UNIVERSITY OF AMSTERDAM

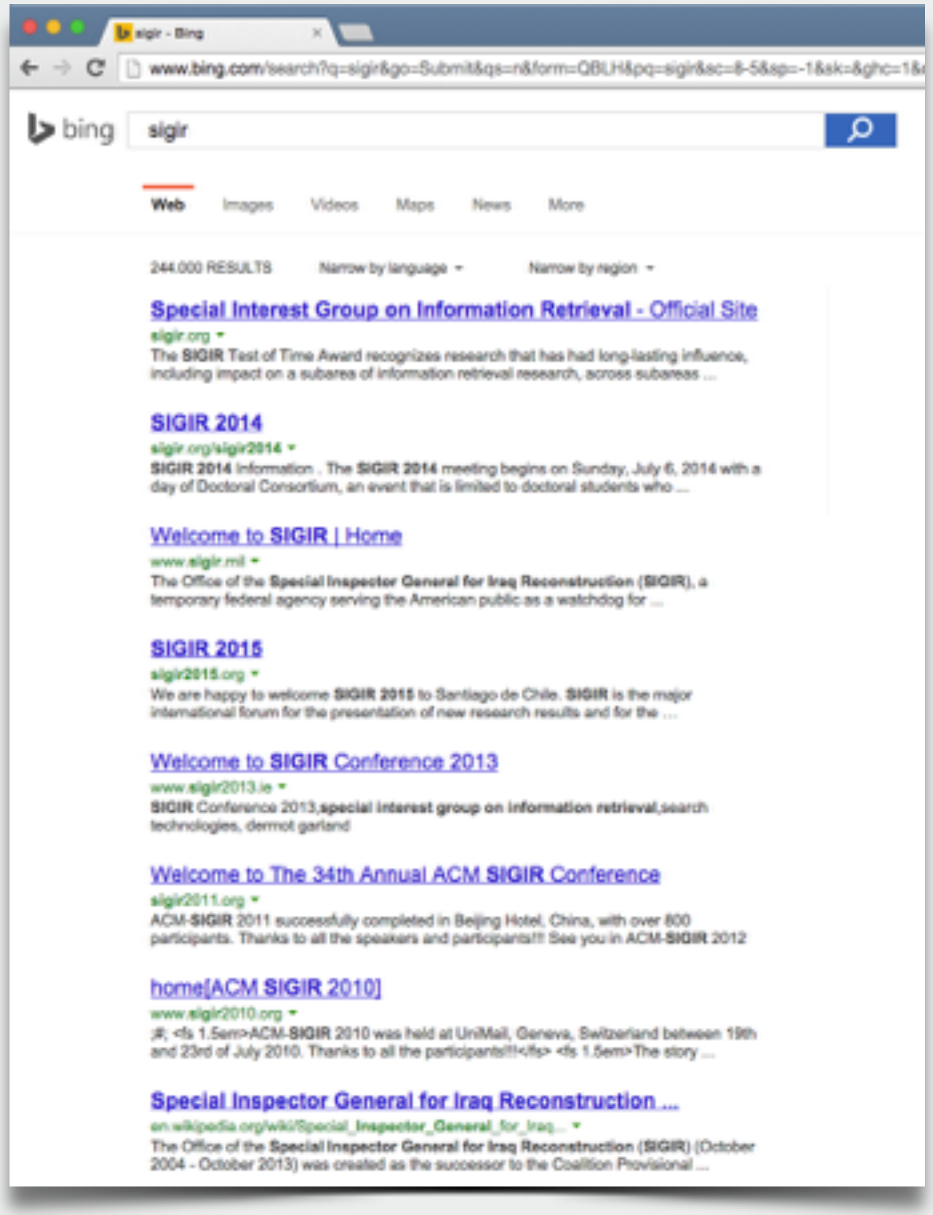
Predicting Search Satisfaction Metrics with Interleaved Comparisons

Anne Schuth
University of Amsterdam
anne.schuth@uva.nl

Katja Hofmann
Microsoft
katja.hofmann@microsoft.com

Filip Radlinski
Microsoft
filiprad@microsoft.com

Motivation - **Evaluation**


sigir - Bing
www.bing.com/search?q=sigir&go=Submit&qs=n&form=QBLH&pq=sigir&sc=8-5&sp=-1&sk=&ghc=1&
bing
sigir
Web
Images
Videos
Maps
News
More

244.000 RESULTS Narrow by language Narrow by region

Special Interest Group on Information Retrieval - Official Site
sigir.org
The **SIGIR** Test of Time Award recognizes research that has had long-lasting influence, including impact on a subarea of information retrieval research, across subareas ...

SIGIR 2014
sigir.org/sigir2014
SIGIR 2014 Information . The **SIGIR 2014** meeting begins on Sunday, July 6, 2014 with a day of Doctoral Consortium, an event that is limited to doctoral students who ...

Welcome to SIGIR | Home
www.sigir.mil
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**), a temporary federal agency serving the American public as a watchdog for ...

SIGIR 2015
sigir2015.org
We are happy to welcome **SIGIR 2015** to Santiago de Chile. **SIGIR** is the major international forum for the presentation of new research results and for the ...

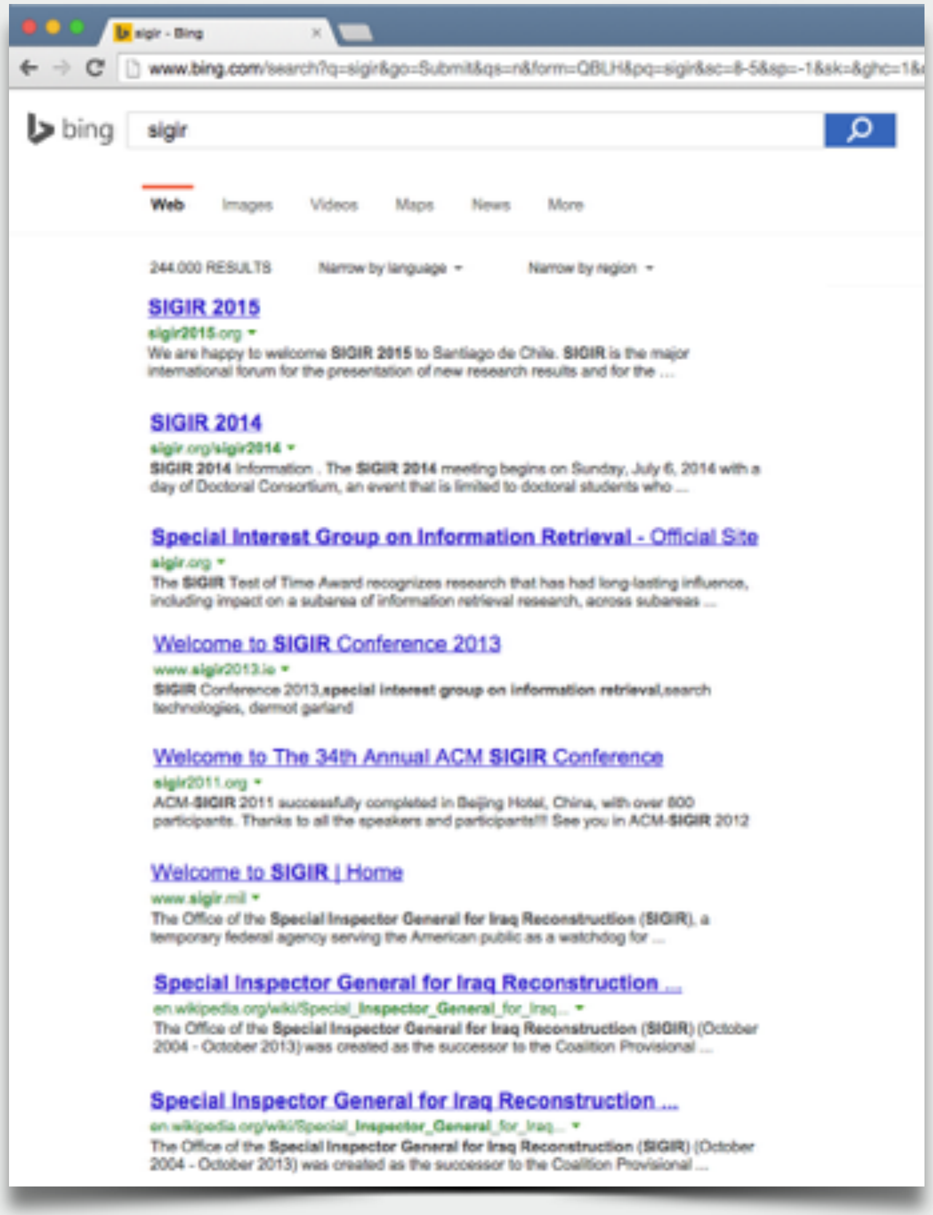
Welcome to SIGIR Conference 2013
www.sigir2013.ie
SIGIR Conference 2013,**special interest group on information retrieval**,search technologies, dermot garland

Welcome to The 34th Annual ACM SIGIR Conference
sigir2011.org
ACM-**SIGIR** 2011 successfully completed in Beijing Hotel, China, with over 800 participants. Thanks to all the speakers and participants!!! See you in ACM-**SIGIR** 2012

home[ACM SIGIR 2010]
www.sigir2010.org
;#; <fs 1.5em>ACM-**SIGIR** 2010 was held at UniMail, Geneva, Switzerland between 19th and 23rd of July 2010. Thanks to all the participants!!!</fs> <fs 1.5em>The story ...

