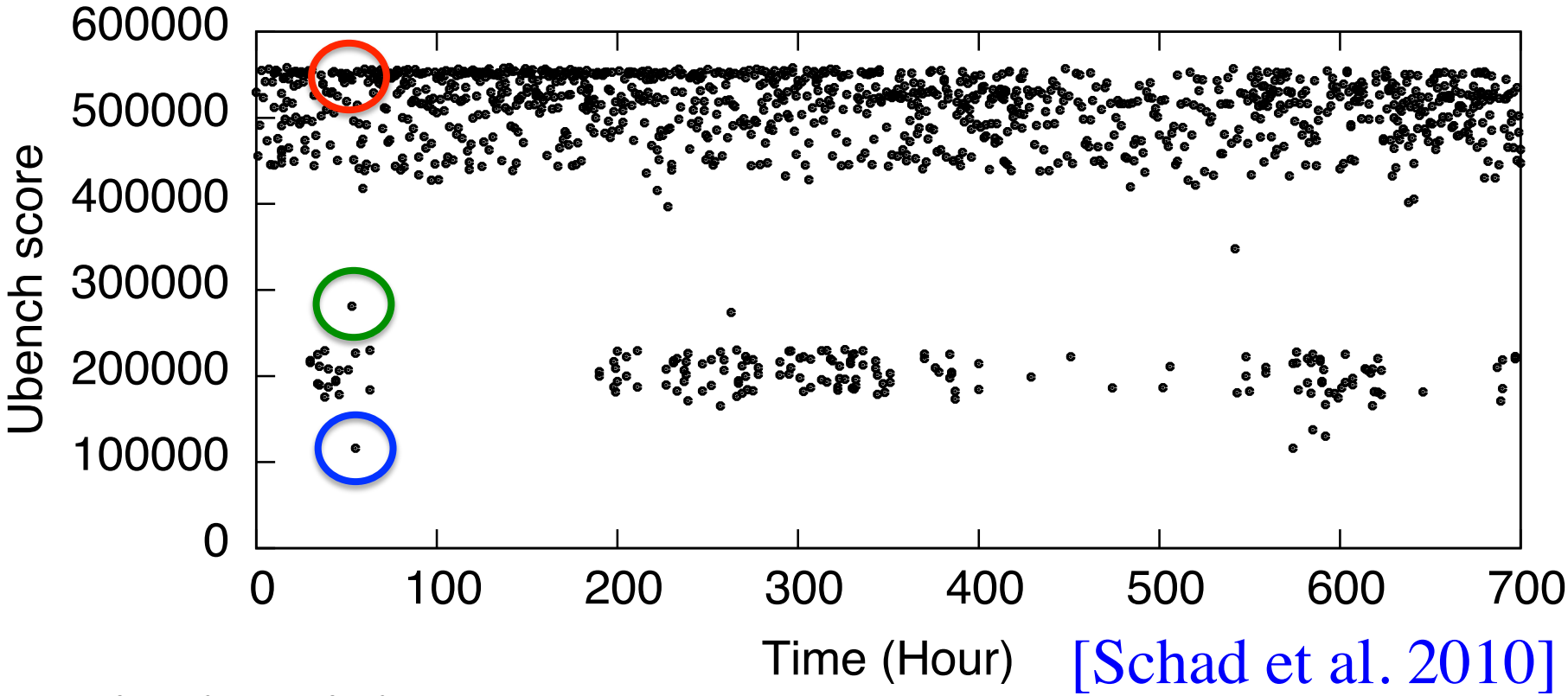
Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

Randomized Multiple Interleaved Trials (RMIT)



Single Trial



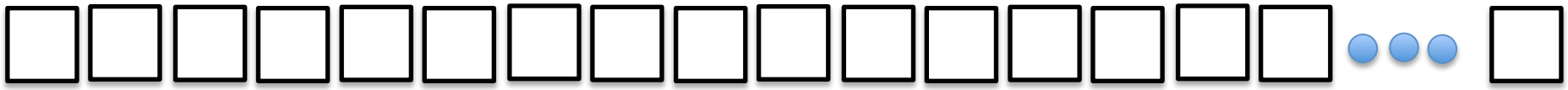
[Schad et al. 2010]

Single Trial

A B C

EC2: CPU performance, US, large instance

Methodology: Multiple Trials

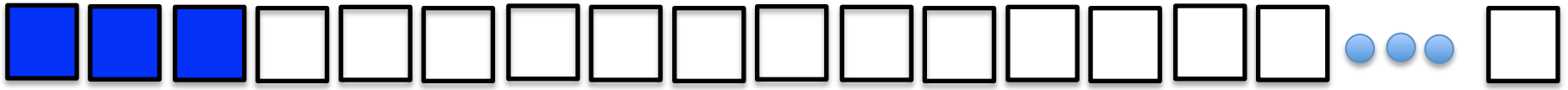


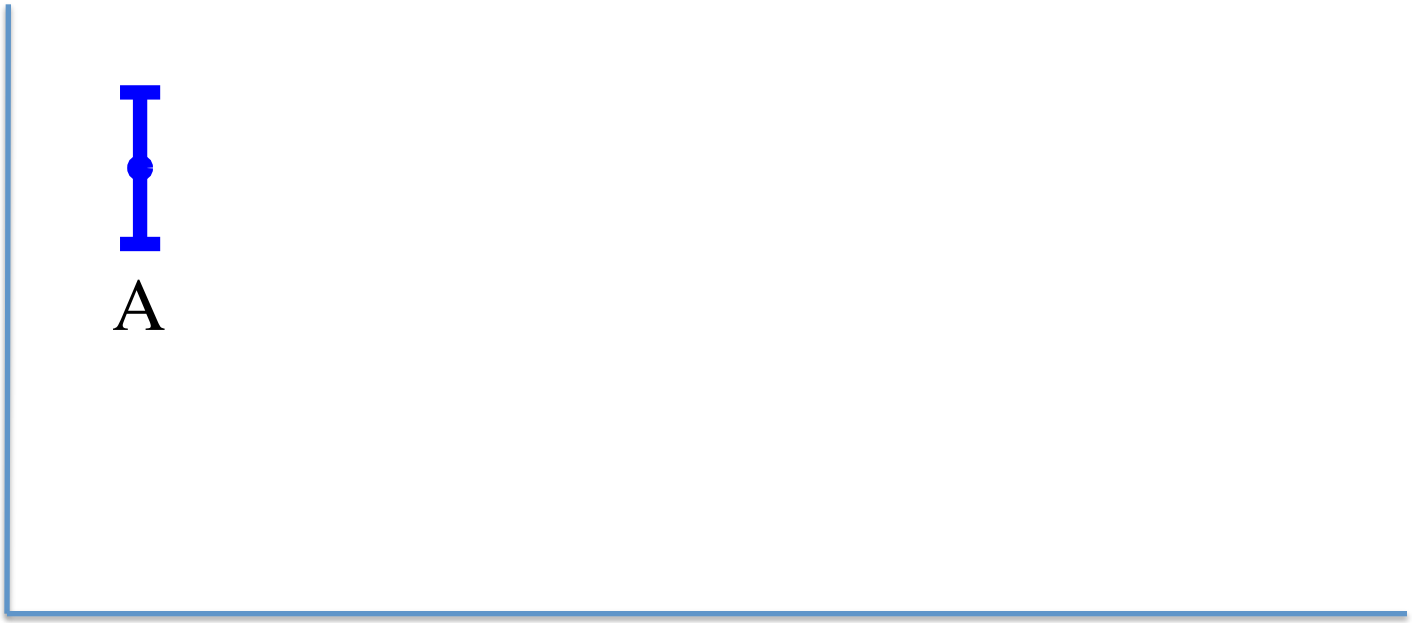
[Schad et al. 2010]

- All data used comes from same set of experiments
- Examine different orderings for trials
  MCT, MIT, RMIT