Special Inspector General for Iraq Reconstruction ...
en.wikipedia.org/wiki/Special_Inspector_General_for_Iraq...
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**) (October 2004 - October 2013) was created as the successor to the Coalition Provisional ...

or


sigir - Bing
www.bing.com/search?q=sigir&go=Submit&qs=n&form=QBLH&pq=sigir&sc=8-5&sp=-1&sk=&ghc=1&
bing
sigir
Web
Images
Videos
Maps
News
More

244.000 RESULTS Narrow by language Narrow by region

SIGIR 2015
sigir2015.org
We are happy to welcome **SIGIR 2015** to Santiago de Chile. **SIGIR** is the major international forum for the presentation of new research results and for the ...

SIGIR 2014
sigir.org/sigir2014
SIGIR 2014 Information . The **SIGIR 2014** meeting begins on Sunday, July 6, 2014 with a day of Doctoral Consortium, an event that is limited to doctoral students who ...

Special Interest Group on Information Retrieval - Official Site
sigir.org
The **SIGIR** Test of Time Award recognizes research that has had long-lasting influence, including impact on a subarea of information retrieval research, across subareas ...

Welcome to SIGIR Conference 2013
www.sigir2013.ie
SIGIR Conference 2013,**special interest group on information retrieval**,search technologies, dermot garland

Welcome to The 34th Annual ACM SIGIR Conference
sigir2011.org
ACM-**SIGIR** 2011 successfully completed in Beijing Hotel, China, with over 800 participants. Thanks to all the speakers and participants!!! See you in ACM-**SIGIR** 2012

Welcome to SIGIR | Home
www.sigir.mil
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**), a temporary federal agency serving the American public as a watchdog for ...

Special Inspector General for Iraq Reconstruction ...
en.wikipedia.org/wiki/Special_Inspector_General_for_Iraq...
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**) (October 2004 - October 2013) was created as the successor to the Coalition Provisional ...

Special Inspector General for Iraq Reconstruction ...
en.wikipedia.org/wiki/Special_Inspector_General_for_Iraq...
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**) (October 2004 - October 2013) was created as the successor to the Coalition Provisional ...

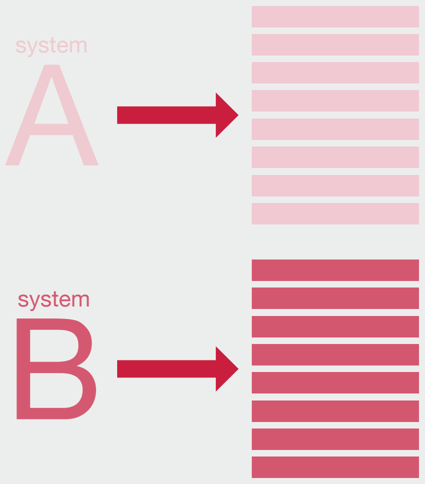
Motivation - **Evaluation**

system
A

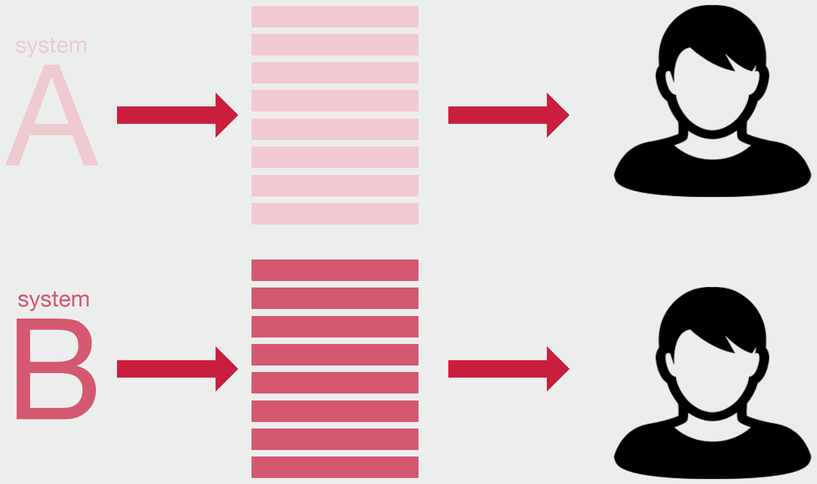
or

system
B

Motivation - **AB Testing**

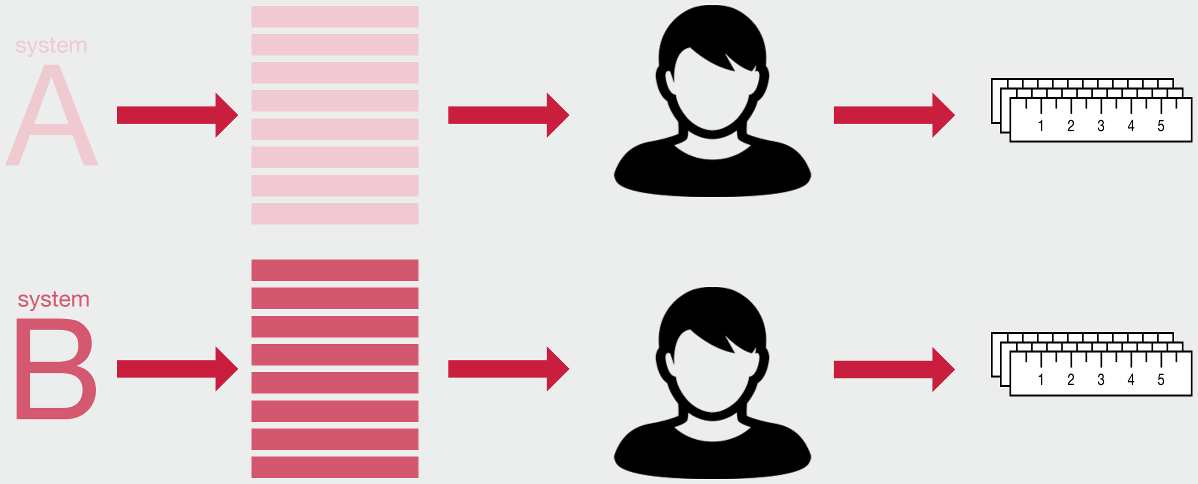


Motivation - **AB Testing**



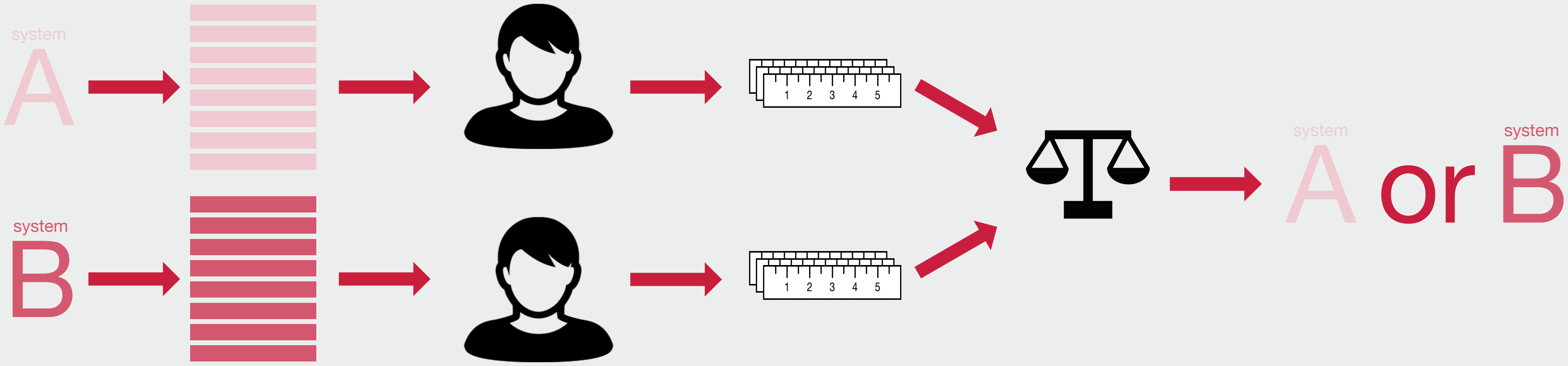
- User population **divided** into two groups

Motivation - **AB Testing**



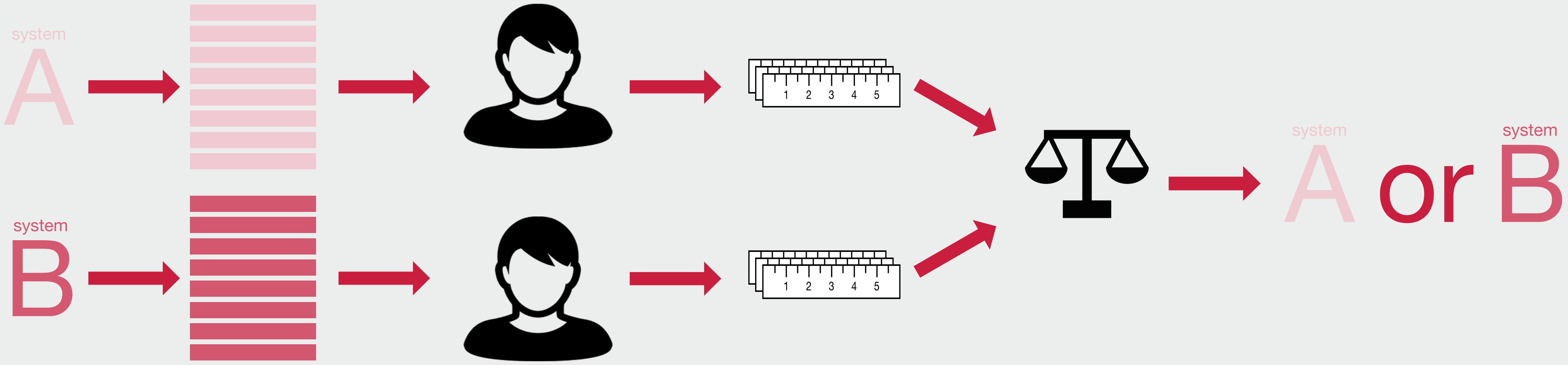
- User population **divided** into two groups
- Trusted and **sophisticated metrics**

Motivation - **AB Testing**



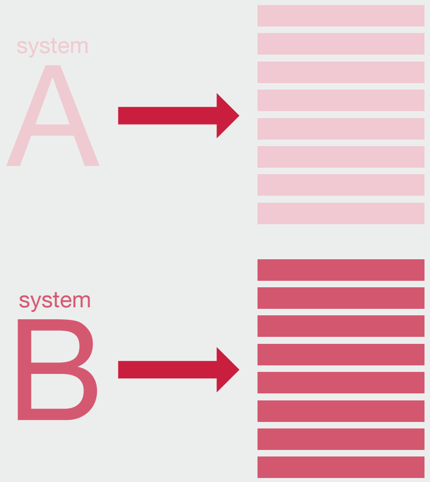
- User population **divided** into two groups
- Trusted and **sophisticated metrics**
- **Difference in metric** indicates the winner

Motivation - **AB Testing**

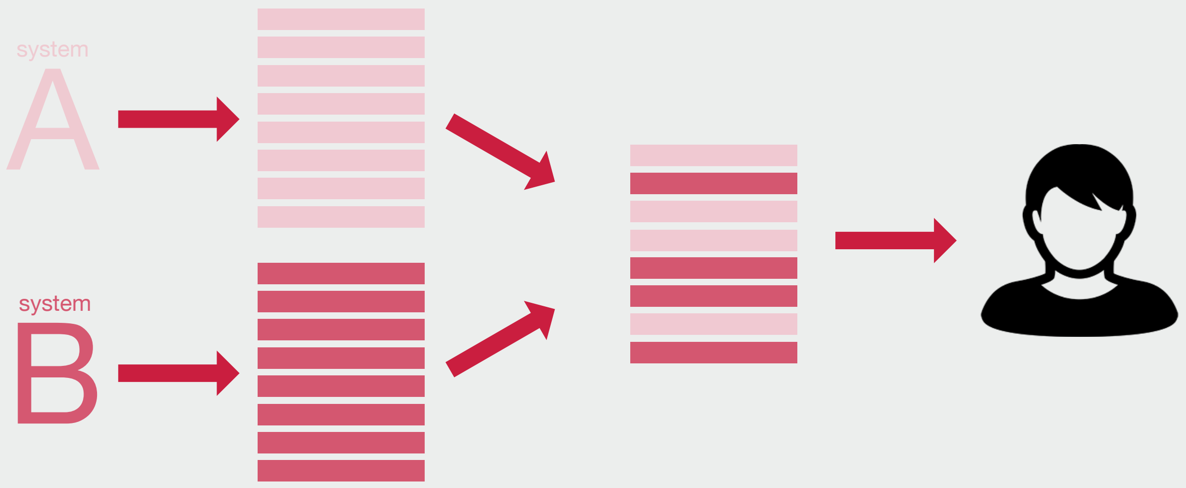


- **User population divided** into two groups
- Trusted and **sophisticated metrics**
- **Difference in metric** indicates the winner
- **Between subject** design
  - Differences between users and their queries
  - **Low sensitivity**, millions of queries