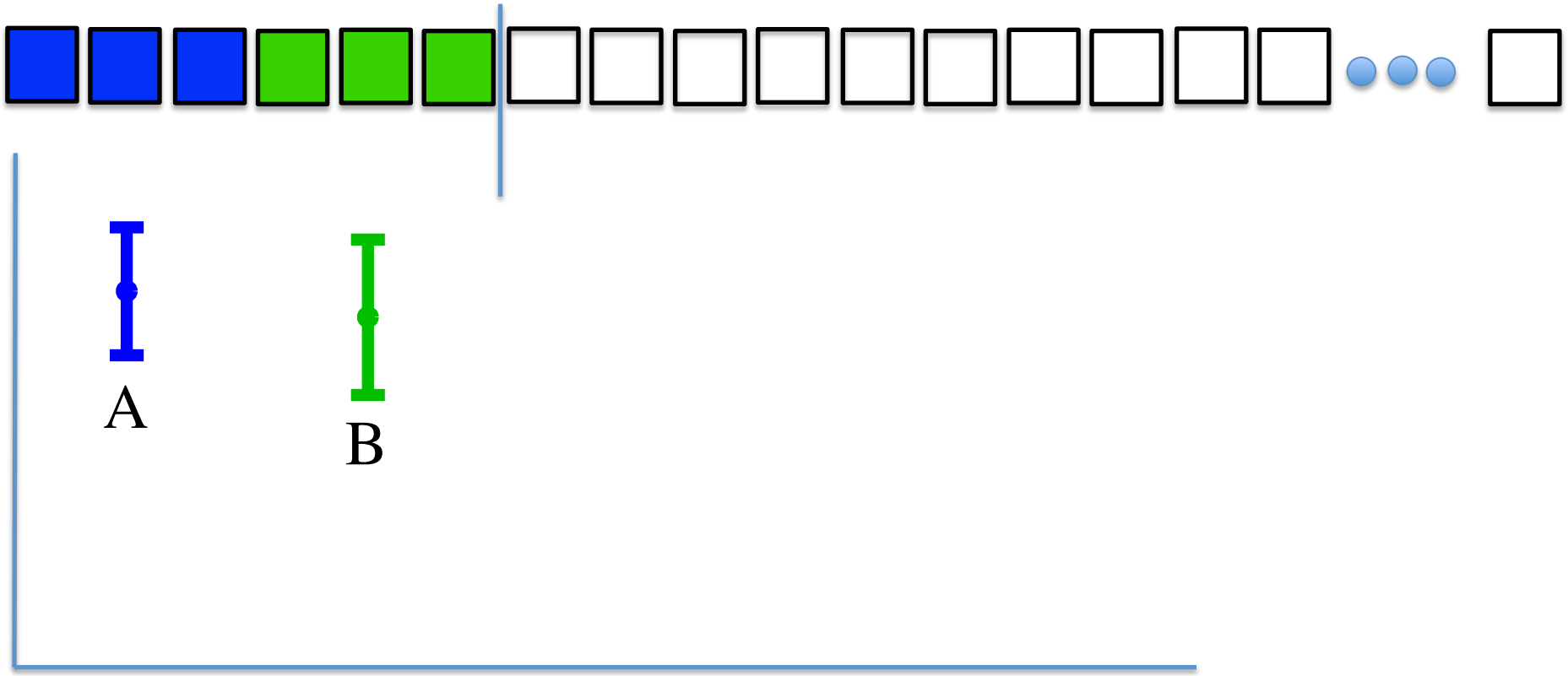
- Consider two different alternatives

Methodology: Consecutive Trials

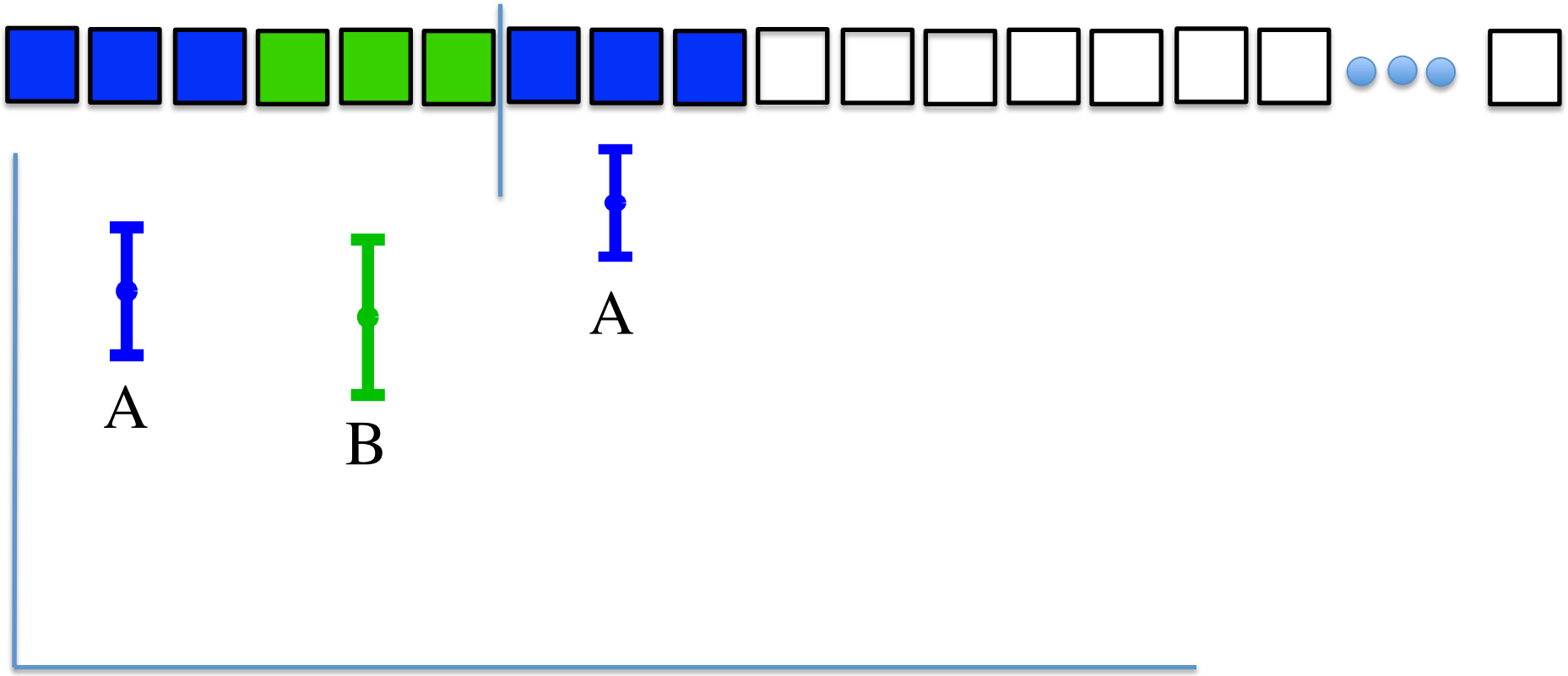




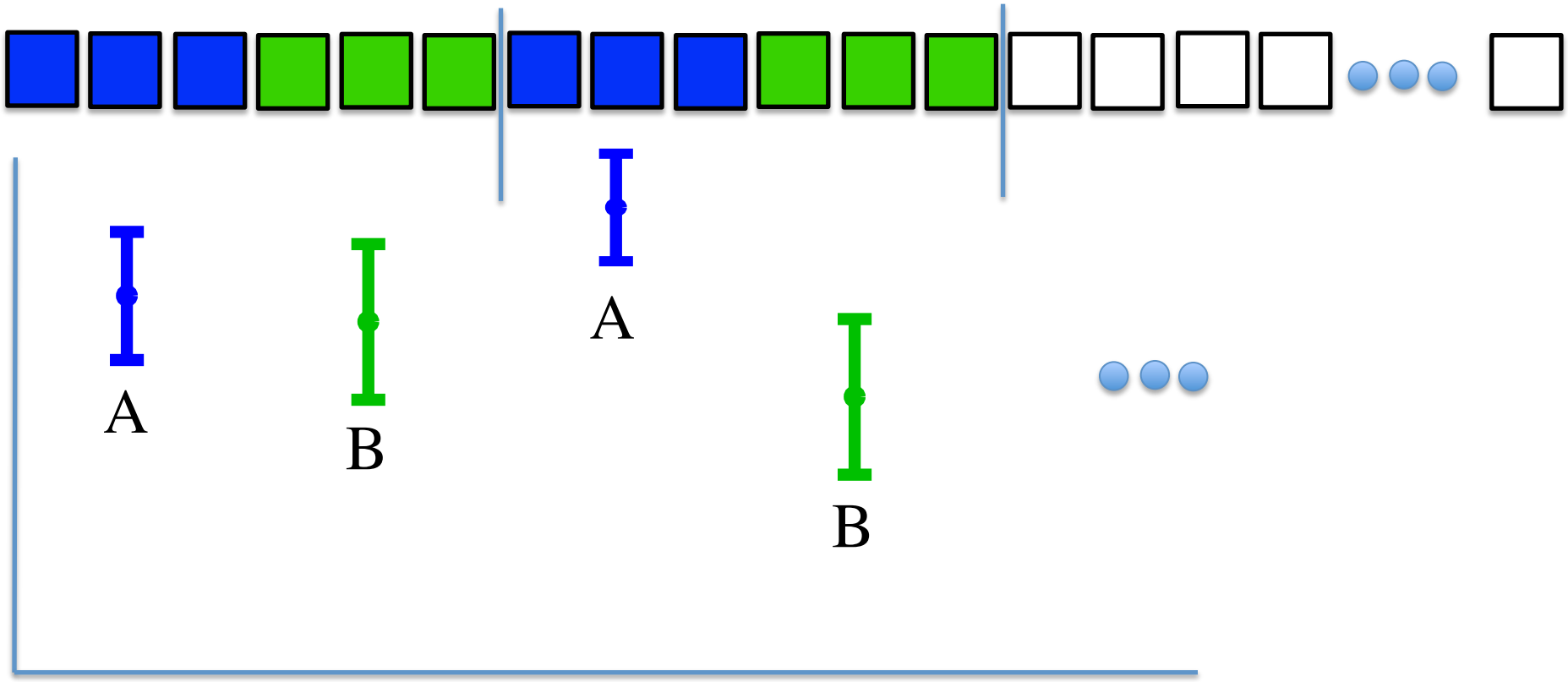
Methodology: Consecutive Trials



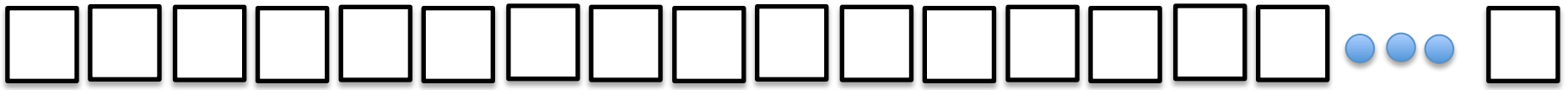
Methodology: Consecutive Trials



Methodology: Consecutive Trials

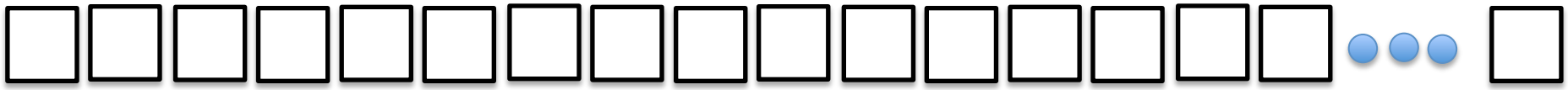


Methodology: Interleaved Trials



[Schad et al. 2010]

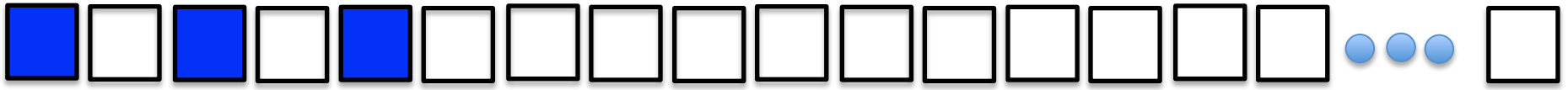
Methodology: Interleaved Trials

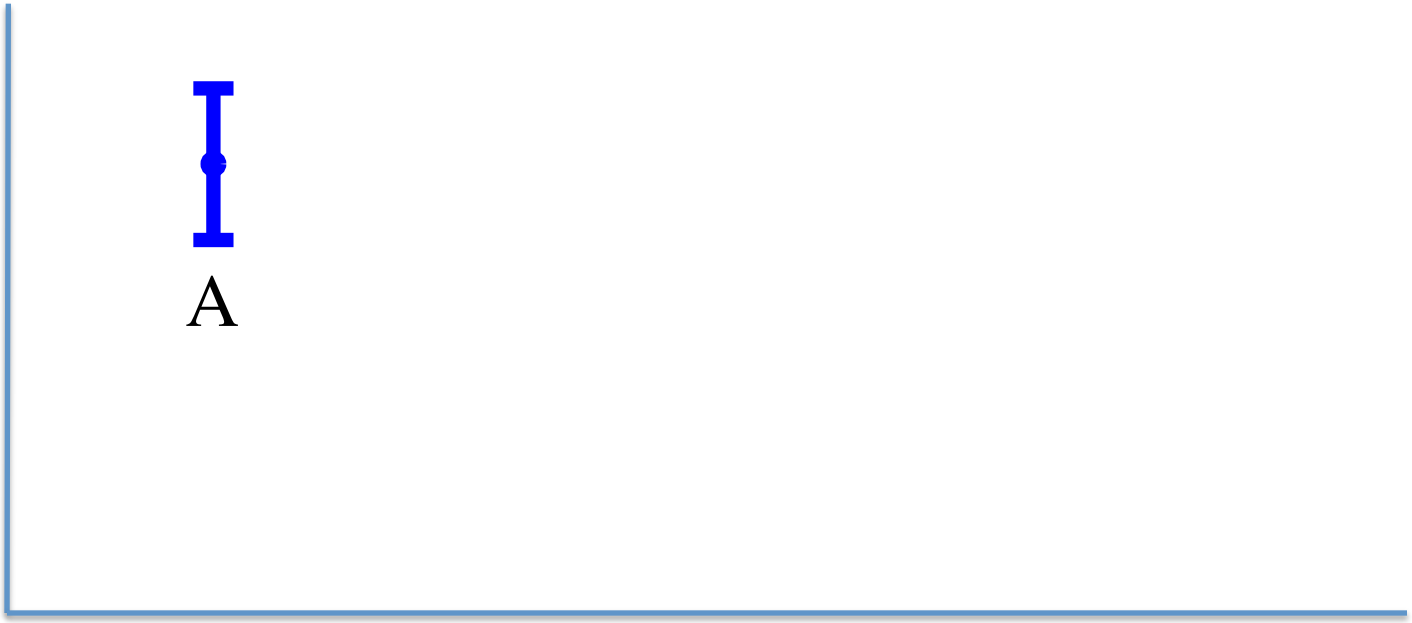


[Schad et al. 2010]

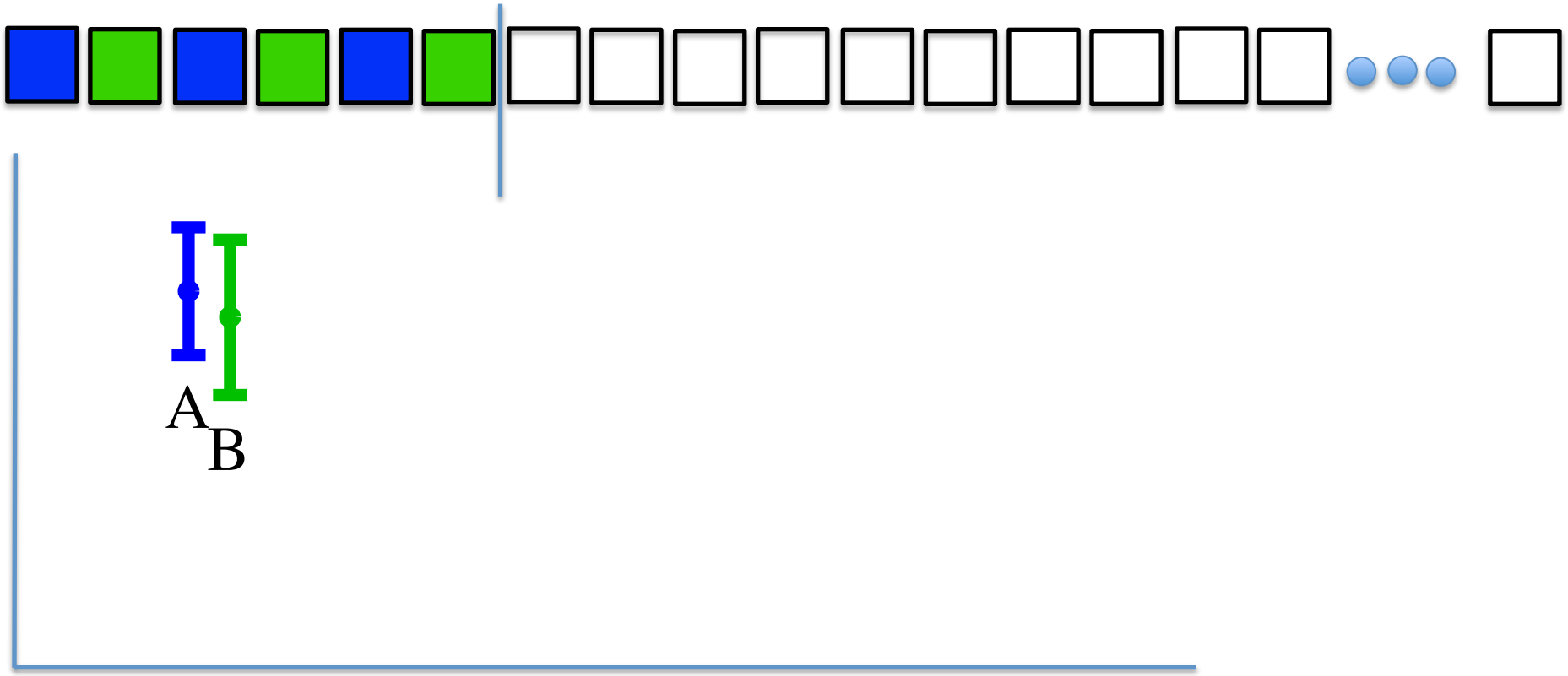
- Use exactly same data but label as different alternatives
- Consider interleaved ordering for conducting trials

Methodology: Interleaved Trials

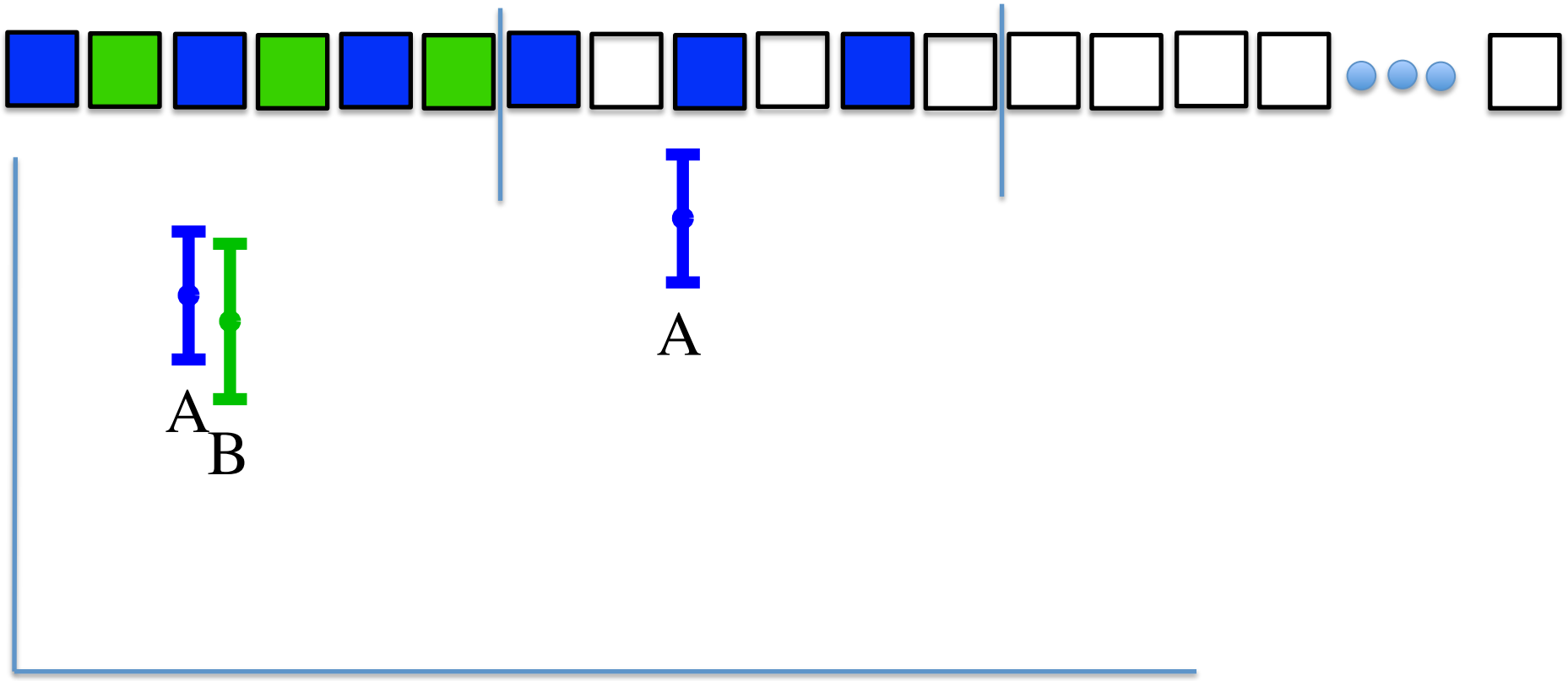




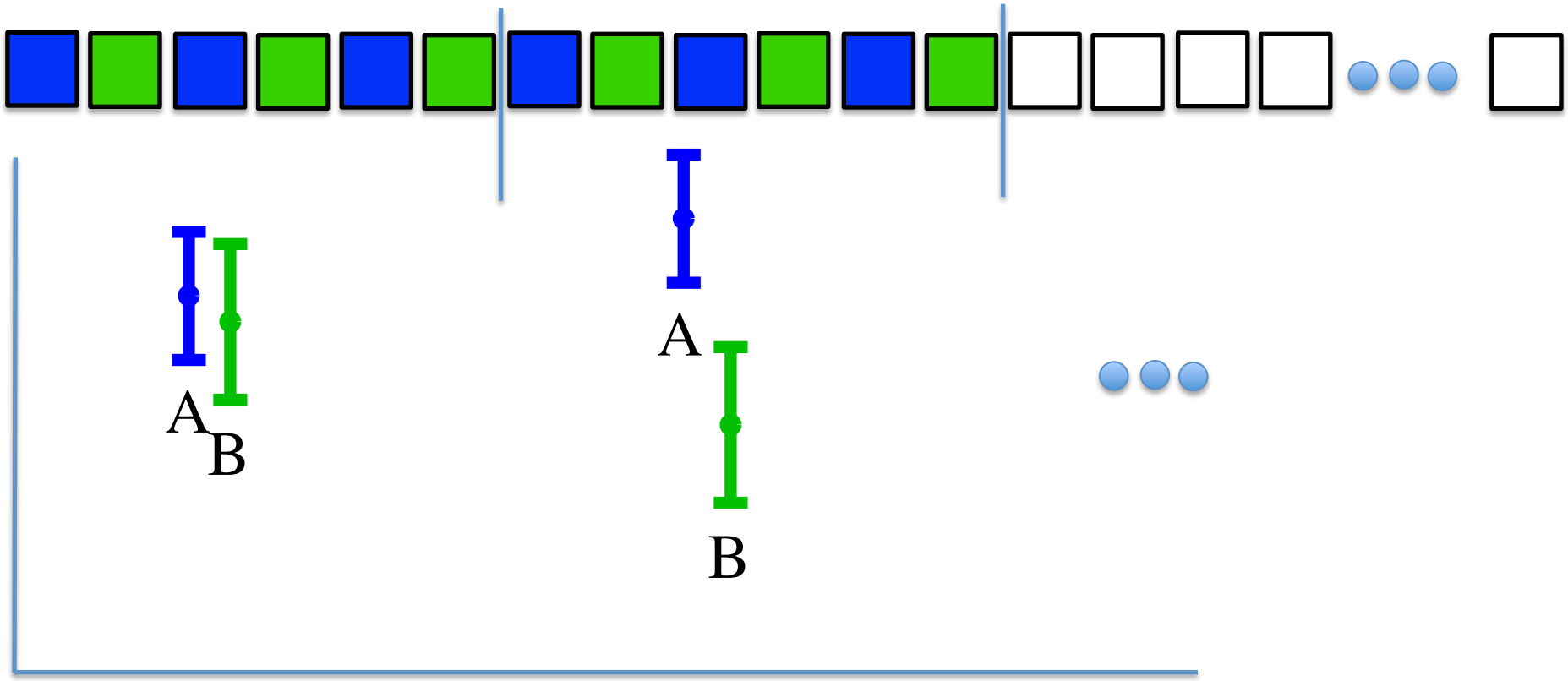
Methodology: Interleaved Trials



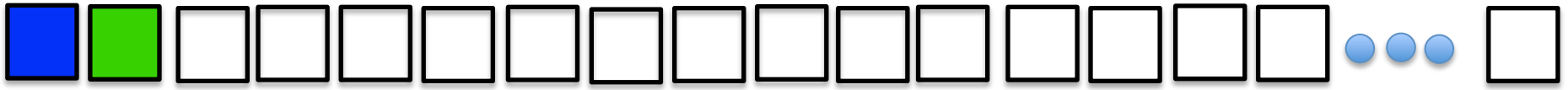
Methodology: Interleaved Trials

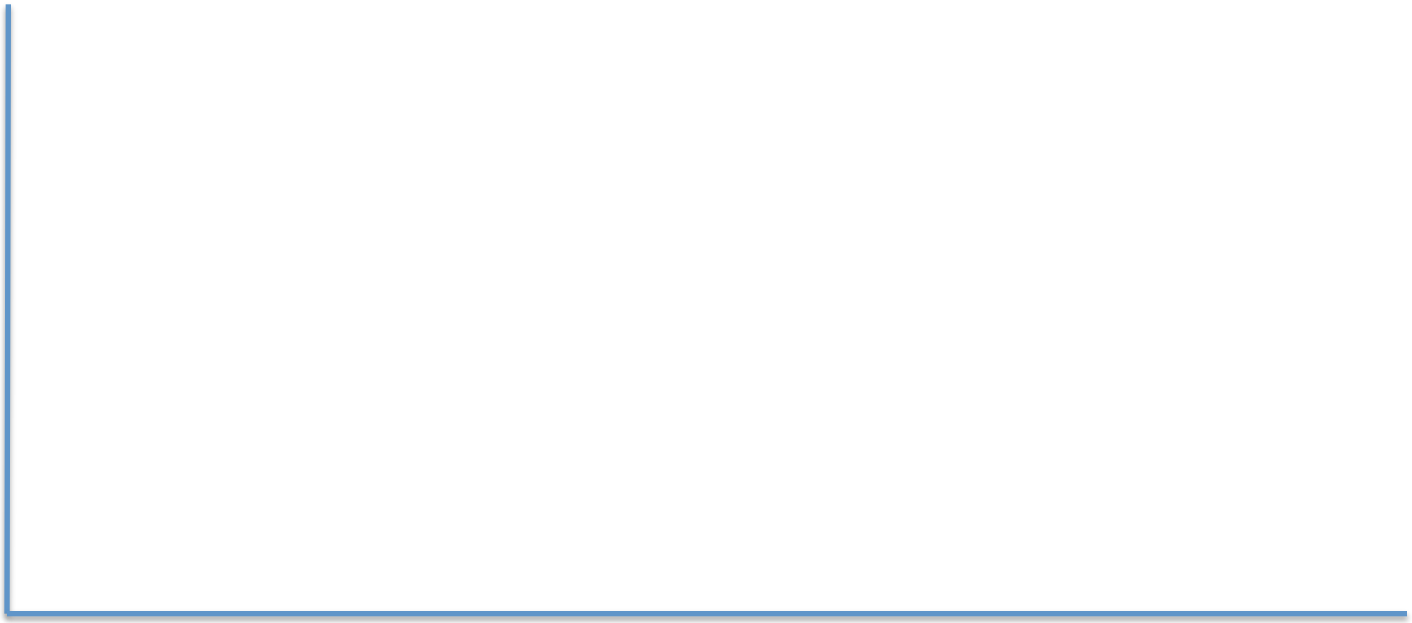


Methodology: Interleaved Trials

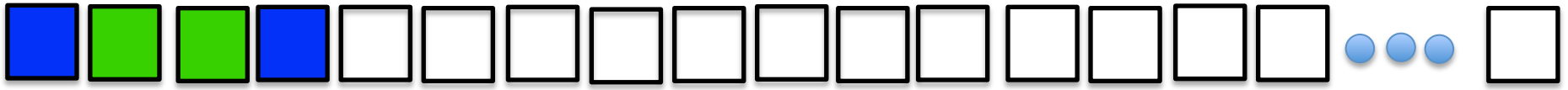


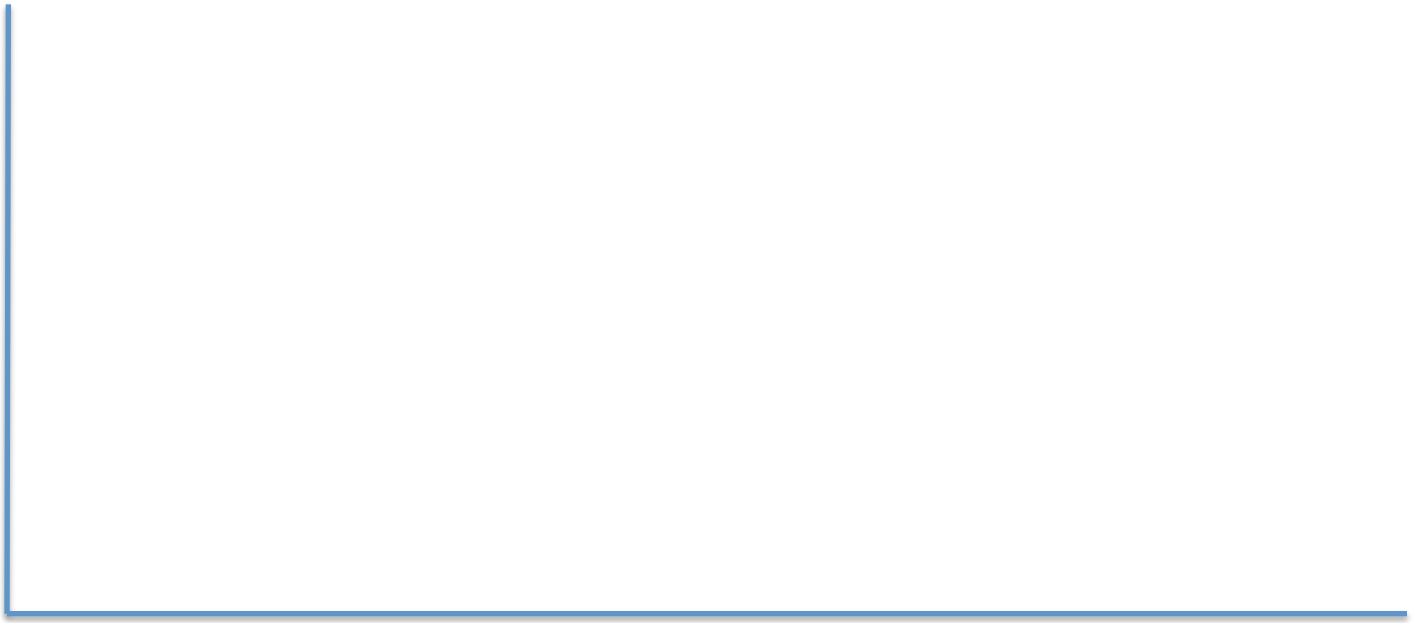
Methodology: Randomized, Interleaved



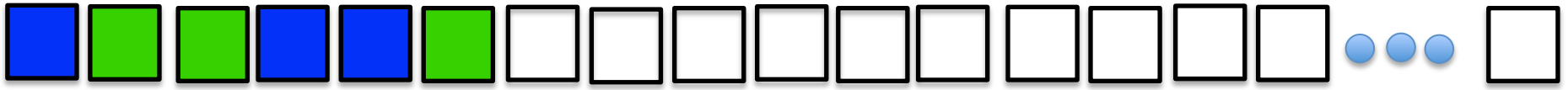


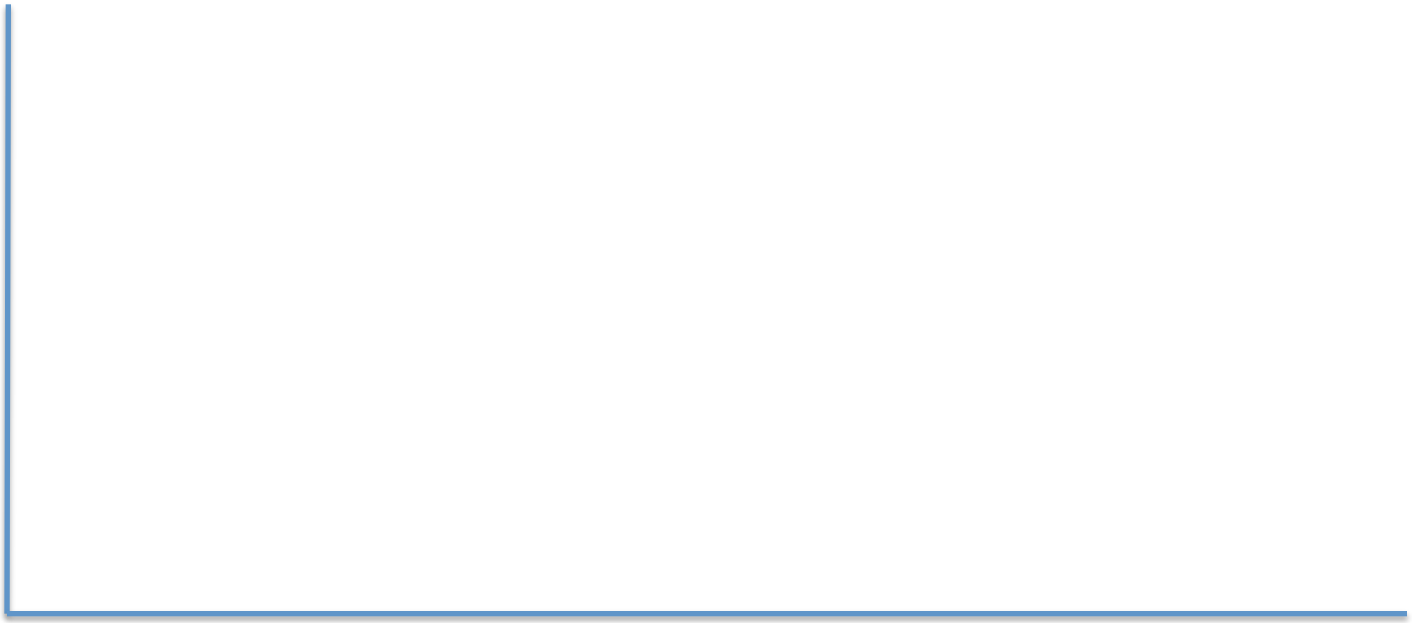
Methodology: Randomized, Interleaved





Methodology: Randomized, Interleaved

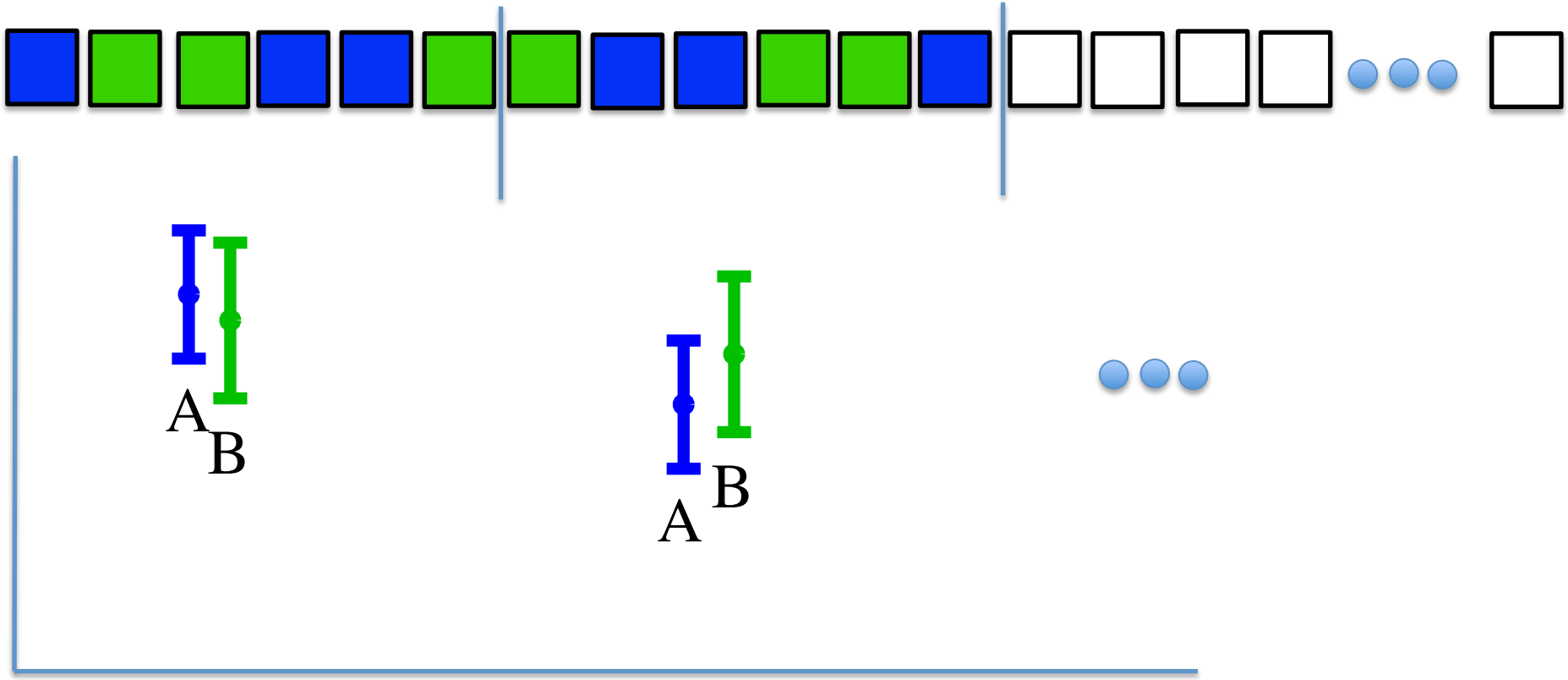