Motivation - **Interleaving**

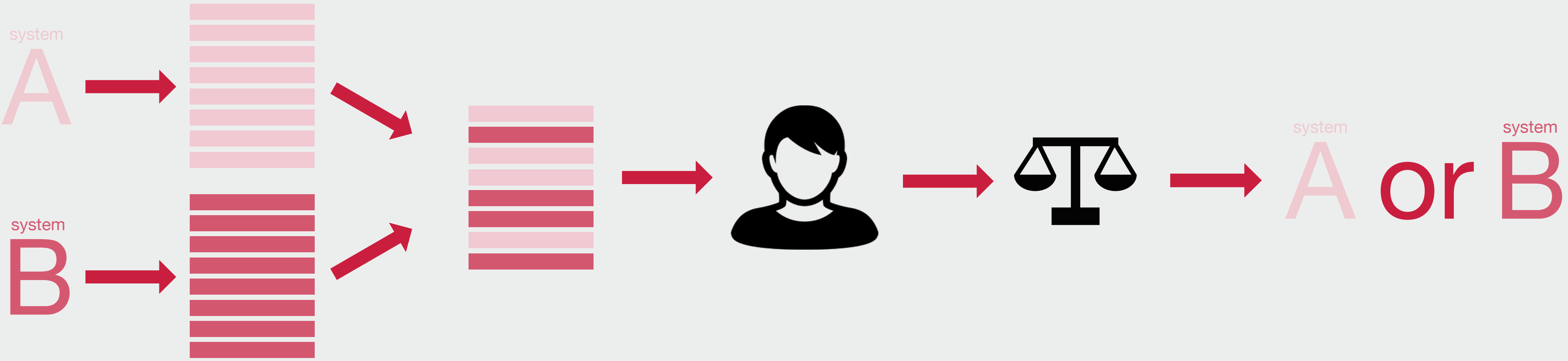


Motivation - **Interleaving**



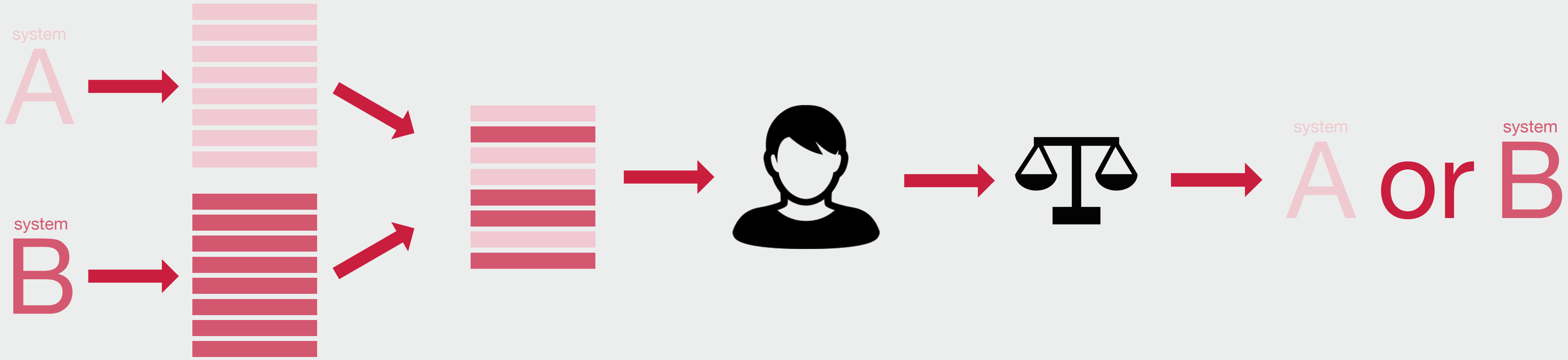
- **Users see both systems**

Motivation - **Interleaving**



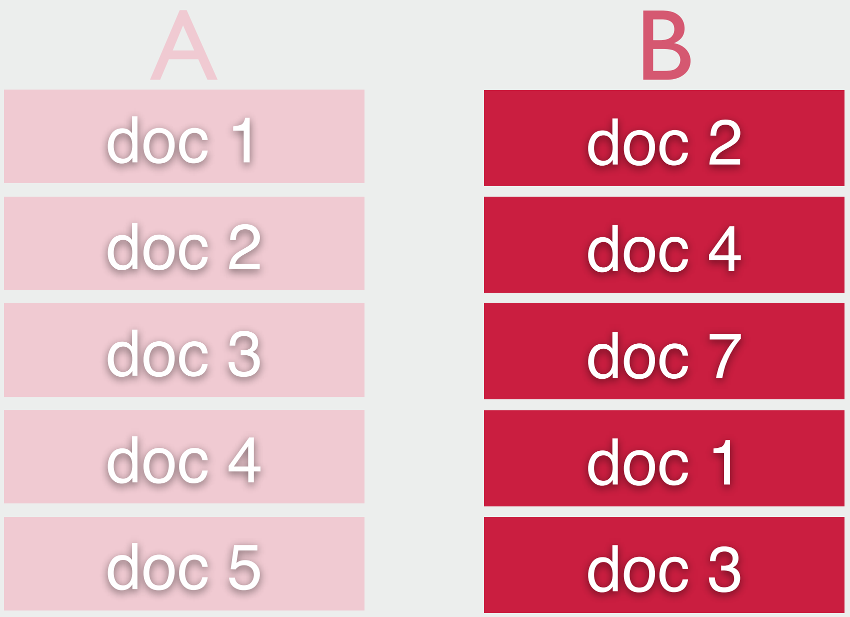
- Users see **both** systems
- **Simple metric:** system with more clicks wins

Motivation - **Interleaving**



- Users see **both** systems
- **Simple metric:** system with more clicks wins
- **Within subject** design
  - **Both systems** now cater for **every user**
  - **High sensitivity**, 10-100x less queries needed (compared to AB Testing)

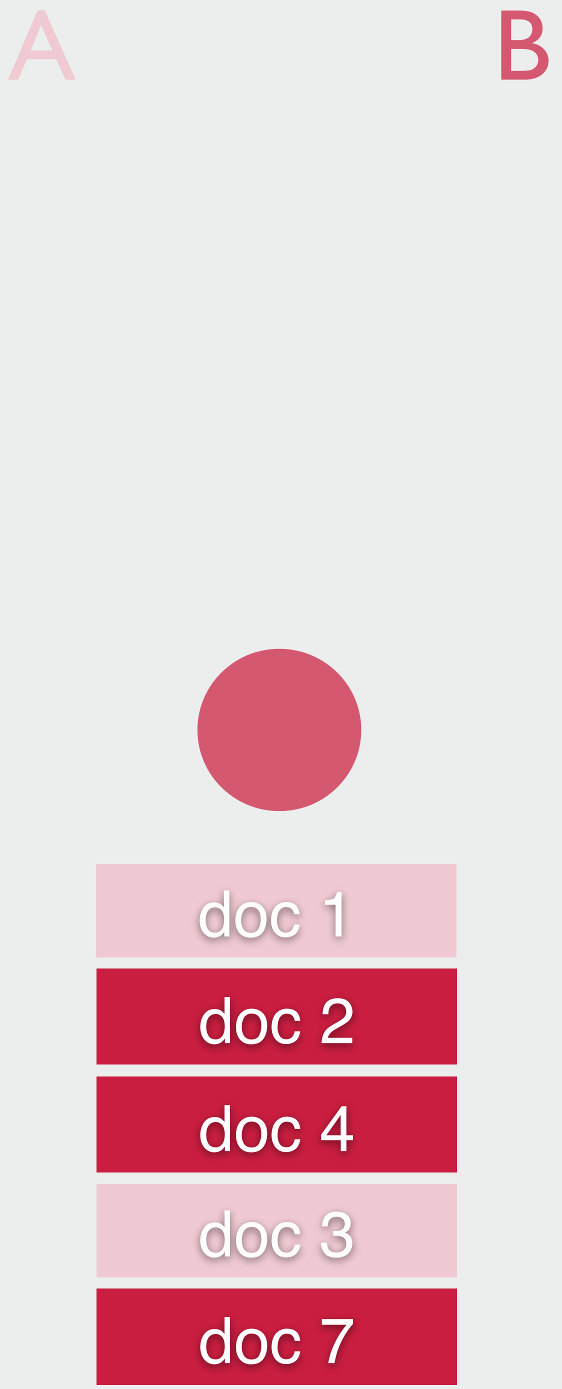
Motivation - **Team Draft Interleaving (TDI)**



A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 7
doc 4	doc 1
doc 5	doc 3

F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008

Motivation - **Team Draft Interleaving (TDI)**



A

B

doc 1

doc 2

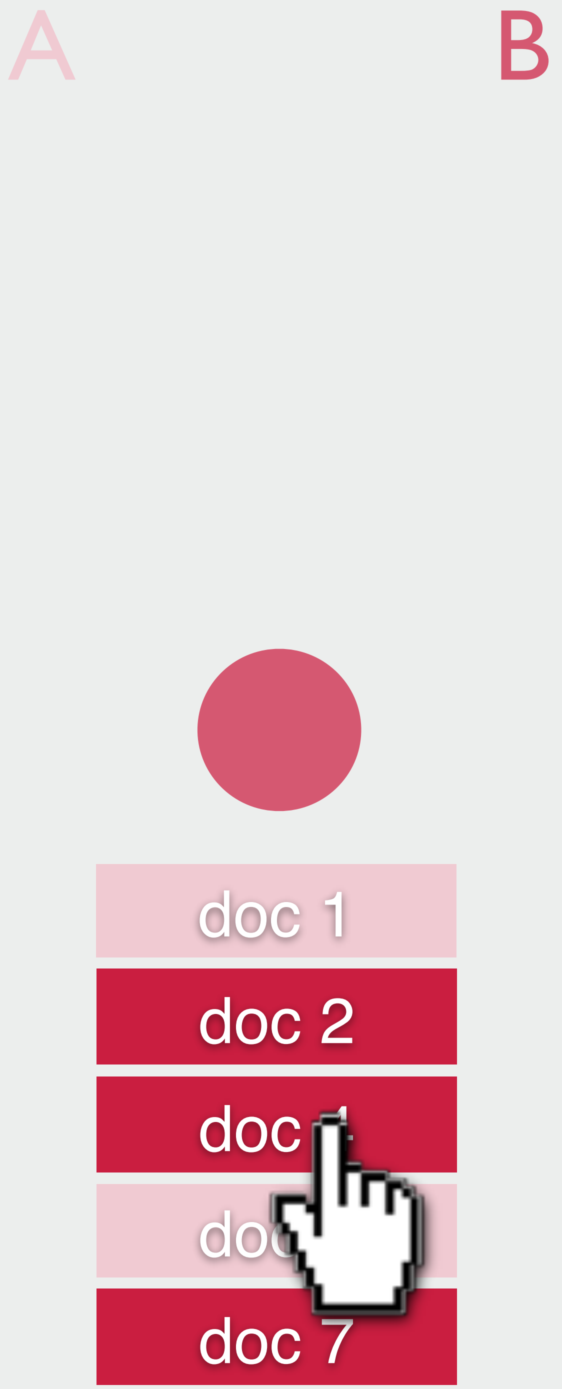
doc 4

doc 3

doc 7

F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008

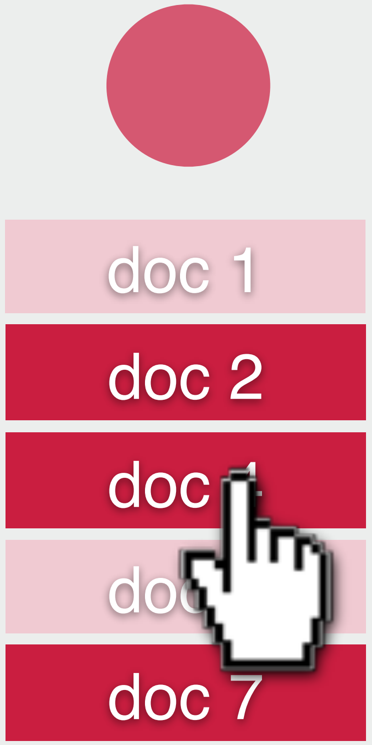
Motivation - **Team Draft Interleaving (TDI)**



F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. 2008

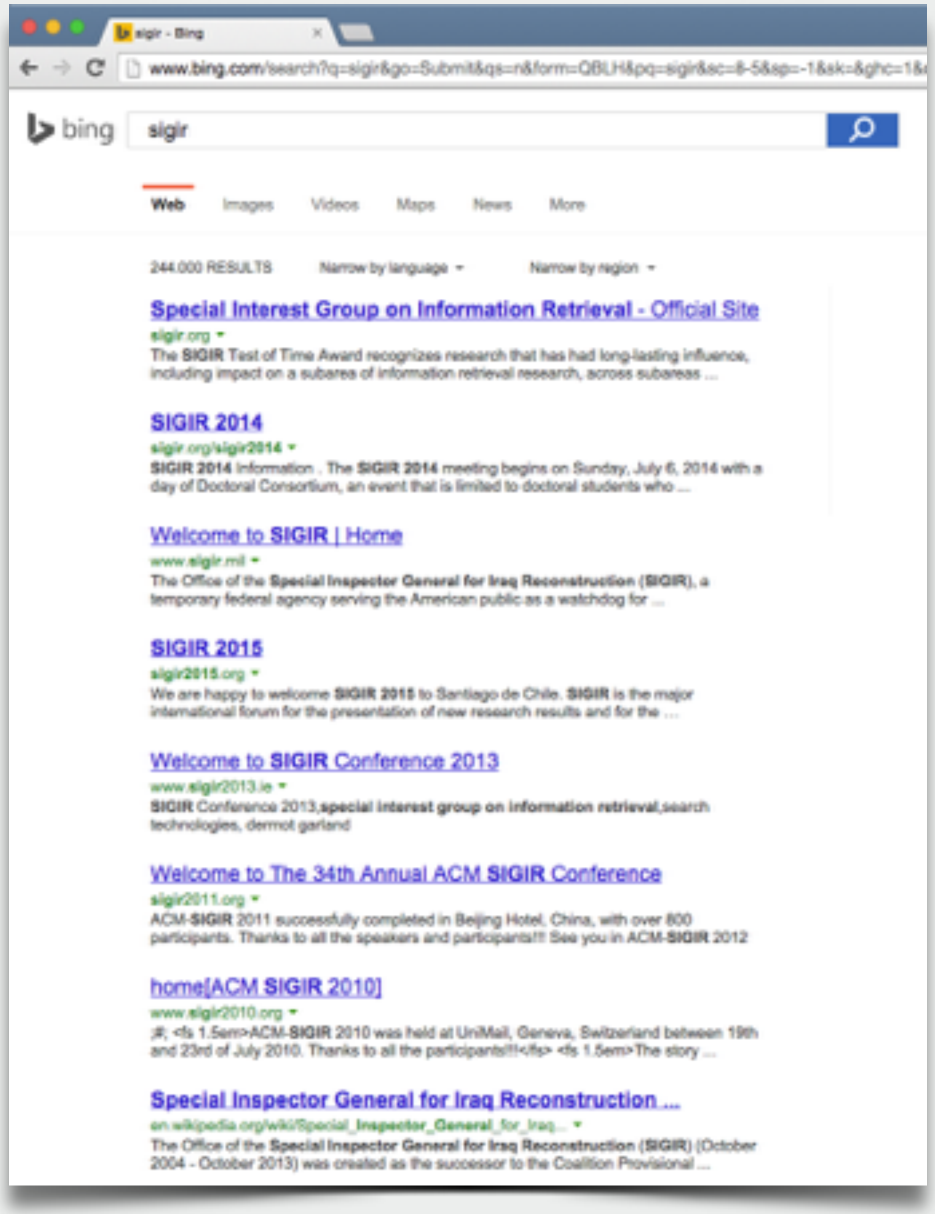
Motivation - **Team Draft Interleaving (TDI)**

- **Infer** winner per query
  - System with more **clicks** wins
  - A < B
- Count **number of wins** over many queries



F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. 2008

Motivation - AB Testing - **As a Gold Standard**



Special Interest Group on Information Retrieval - Official Site
sigir.org
The **SIGIR** Test of Time Award recognizes research that has had long-lasting influence, including impact on a subarea of information retrieval research, across subareas ...

SIGIR 2014
sigir.org/sigir2014
SIGIR 2014 Information . The **SIGIR 2014** meeting begins on Sunday, July 6, 2014 with a day of Doctoral Consortium, an event that is limited to doctoral students who ...

Welcome to SIGIR | Home
www.sigir.mil
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**), a temporary federal agency serving the American public as a watchdog for ...

SIGIR 2015
sigir2015.org
We are happy to welcome **SIGIR 2015** to Santiago de Chile. **SIGIR** is the major international forum for the presentation of new research results and for the ...