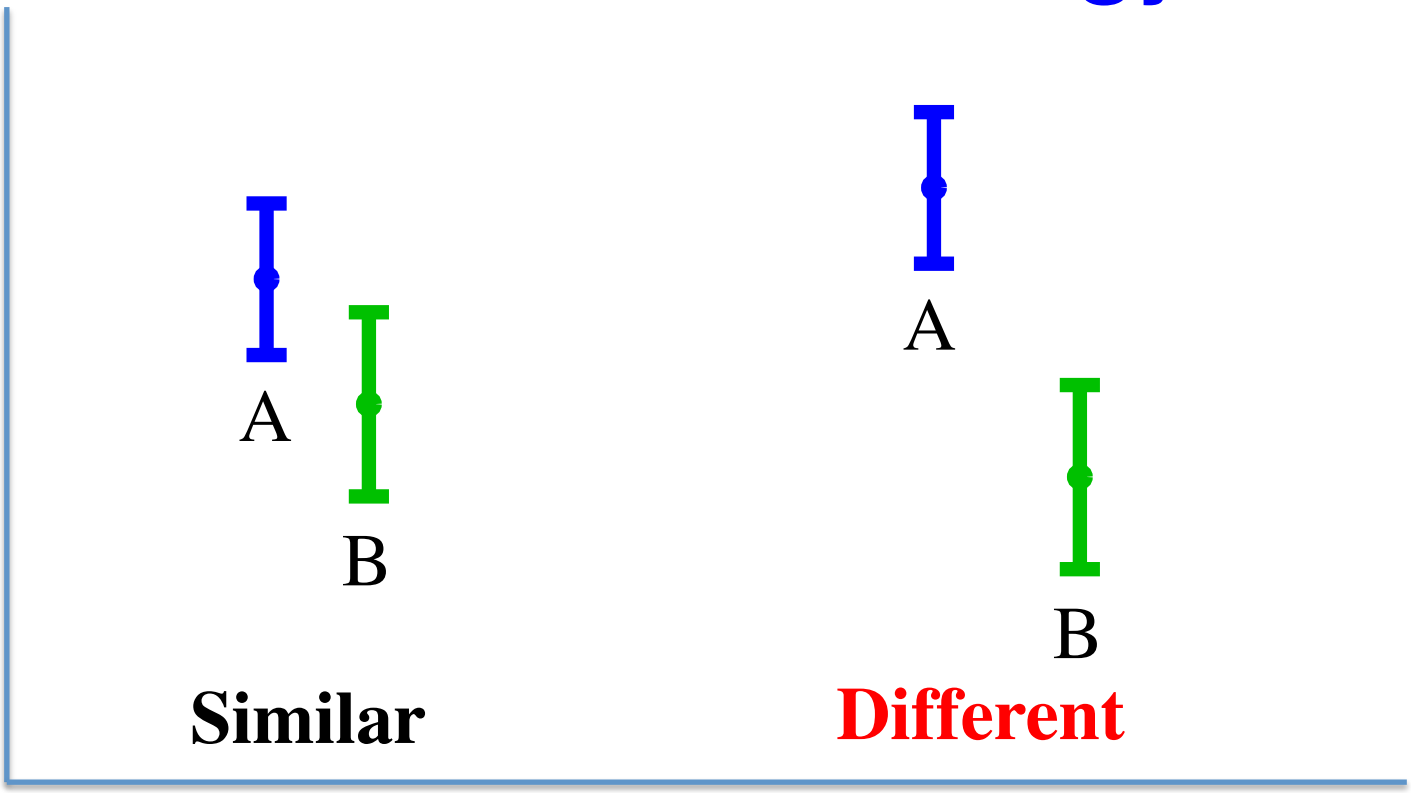
Methodology: Randomized, Interleaved

Methodology: Randomized, Interleaved

Methodology: Randomized, Interleaved

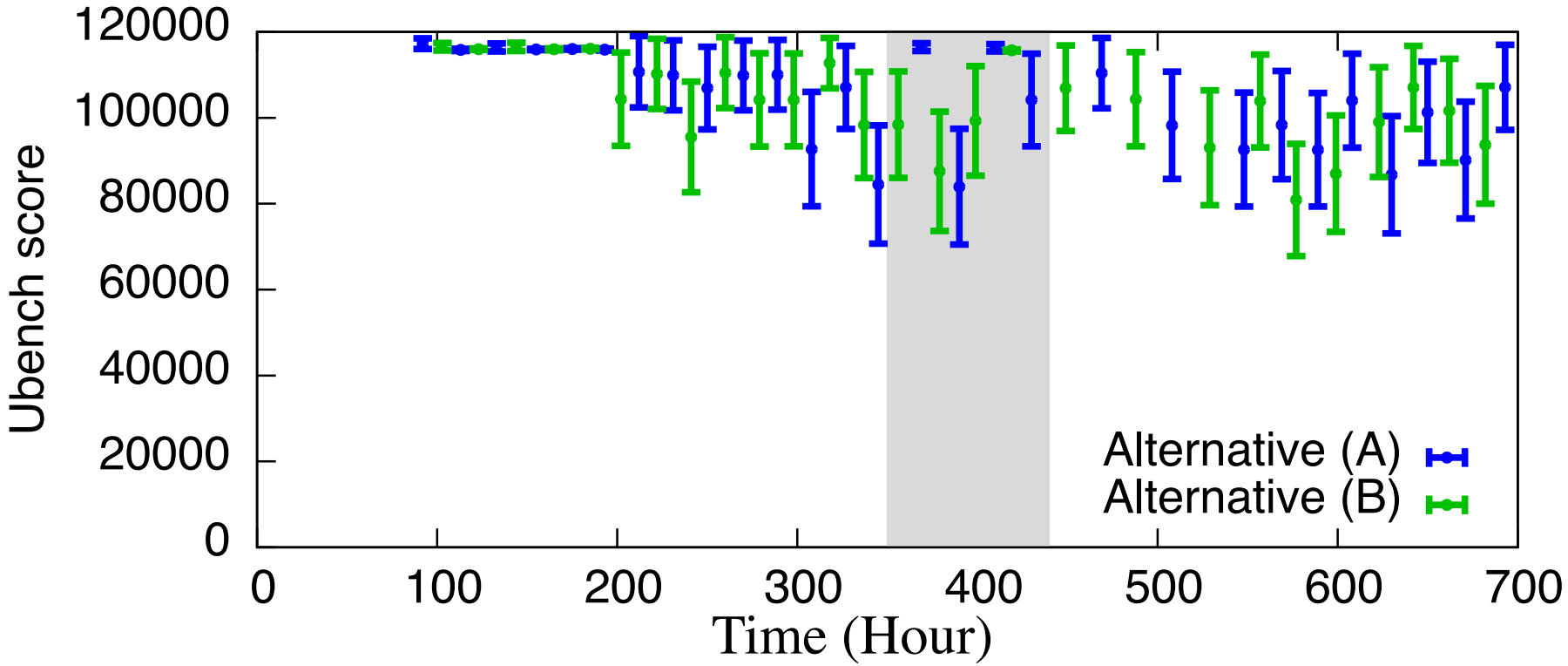


Methodology



- 20 trials of each alternative (95% Confidence Intervals)
  - All results SHOULD BE SIMILAR
  - If not, approach to running experiments is flawed

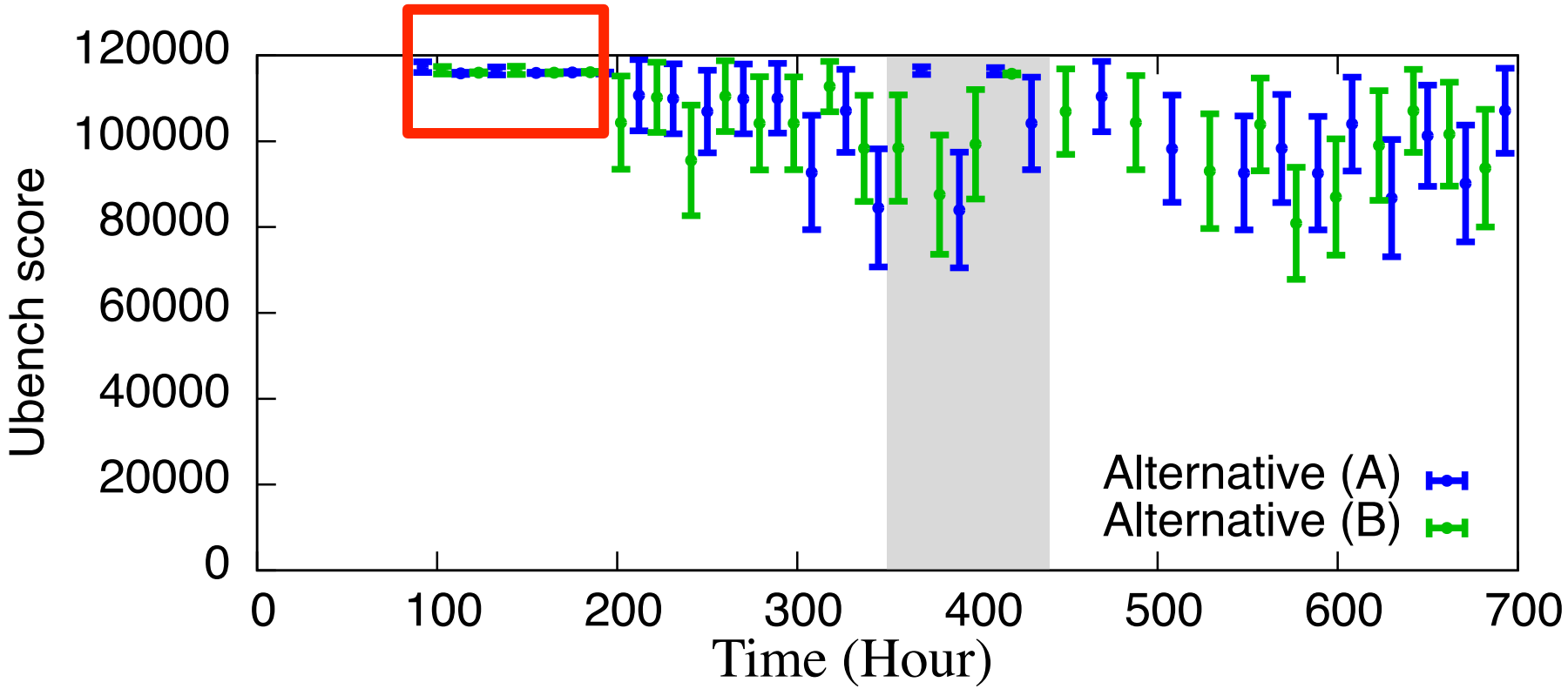
Multiple Consecutive Trials



Multiple Consecutive Trials (MCT)

EC2: CPU performance, US, small instance

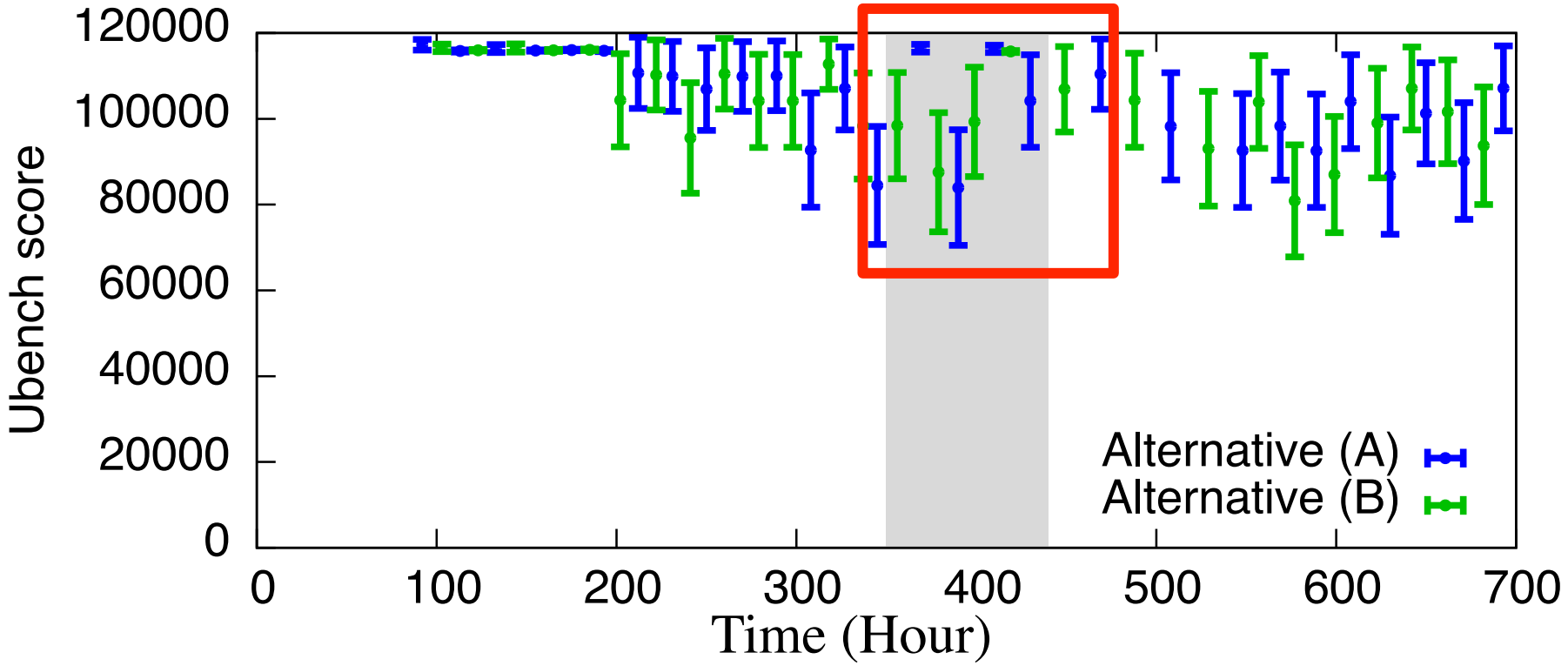
Multiple Consecutive Trials



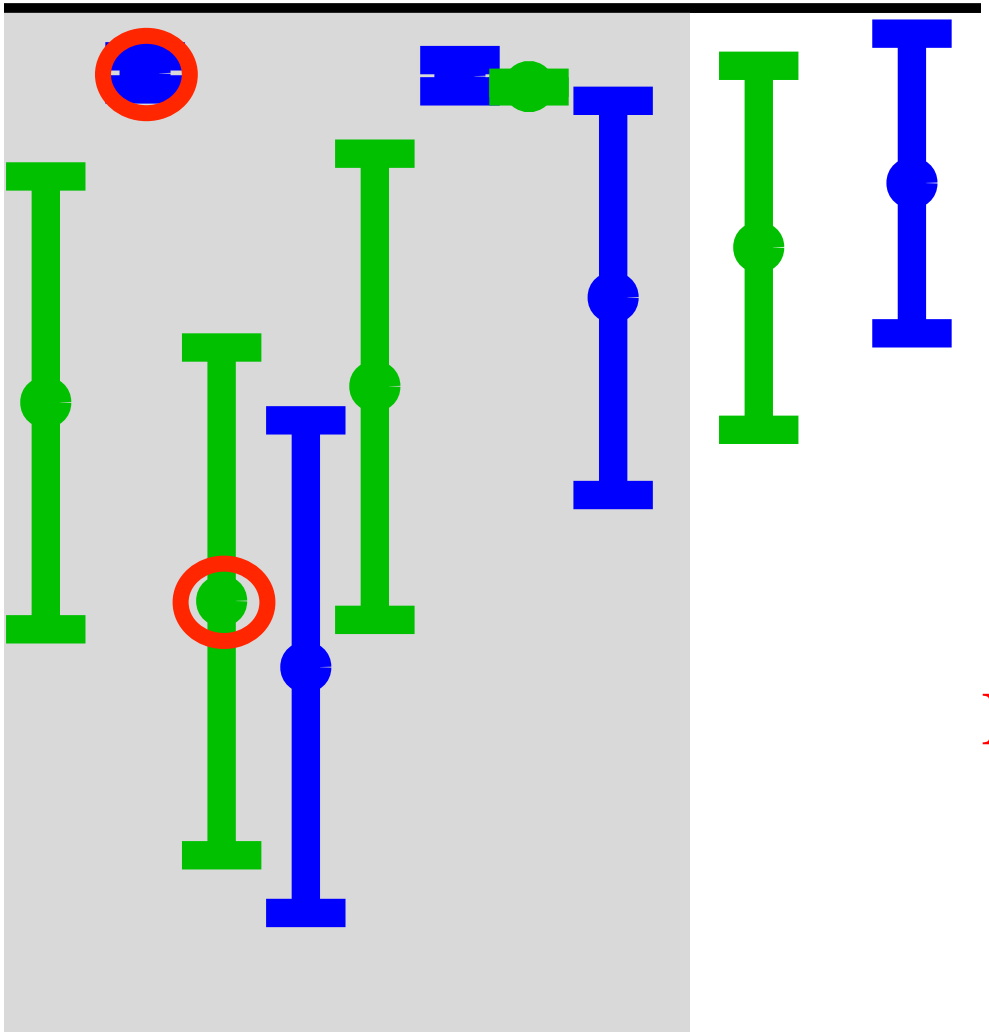
Multiple Consecutive Trials (MCT)

EC2: CPU performance, US, small instance

Multiple Consecutive Trials

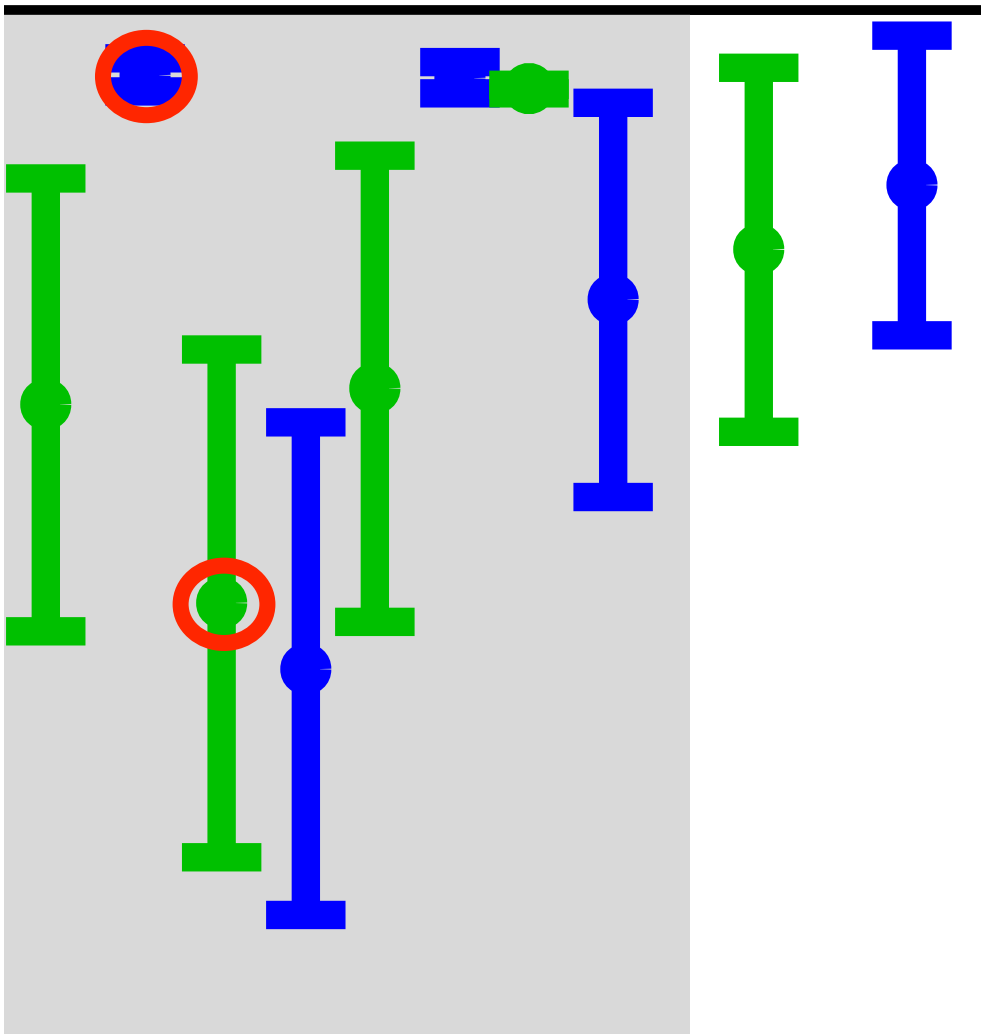


Multiple Consecutive Trials



Flawed

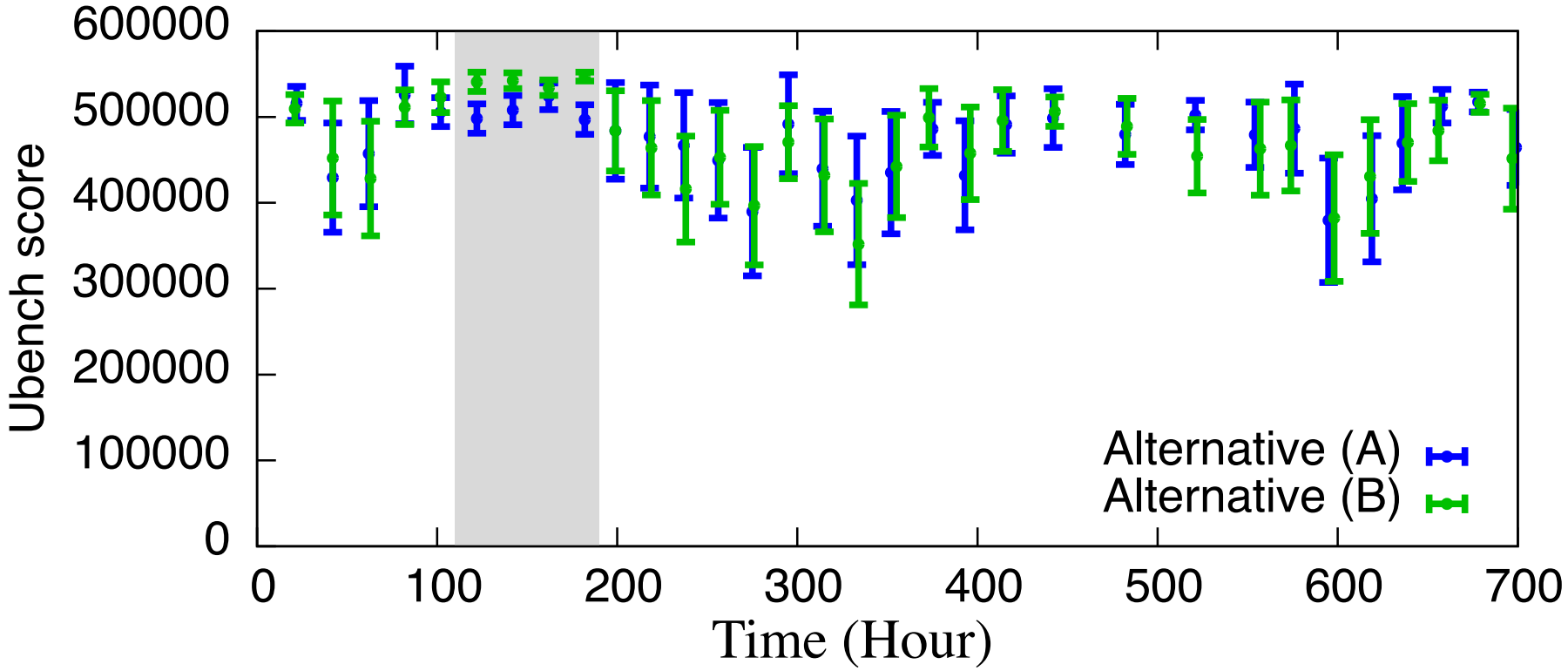
Multiple Consecutive Trials



Flawed

CI's provide false sense of validity and rigour

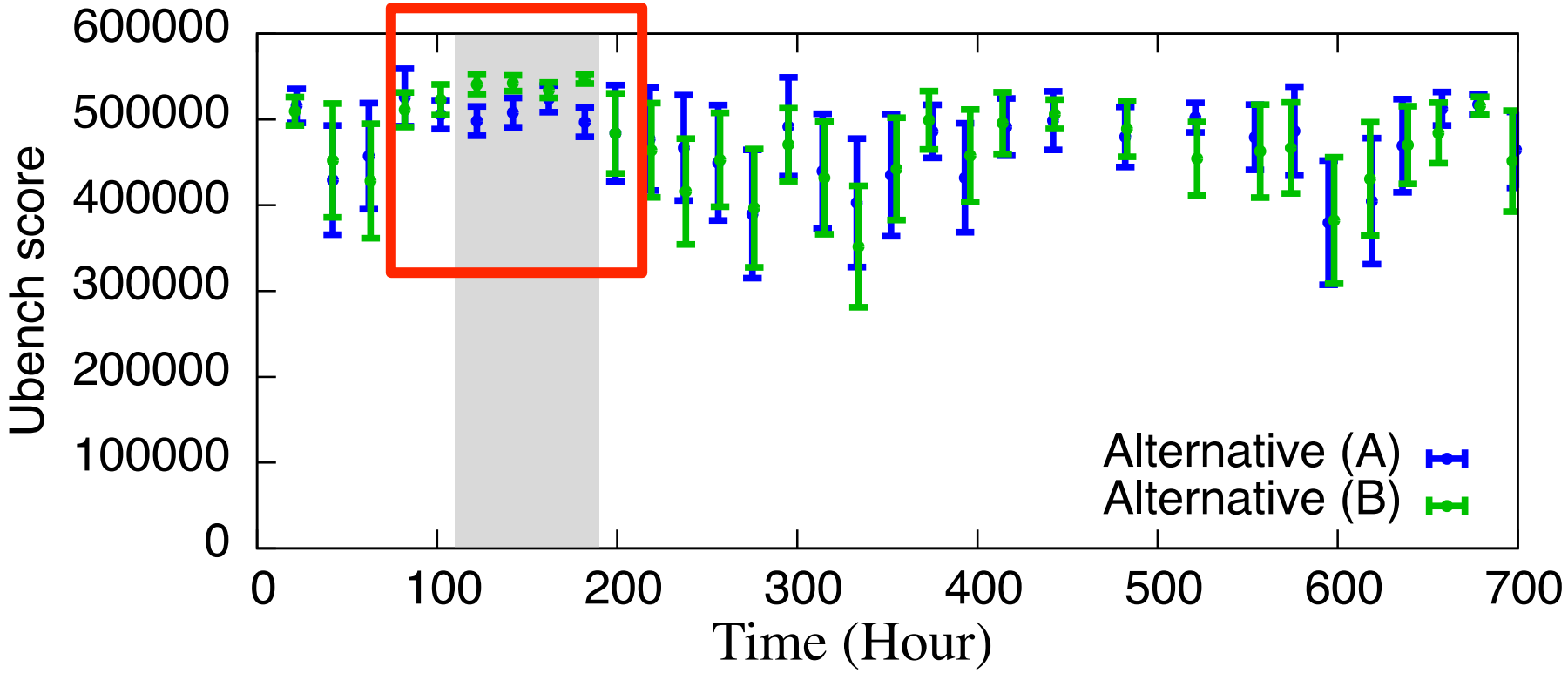
Multiple Interleaved Trials



Multiple Interleaved Trials (MIT)

EC2: CPU performance, US, large instance

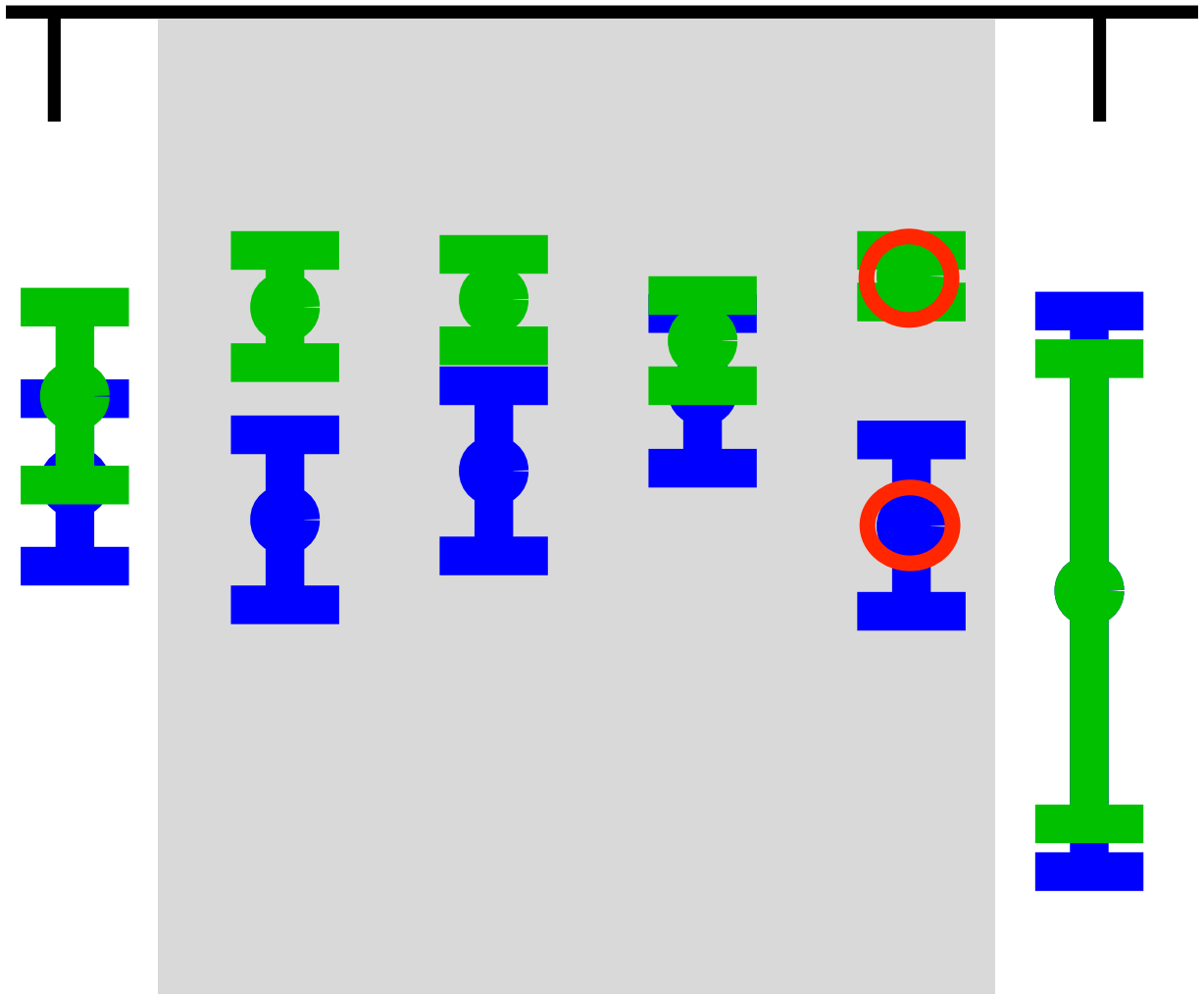
Multiple Interleaved Trials



Multiple Interleaved Trials (MIT)

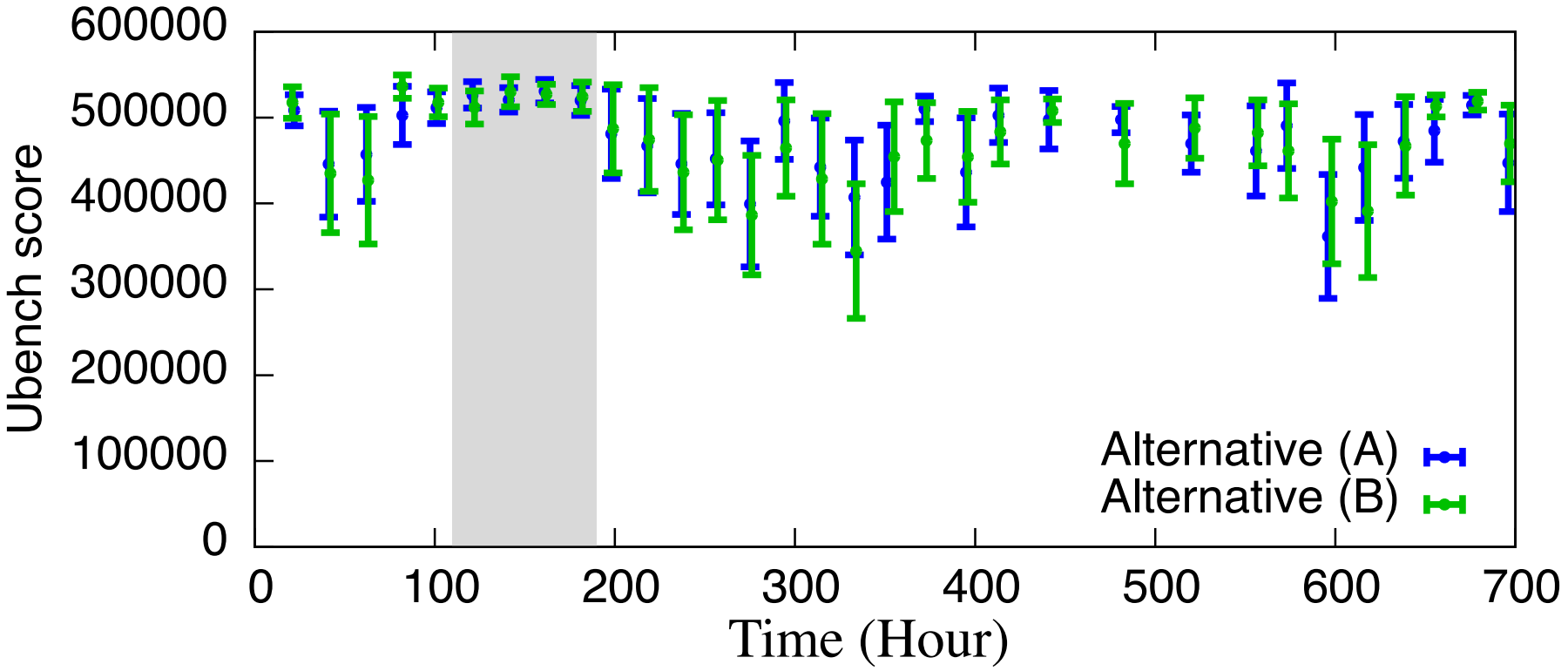
EC2: CPU performance, US, large instance