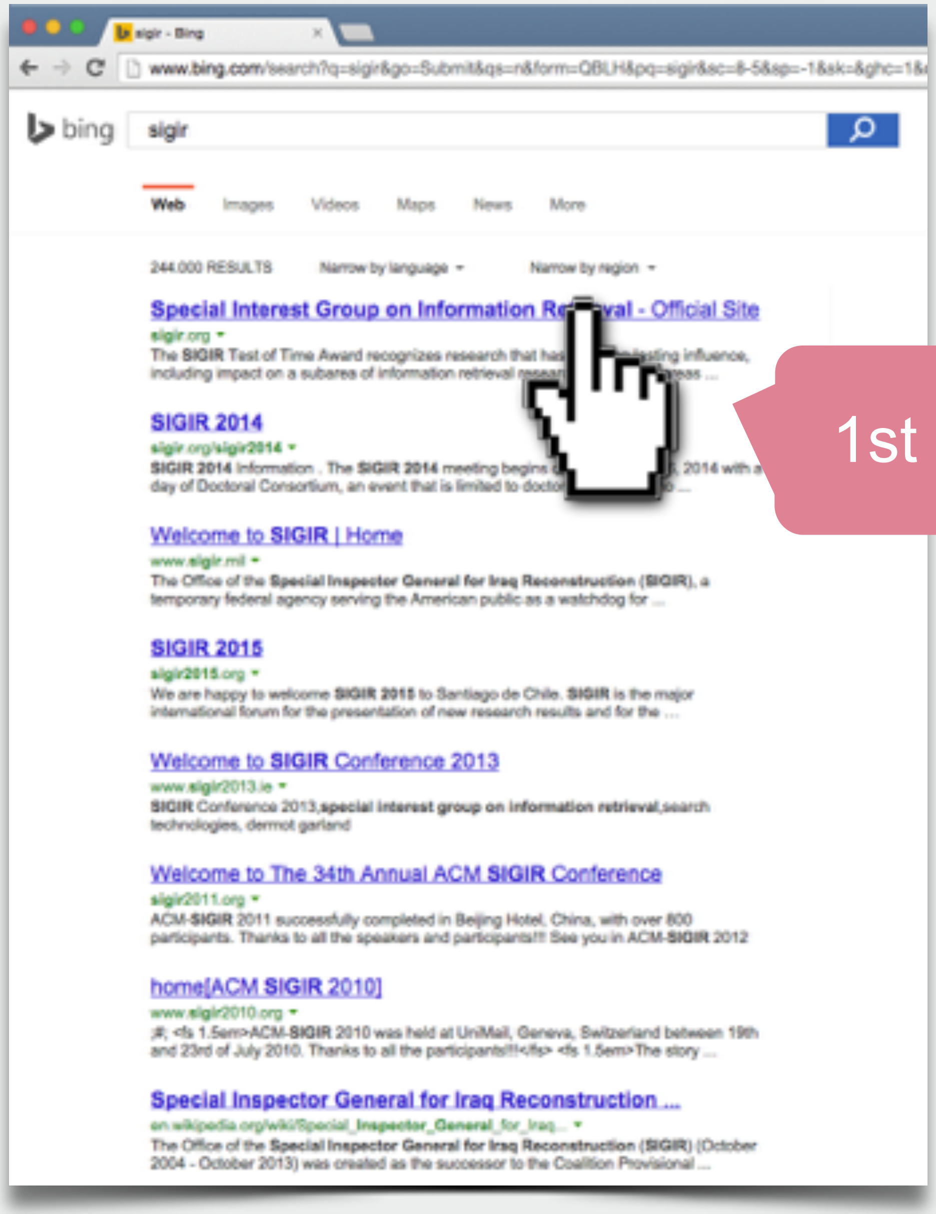
Welcome to SIGIR Conference 2013
www.sigir2013.ie
SIGIR Conference 2013,**special interest group on information retrieval**,search technologies, dermot garland

Welcome to The 34th Annual ACM SIGIR Conference
sigir2011.org
ACM-**SIGIR** 2011 successfully completed in Beijing Hotel, China, with over 800 participants. Thanks to all the speakers and participants!!! See you in ACM-**SIGIR** 2012

home[ACM SIGIR 2010]
www.sigir2010.org
;#; <fs 1.5em>ACM-**SIGIR** 2010 was held at UniMail, Geneva, Switzerland between 19th and 23rd of July 2010. Thanks to all the participants!!!</fs> <fs 1.5em>The story ...

Special Inspector General for Iraq Reconstruction ...
en.wikipedia.org/wiki/Special_Inspector_General_for_Iraq...
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**) (October 2004 - October 2013) was created as the successor to the Coalition Provisional ...

Motivation - AB Testing - **As a Gold Standard**



sigir - Bing
www.bing.com/search?q=sigir&go=Submit&qs=n&form=QBLH&pq=sigir&sc=8-5&sp=-1&sk=&ghc=1&
bing
sigir
Web
Images
Videos
Maps
News
More

244.000 RESULTS Narrow by language Narrow by region

Special Interest Group on Information Re[illegible]val - Official Site
sigir.org
The **SIGIR** Test of Time Award recognizes research that ha[illegible]sting influence, including impact on a subarea of information retrieval [illegible]eas ...

SIGIR 2014
sigir.org/sigir2014
SIGIR 2014 Information . The **SIGIR 2014** meeting begins [illegible], 2014 with a day of Doctoral Consortium, an event that is limited to doctor[illegible] ...

Welcome to **SIGIR** | Home
www.sigir.mil
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**), a temporary federal agency serving the American public as a watchdog for ...

SIGIR 2015
sigir2015.org
We are happy to welcome **SIGIR 2015** to Santiago de Chile. **SIGIR** is the major international forum for the presentation of new research results and for the ...

Welcome to **SIGIR** Conference 2013
www.sigir2013.ie
SIGIR Conference 2013,**special interest group on information retrieval**,search technologies, dermot garland

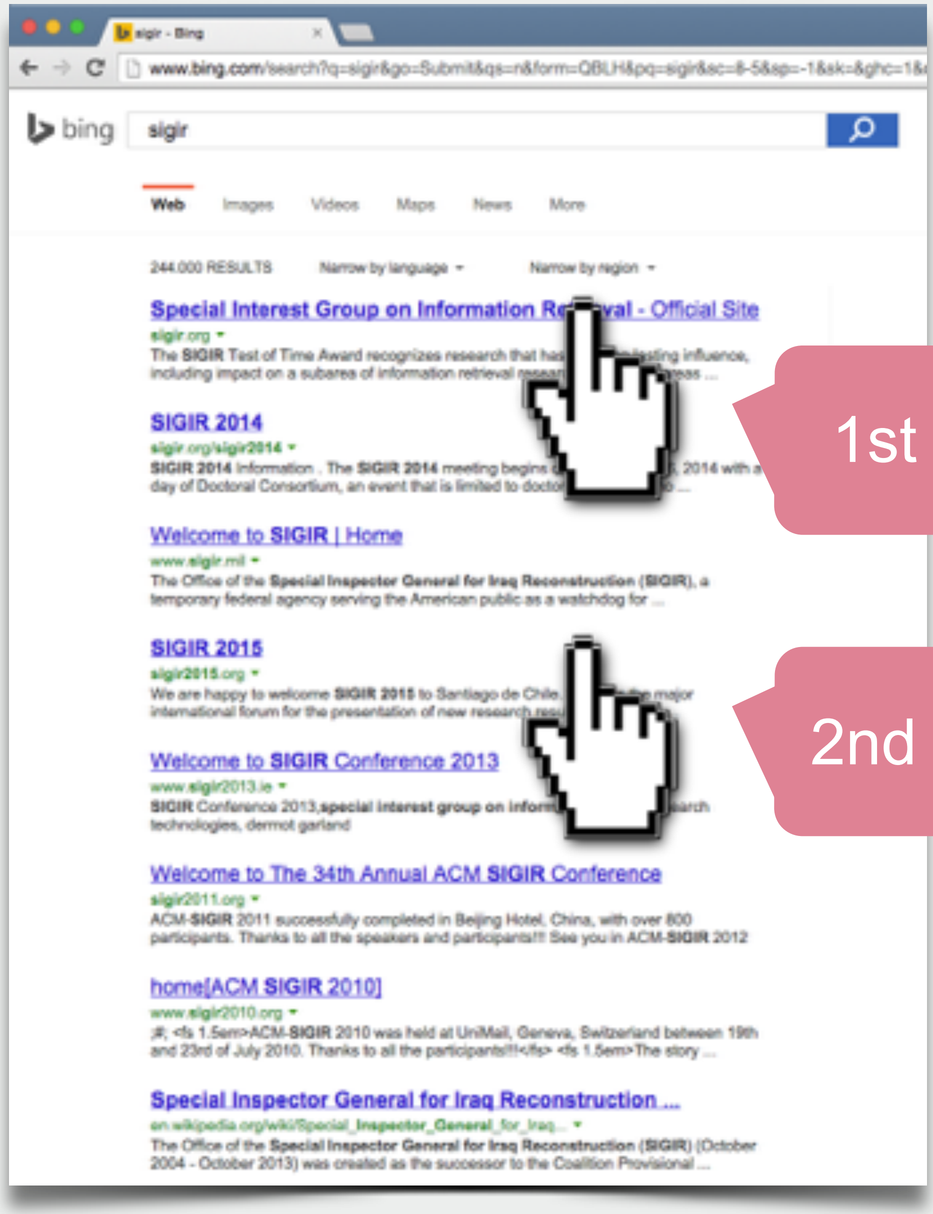
Welcome to The 34th Annual ACM **SIGIR** Conference
sigir2011.org
ACM-**SIGIR** 2011 successfully completed in Beijing Hotel, China, with over 800 participants. Thanks to all the speakers and participants!!! See you in ACM-**SIGIR** 2012

home[ACM **SIGIR** 2010]
www.sigir2010.org
;#; <fs 1.5em>ACM-**SIGIR** 2010 was held at UniMail, Geneva, Switzerland between 19th and 23rd of July 2010. Thanks to all the participants!!!</fs> <fs 1.5em>The story ...