Multiple Interleaved Trials



Flawed

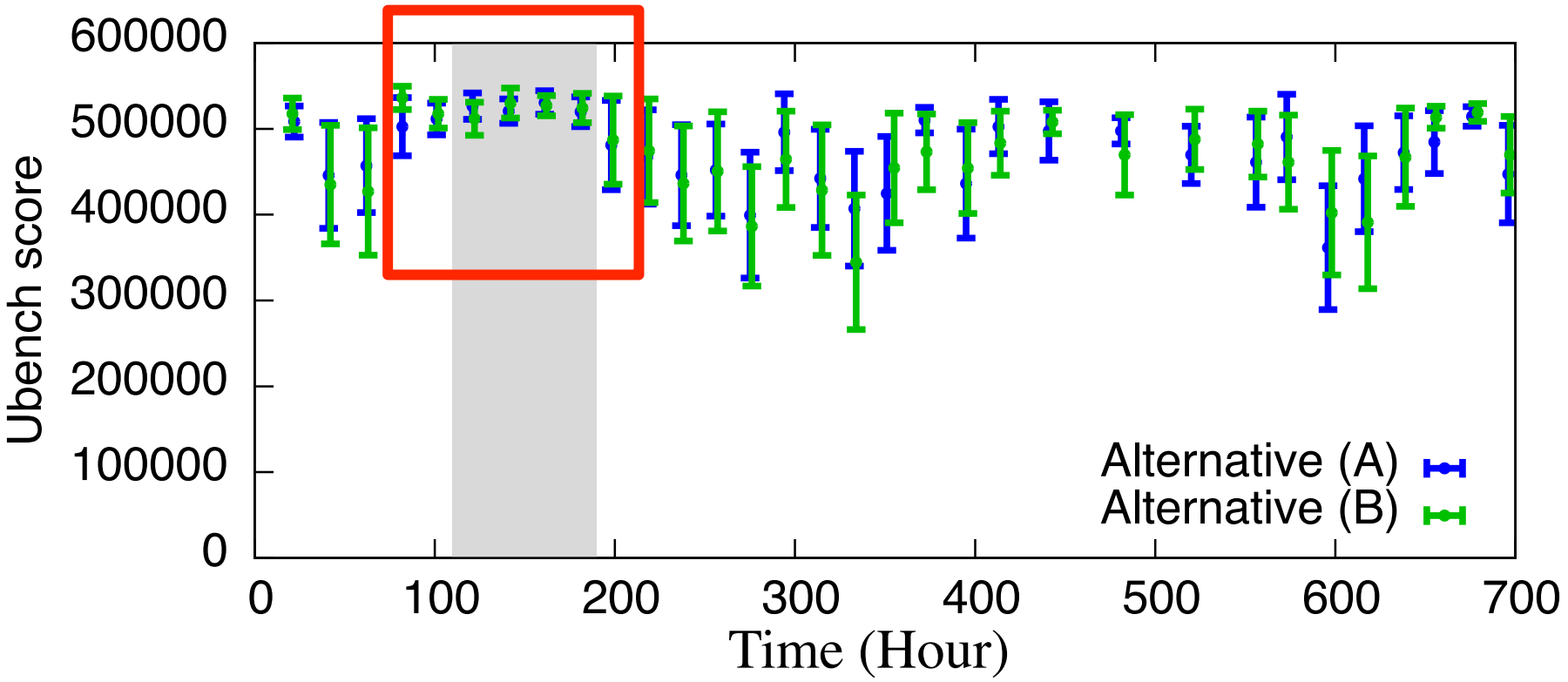
Randomized Multiple Interleaved Trials



Randomized Multiple Interleaved Trials (RMIT)

EC2: CPU performance, US, large instance

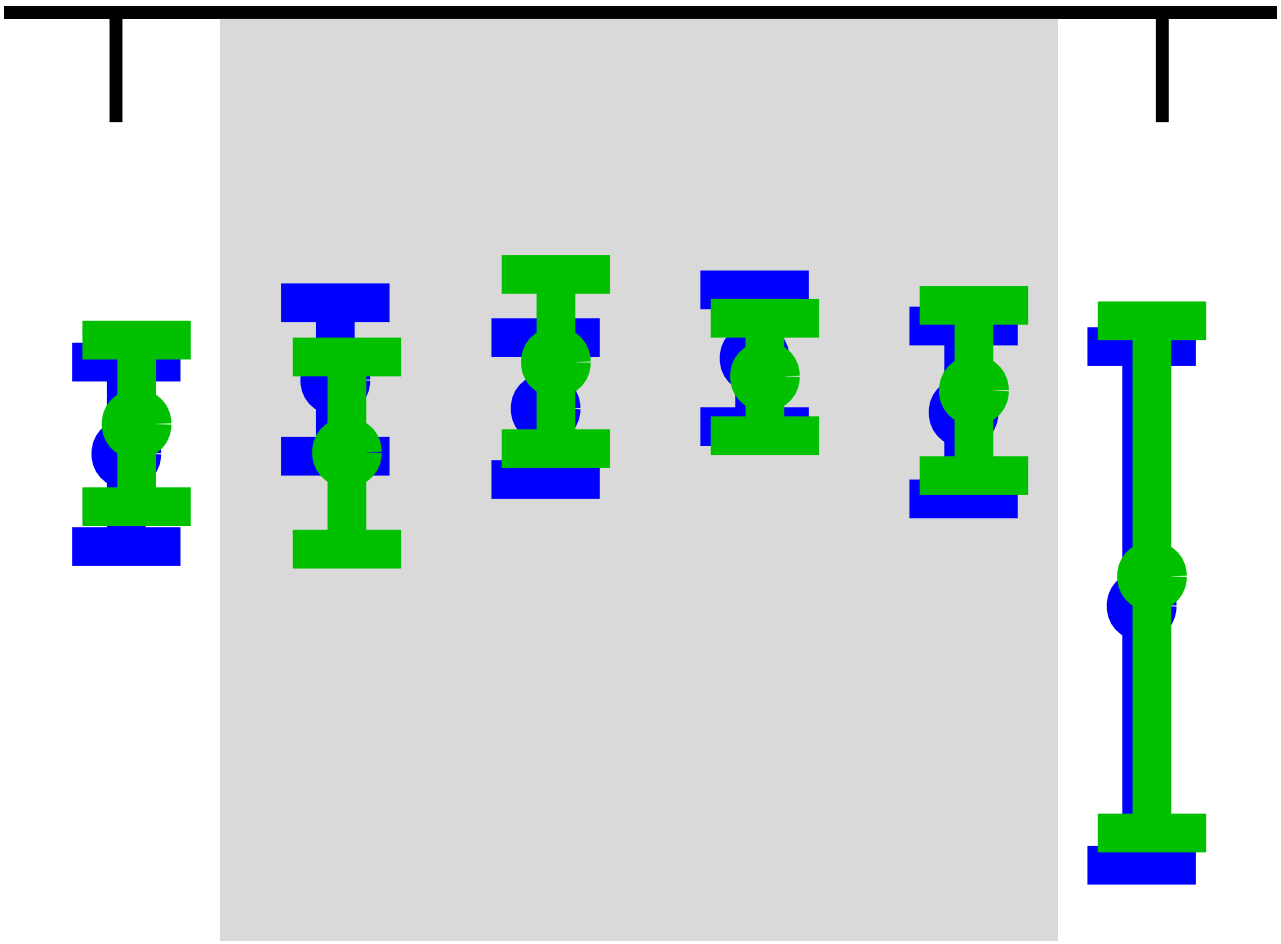
Randomized Multiple Interleaved Trials



Randomized Multiple Interleaved Trials (RMIT)

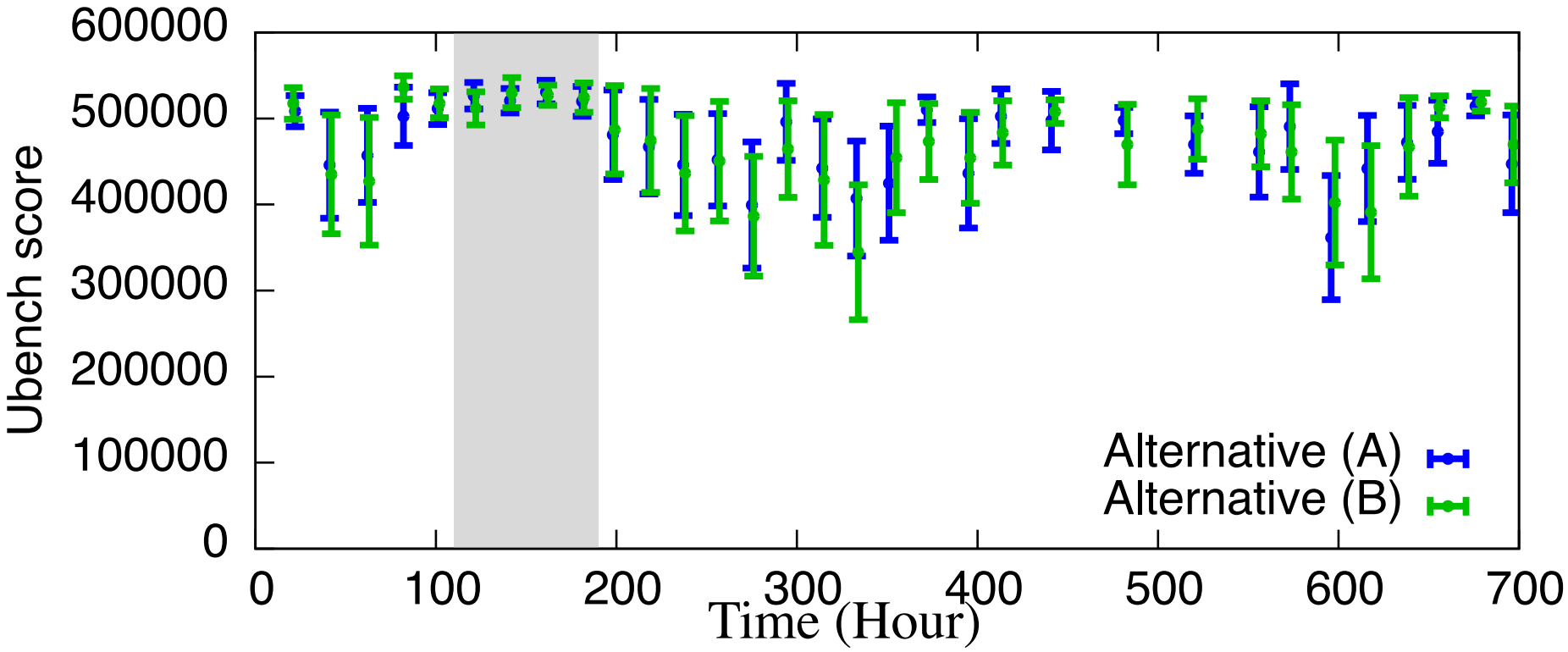
EC2: CPU performance, US, large instance

Randomized Multiple Interleaved Trials



All Are Now Similar!

Randomized Multiple Interleaved Trials



All Are Now Similar!

RMIT approach is sound

University of Waterloo

Conclusions

- When comparing competing alternatives
- Use Randomized Multiple Interleaved Trials (**RMIT**)
  - Or risk invalid comparisons and incorrect conclusions

Because:

- Performance in clouds is highly variable
- Environment changes
- Not possible to know about or detect changes!

Conclusions

- When comparing competing alternatives
- Use Randomized Multiple Interleaved Trials (**RMIT**)
  - Or risk invalid comparisons and incorrect conclusions

Because:

- Performance in clouds is highly variable
- Environment changes
- Not possible to know about or detect changes!

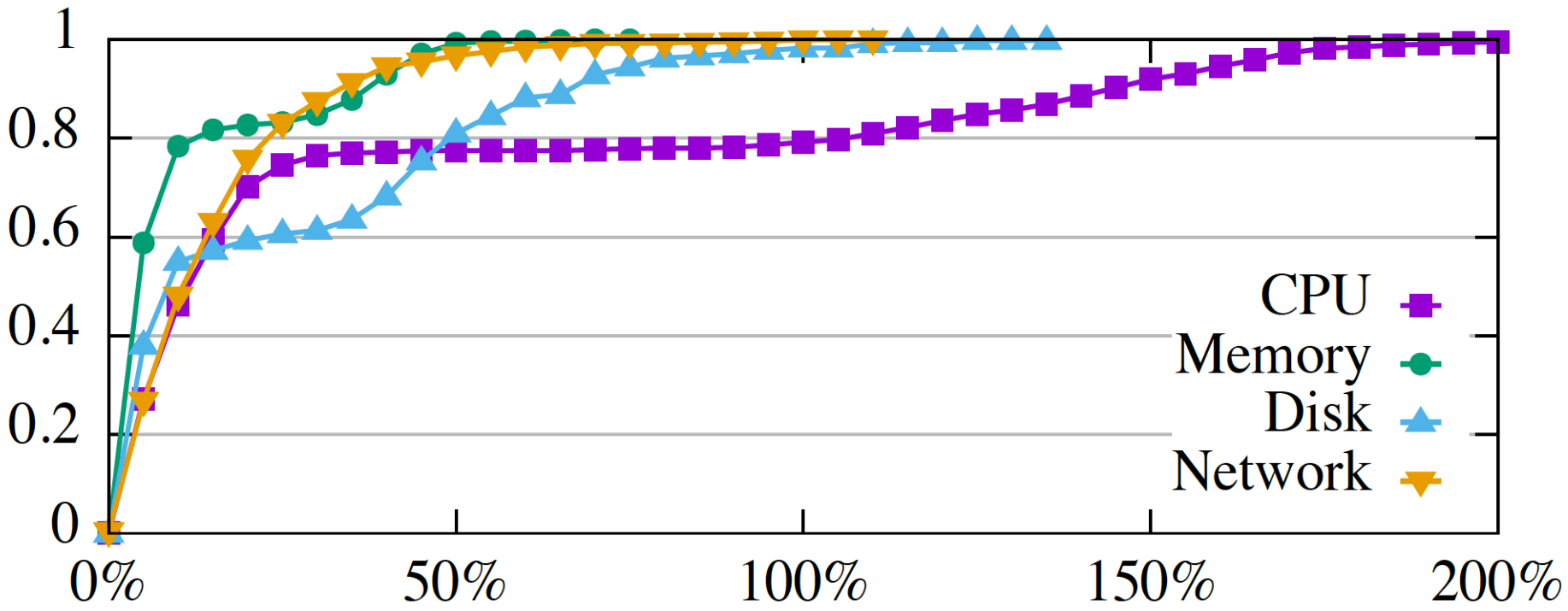
Questions?

Data used is available: URL is in the paper

University of Waterloo

The End

CDF

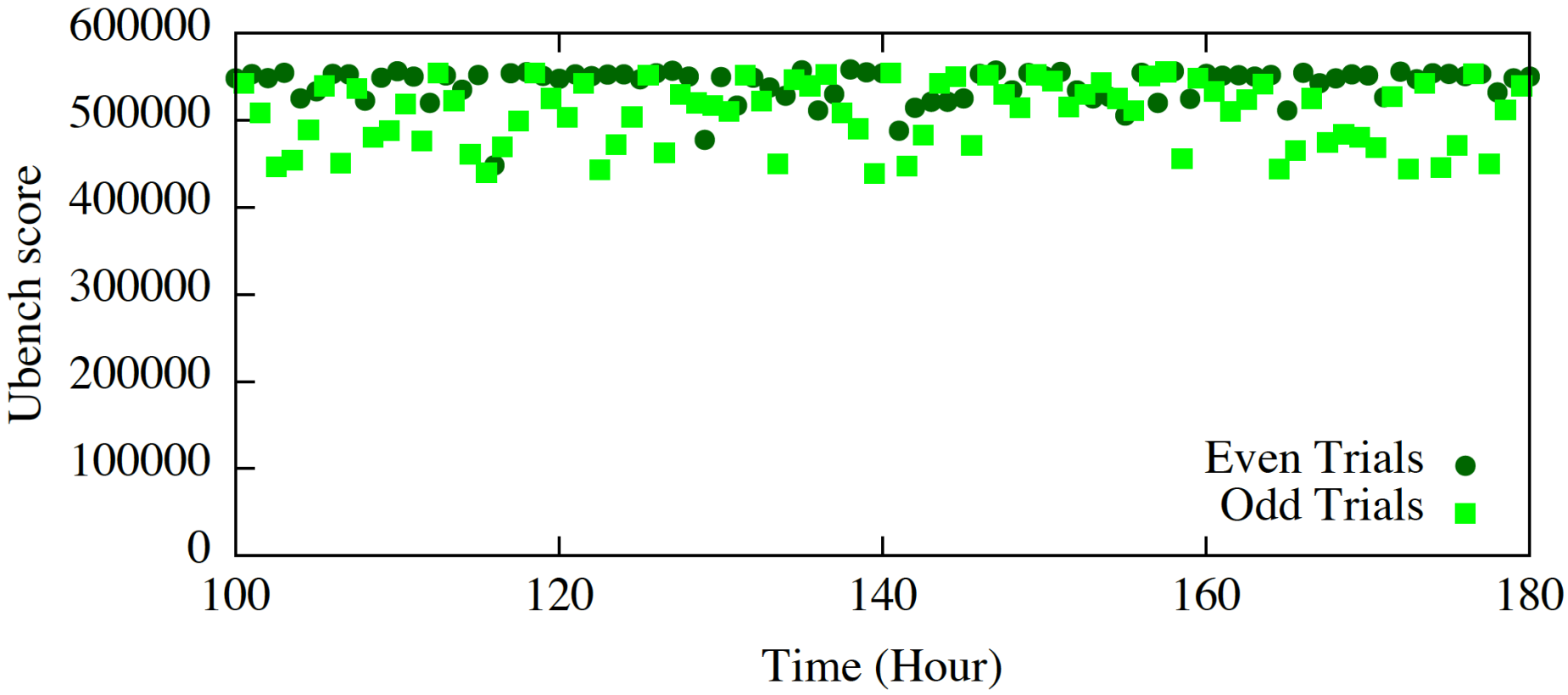


Percentage difference between consecutive trials

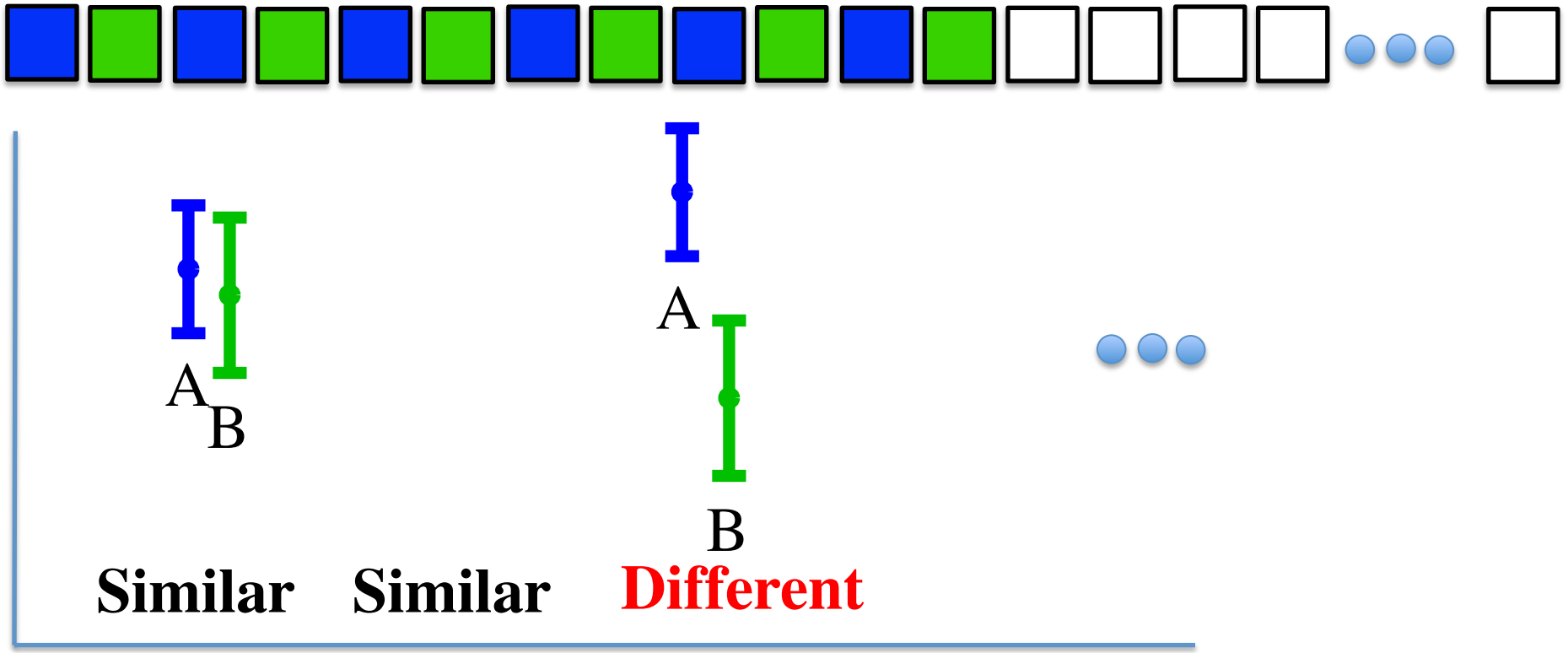
Efficacy of MCT

Benchmark	Instance	Region	RFE	Max Diff
CPU	Small	US	10.2	37.8
		EU	9.8	27.3
	Large	US	7.9	30.1
		EU	14.9	36.5
Memory	Small	US	9.8	15.1
		EU	6.7	8.8
	Large	US	10.2	16.8
		EU	7.0	11.1
Disk I/O	Small	US	4.0	15.3
		EU	0.0	NA
	Large	US	5.9	23.8
		EU	0.0	NA
Network	Mixed	US	6.1	19.6
		EU	5.1	12.5

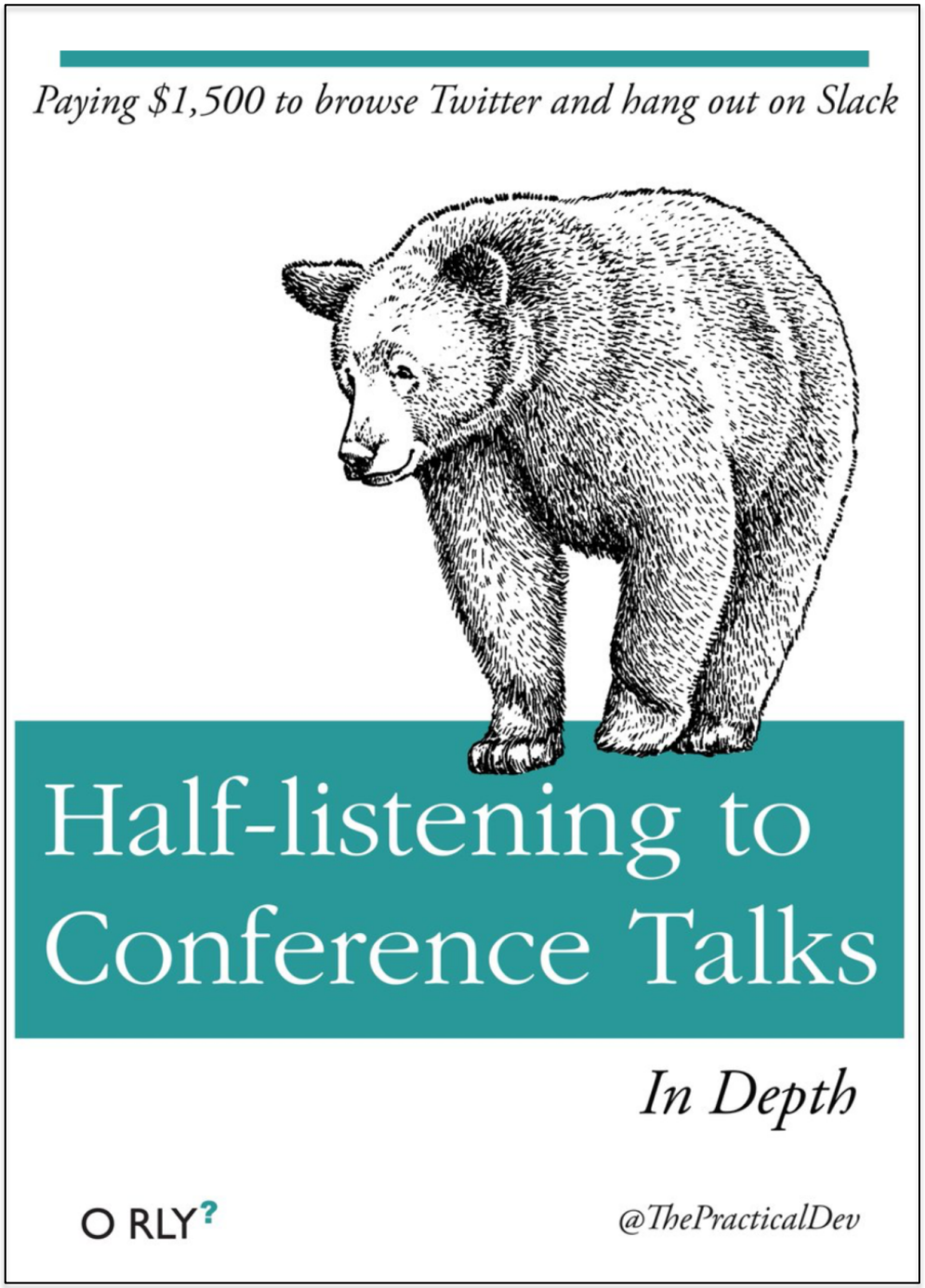
CPU Performance Large Instance, US



Methodology: Interleaved Trials



- 20 trials of each alternative (95% Confidence Intervals)
  - All results SHOULD BE SIMILAR
  - If not, approach to running experiments is flawed

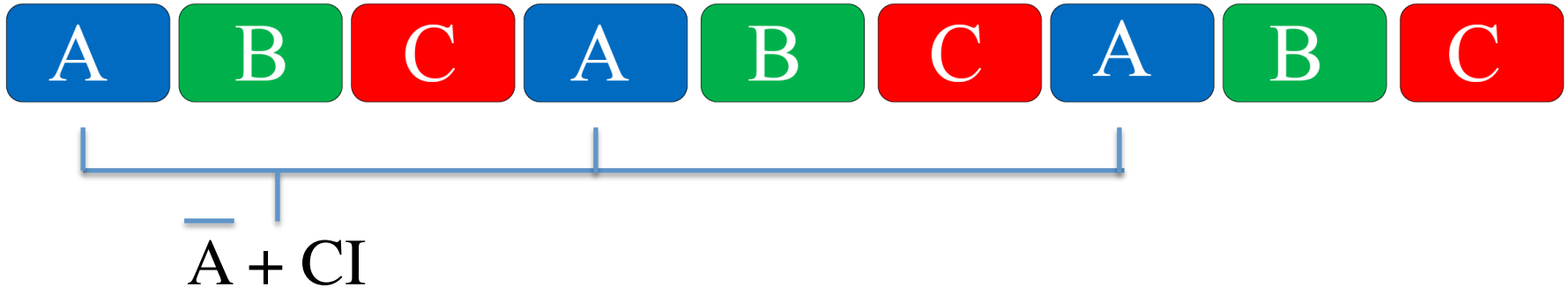

Paying $1,500 to browse Twitter and hang out on Slack
Half-listening to Conference Talks
In Depth
O RLY?
@ThePracticalDev

Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

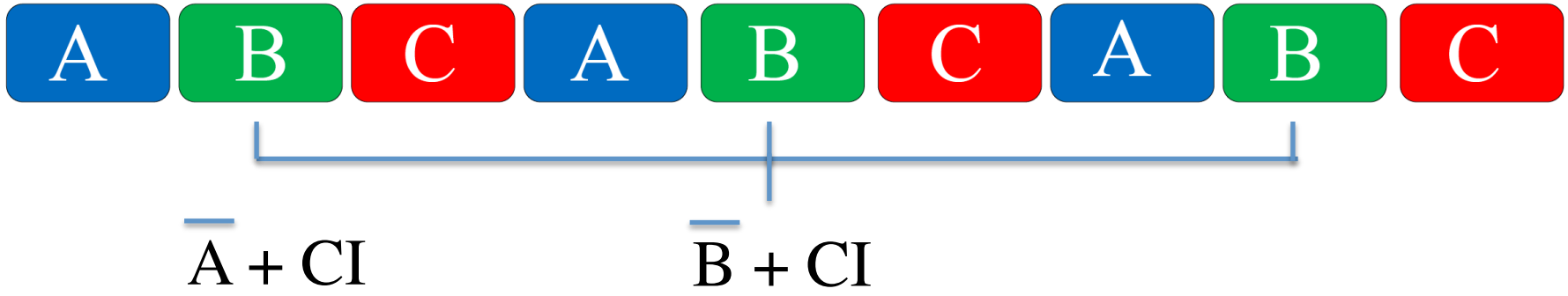


Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

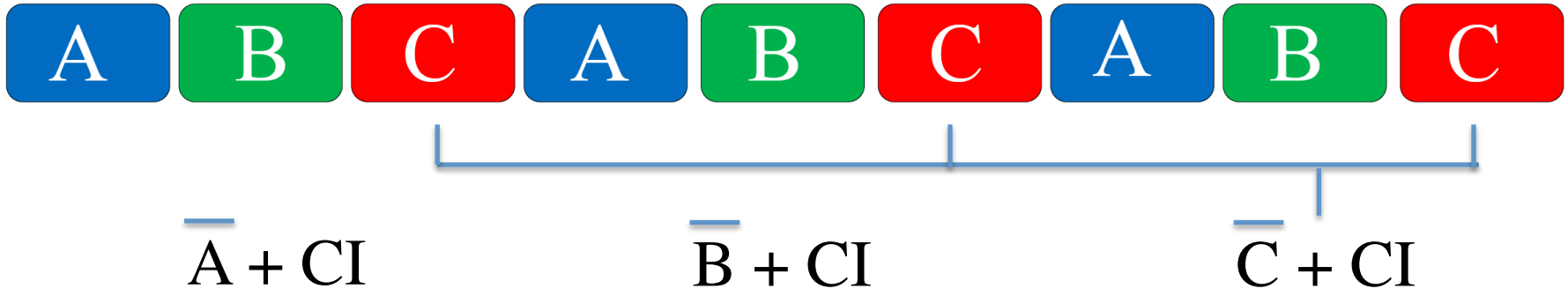


Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

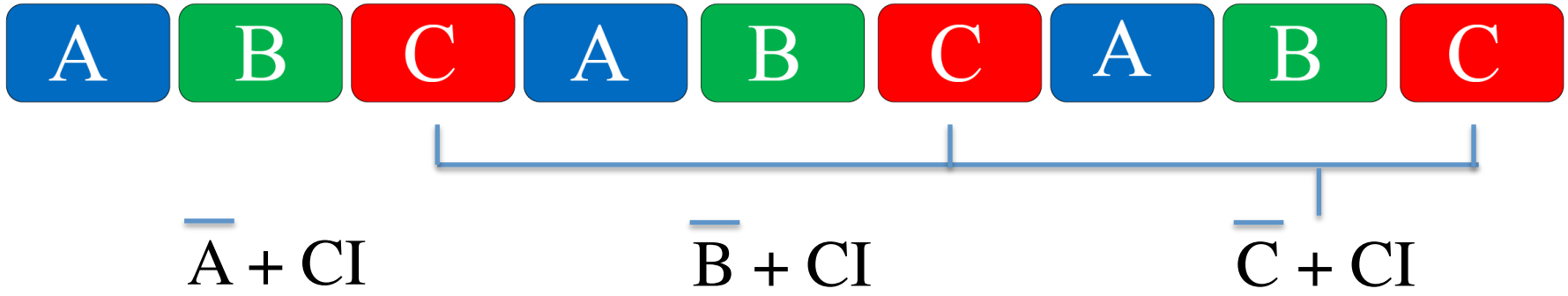


Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)



Potential issue: Periodic changes in cloud environment

Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

Randomized Multiple Interleaved Trials (RMIT)