Special Inspector General for Iraq Reconstruction ...
en.wikipedia.org/wiki/Special_Inspector_General_for_Iraq...
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**) (October 2004 - October 2013) was created as the successor to the Coalition Provisional ...

1st click, 5sec dwell time

Motivation - AB Testing - **As a Gold Standard**


sigir - Bing
www.bing.com/search?q=sigir&go=Submit&qs=n&form=QBLH&pq=sigir&sc=8-5&sp=-1&sk=&ghc=1&
bing
sigir
Web
Images
Videos
Maps
News
More

244.000 RESULTS   Narrow by language   Narrow by region

Special Interest Group on Information Re val - Official Site
sigir.org
The **SIGIR** Test of Time Award recognizes research that has sting influence, including impact on a subarea of information retrieval resear eas ...

SIGIR 2014
sigir.org/sigir2014
SIGIR 2014 Information . The **SIGIR 2014** meeting begins , 2014 with a day of Doctoral Consortium, an event that is limited to doctor o ...

Welcome to **SIGIR** | Home
www.sigir.mil
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**), a temporary federal agency serving the American public as a watchdog for ...

SIGIR 2015
sigir2015.org
We are happy to welcome **SIGIR 2015** to Santiago de Chile, major international forum for the presentation of new research resu

Welcome to **SIGIR** Conference 2013
www.sigir2013.ie
SIGIR Conference 2013,**special interest group on inform** earch technologies, dermot garland

Welcome to The 34th Annual ACM **SIGIR** Conference
sigir2011.org
ACM-**SIGIR** 2011 successfully completed in Beijing Hotel, China, with over 800 participants. Thanks to all the speakers and participants!!! See you in ACM-**SIGIR** 2012

home[ACM **SIGIR** 2010]
www.sigir2010.org
;#; <fs 1.5em>ACM-**SIGIR** 2010 was held at UniMail, Geneva, Switzerland between 19th and 23rd of July 2010. Thanks to all the participants!!!</fs> <fs 1.5em>The story ...

Special Inspector General for Iraq Reconstruction ...
en.wikipedia.org/wiki/Special_Inspector_General_for_Iraq...
The Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**) (October 2004 - October 2013) was created as the successor to the Coalition Provisional ...

1st click, 5sec dwell time

2nd click, user stays away

Motivation - AB Testing - **Metrics**

AB Metric	Description

Motivation - AB Testing - **Metrics**

AB Metric	Description
AB	Fraction queries with at least one **click**

Motivation - AB Testing - **Metrics**

AB Metric	Description
AB	Fraction queries with at least one **click**
AB@1	Fraction queries with at least one **click on 1st position**

Motivation - AB Testing - **Metrics**

AB Metric	Description
AB	Fraction queries with at least one **click**
AB@1	Fraction queries with at least one **click on 1st positi**
AB_S	Fraction queries with at least one **SAT click**

Classifier predicting **SAT probability** with a **threshold**

Motivation - AB Testing - **Metrics**

AB Metric	Description
AB	Fraction queries with at least one **click**
AB@1	Fraction queries with at least one **click on 1st positio**
AB$_S$	Fraction queries with at least one **SAT click**
AB$_S$@1	Fraction queries with at least one **SAT click on 1st position**

Classifier predicting **SAT probability** with a **threshold**

Motivation - AB Testing - **Metrics**

AB Metric	Description
AB	Fraction queries with at least one **click**
AB@1	Fraction queries with at least one **click on 1st positio**
AB_S	Fraction queries with at least one **SAT click**
AB_S@1	Fraction queries with at least one **SAT click on 1st position**
AB_T	**Time** from the query issue **until first click**

Classifier predicting **SAT probability** with a **threshold**

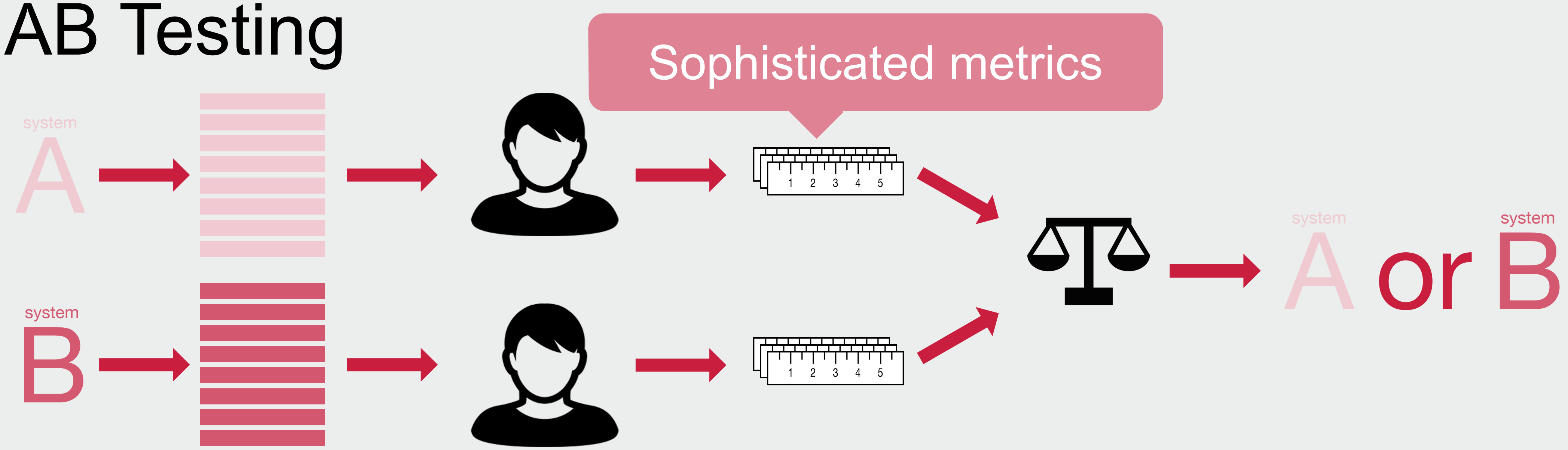
Motivation - AB Testing - **Metrics**

AB Metric	Description
AB	Fraction queries with at least one **click**
AB@1	Fraction queries with at least one **click on 1st positio**
AB_S	Fraction queries with at least one **SAT click**
AB_S@1	Fraction queries with at least one **SAT click on 1st position**
AB_T	**Time** from the query issue **until first click**
AB_T@1	**Time** from the query issue **until first click on 1st position**
$AB_{T,S}$	**Time** from the query issue **until first SAT click**
$AB_{T,S}$@1	**Time** from the query issue **until first SAT click on 1st position**