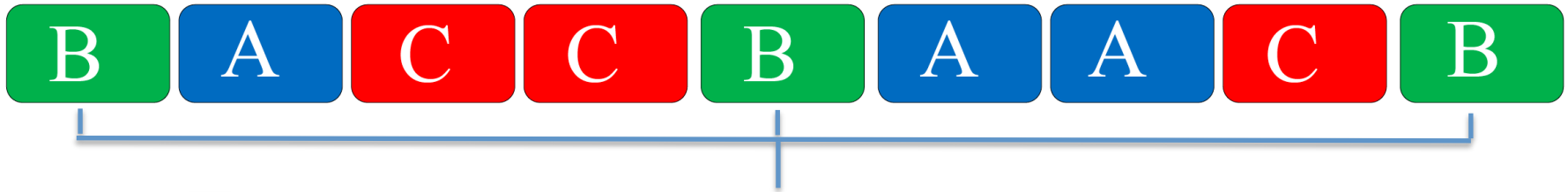
Proposed Approaches

Single Trial

Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

Randomized Multiple Interleaved Trials (RMIT)



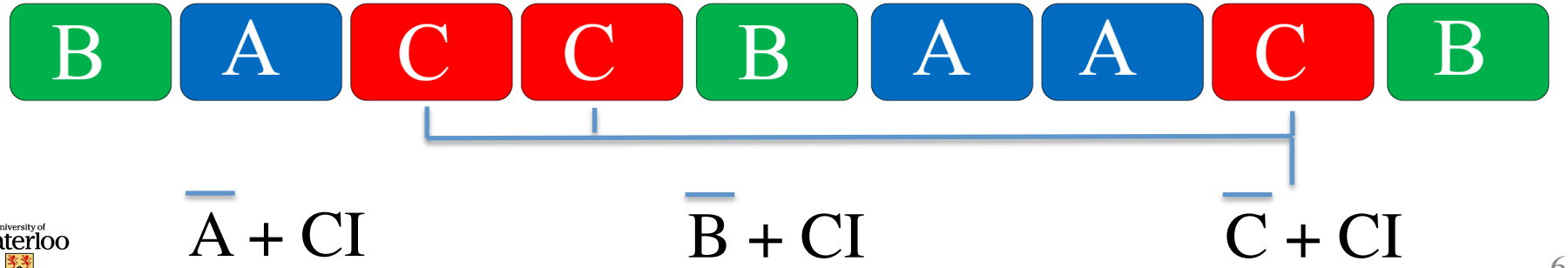
Proposed Approaches

Single Trial

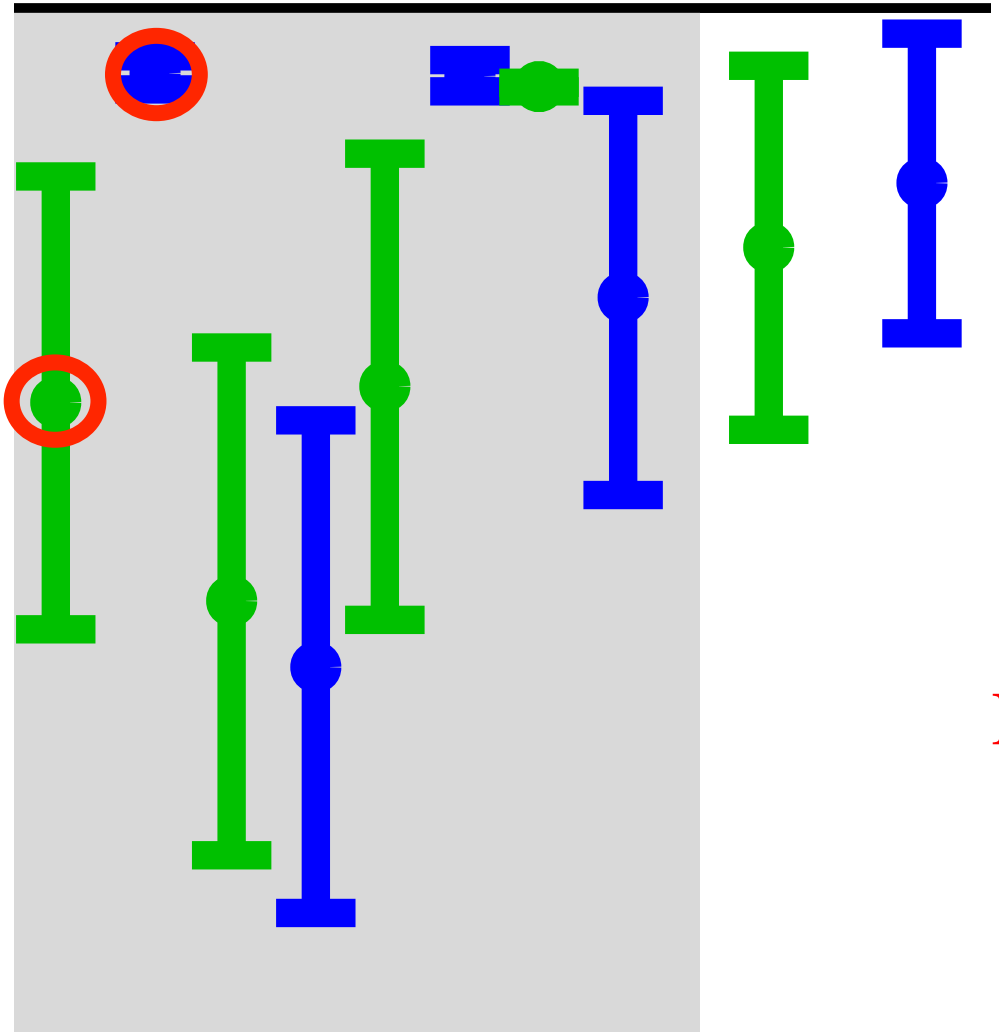
Multiple Consecutive Trials (MCT)

Multiple Interleaved Trials (MIT)

Randomized Multiple Interleaved Trials (RMIT)

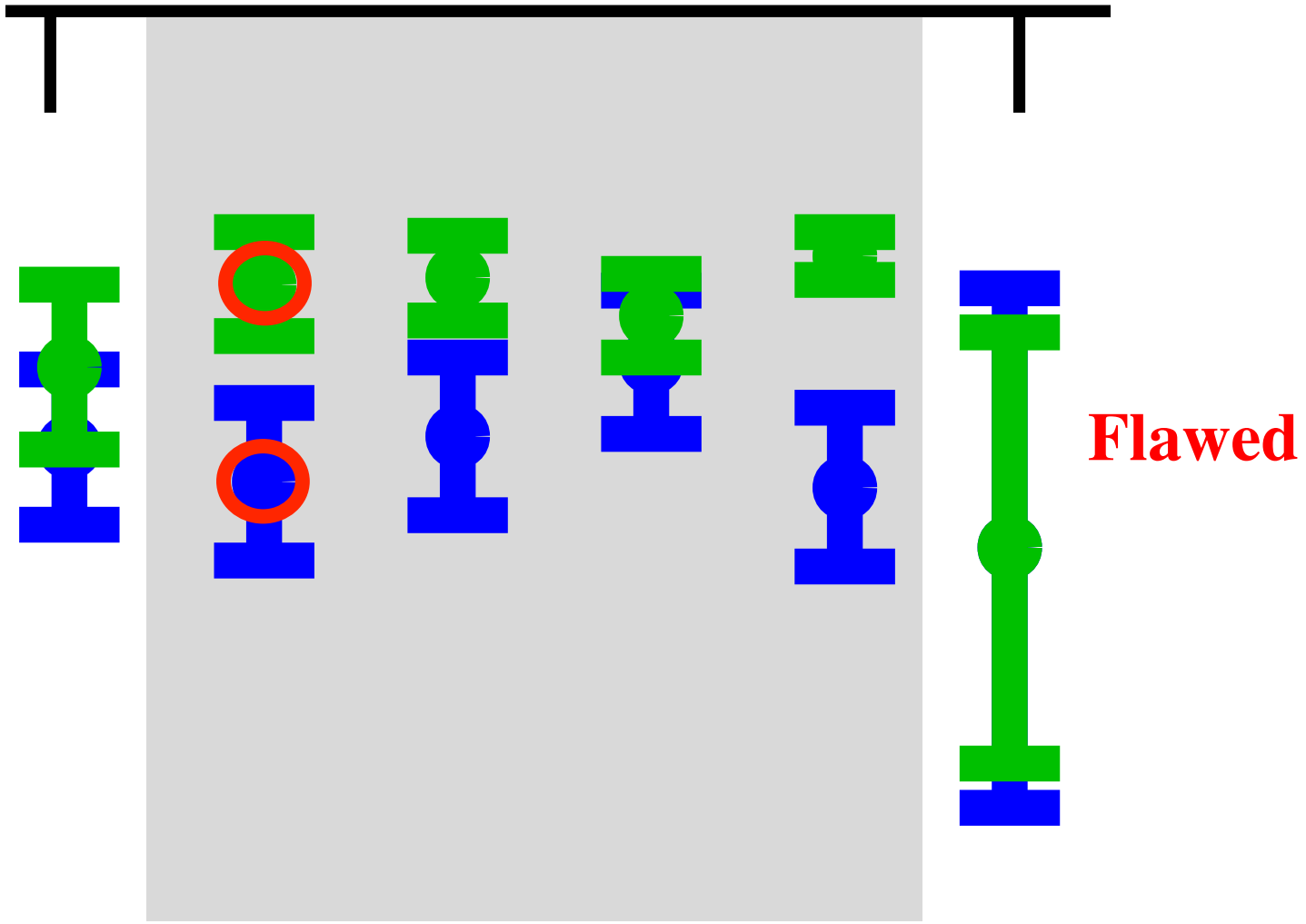


Multiple Consecutive Trials

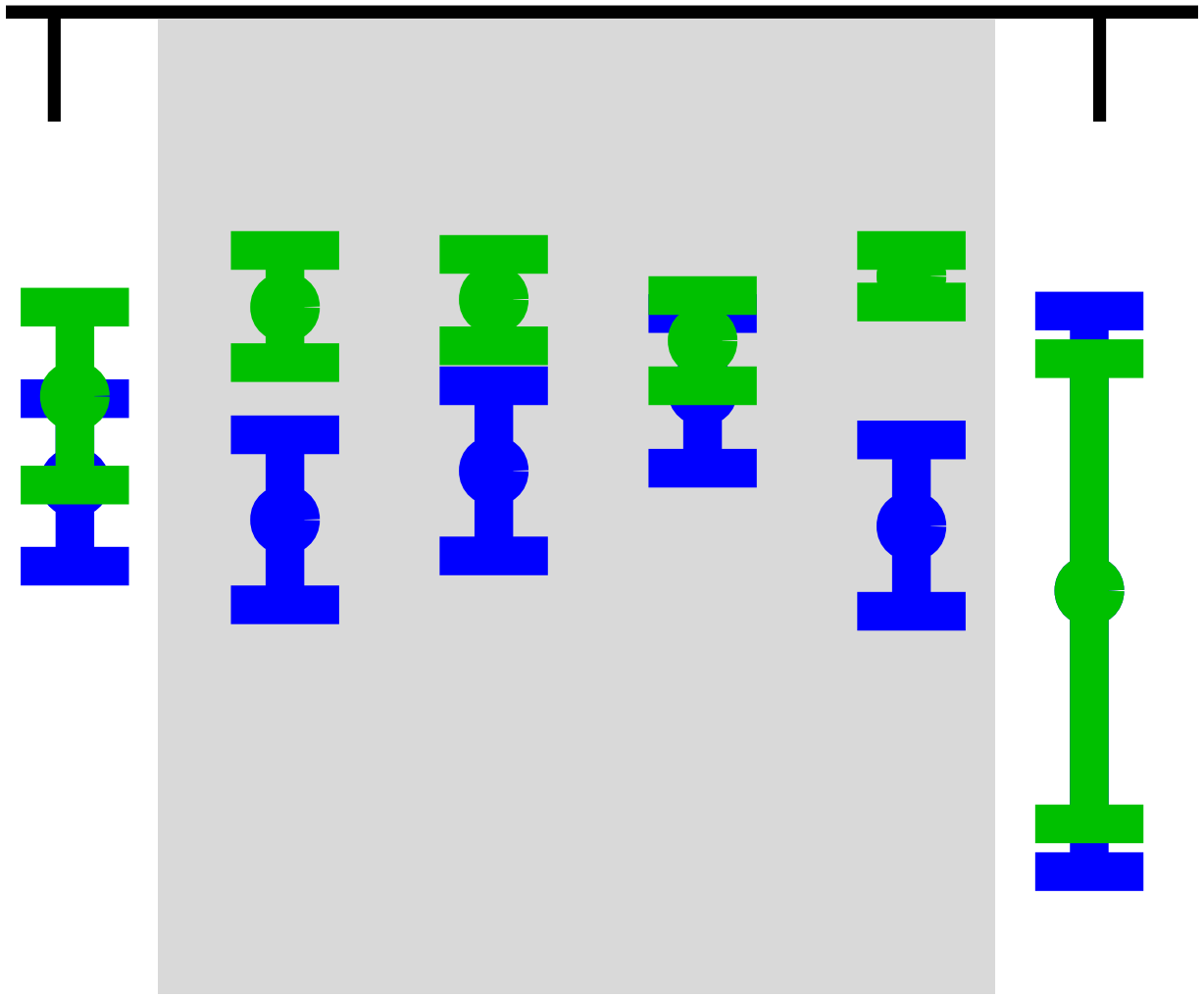
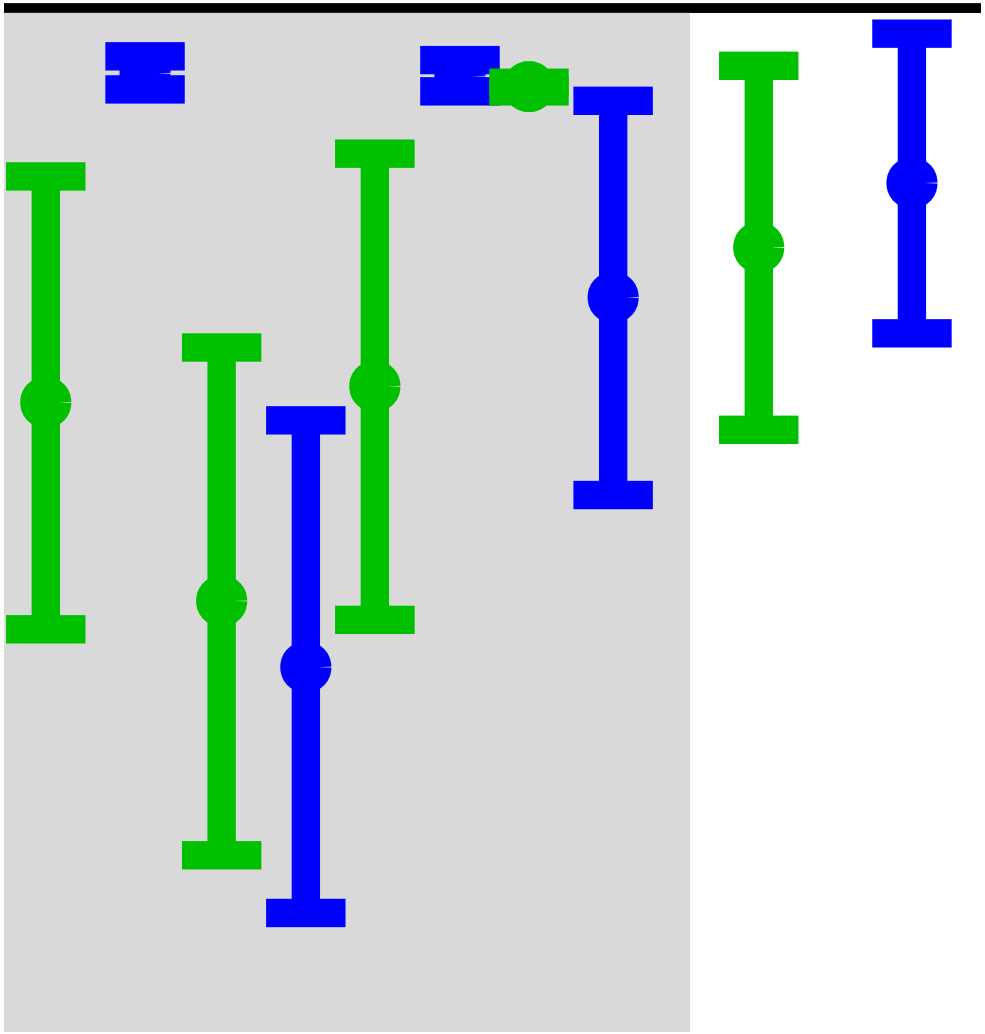


Flawed

Multiple Interleaved Trials



Multiple Interleaved Trials



Multiple Consecutive Trials