Classifier predicting **SAT probability** with a **threshold**

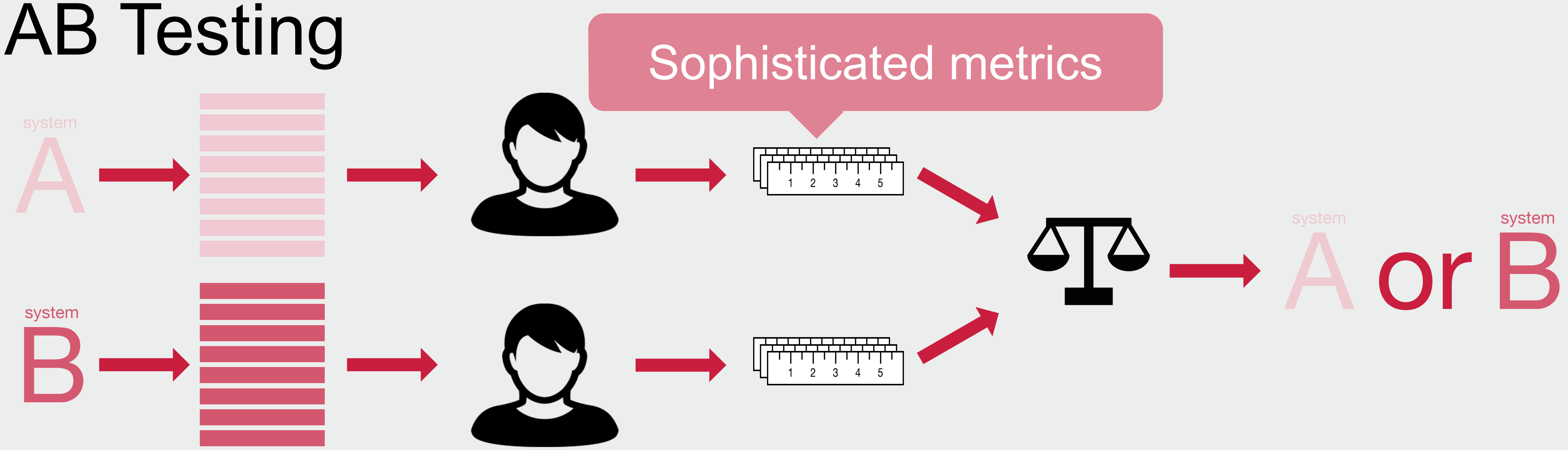
Motivation - **Agreement**

- AB Testing



Motivation - **Agreement**

- AB Testing



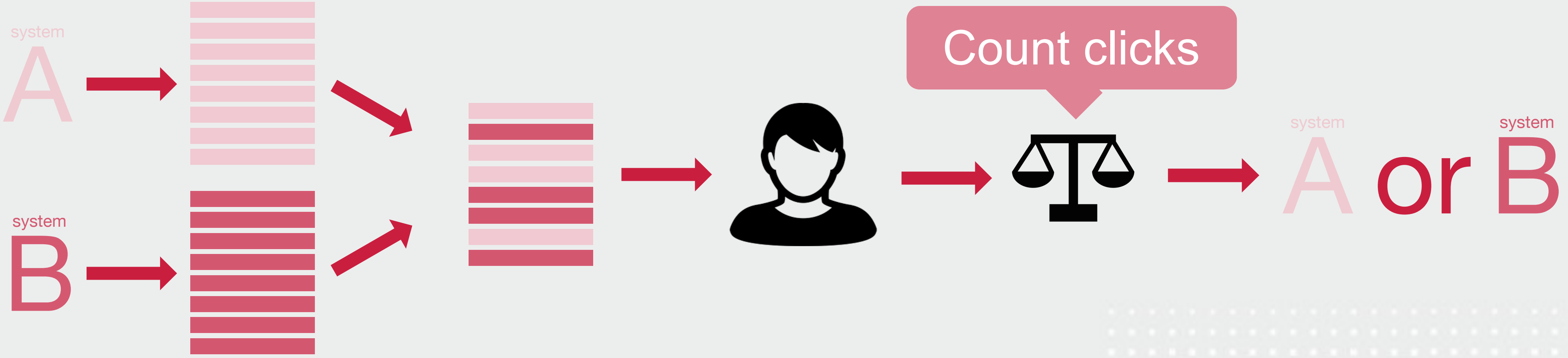
system A

system B

Sophisticated metrics

system A or system B

- Interleaving



system A

system B

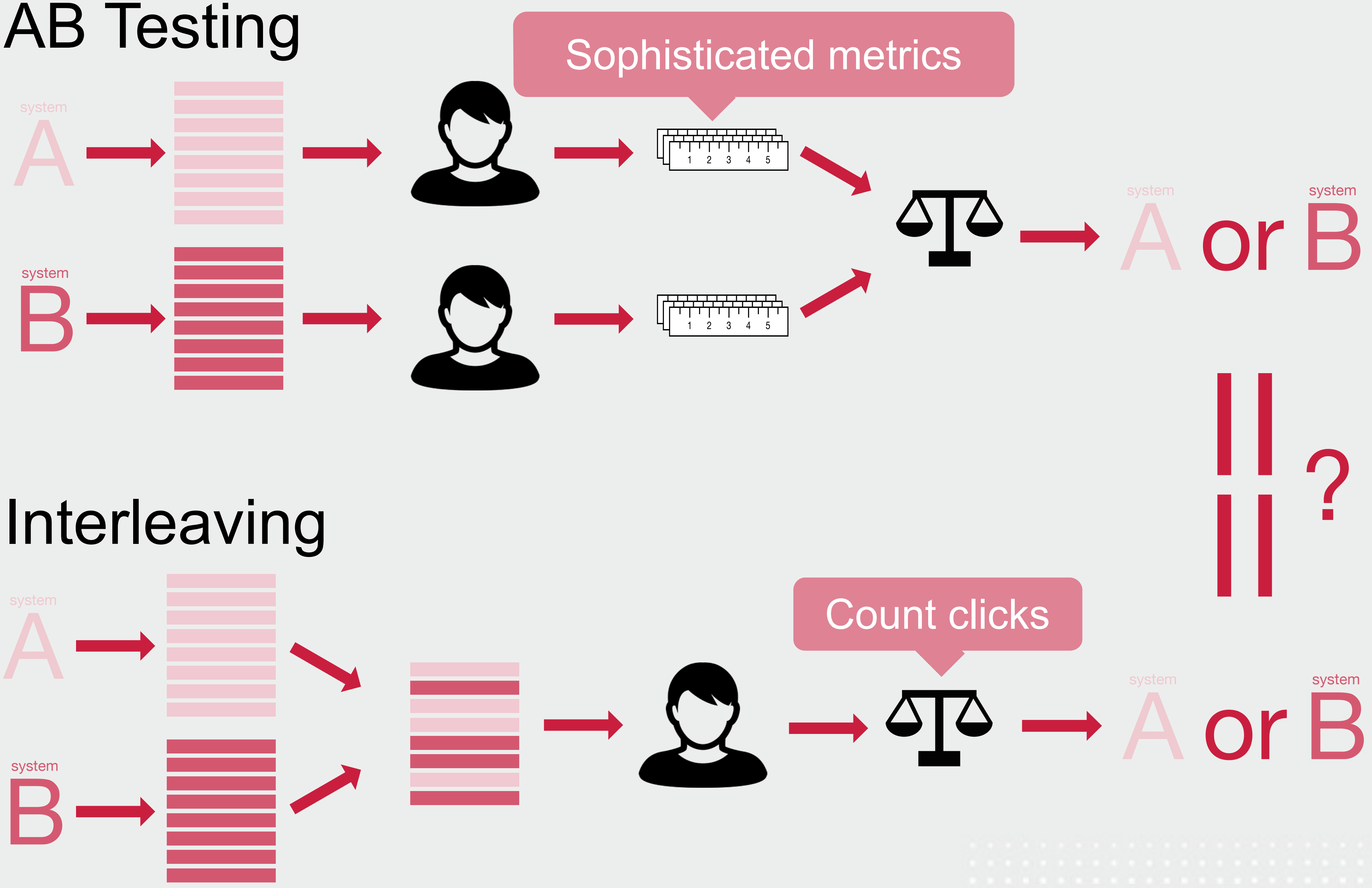
Count clicks

system A or system B

Motivation - **Agreement**

- AB Testing

- Interleaving



Outline

Motivation

Data + analysis

Methods + results

Conclusions

Data - **Properties**

Data - **Properties**

- **38 ranker pairs**

Data - Properties

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop
  - 9 months in 2014

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop
  - 9 months in 2014
  - United States locale

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop
  - 9 months in 2014
  - United States locale
- **Click volume**

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop
  - 9 months in 2014
  - United States locale
- **Click volume**
  - AB: ~1 week, **high** volume

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop
  - 9 months in 2014
  - United States locale
- **Click volume**
  - AB: ~1 week, **high** volume
  - Interleaving: ~4 days, **low** volume

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop
  - 9 months in 2014
  - United States locale
- **Click volume**
  - AB: ~1 week, **high** volume
  - Interleaving: ~4 days, **low** volume
  - **~80 times** more queries for AB

Data - **Properties**

- **38 ranker pairs**
  - AB Tested + Interleaved (TDI)
  - only **ranking** changes
  - bing.com, web, desktop
  - 9 months in 2014
  - United States locale
- **Click volume**
  - AB: ~1 week, **high** volume
  - Interleaving: ~4 days, **low** volume
  - **~80 times** more queries for AB
  - **~3 billion clicks**

Data - Analysis - **Agreement**

- **Interleaving (TDI)** does **not agree** well with **AB metrics**

AB Metric	Interleaving (TDI)
AB	0.63

Data - Analysis - **Agreement**

- **Interleaving (TDI)** does **not agree** well with **AB metrics**

AB Metric	Interleaving (TDI)
AB	0.63
AB@1	**0.71**
AB$_S$	**0.71**
AB$_S$@1	**0.76**
AB$_T$	0.53
AB$_T$@1	0.45
AB$_{T,S}$	0.47
AB$_{T,S}$@1	0.42

Significantly different from random

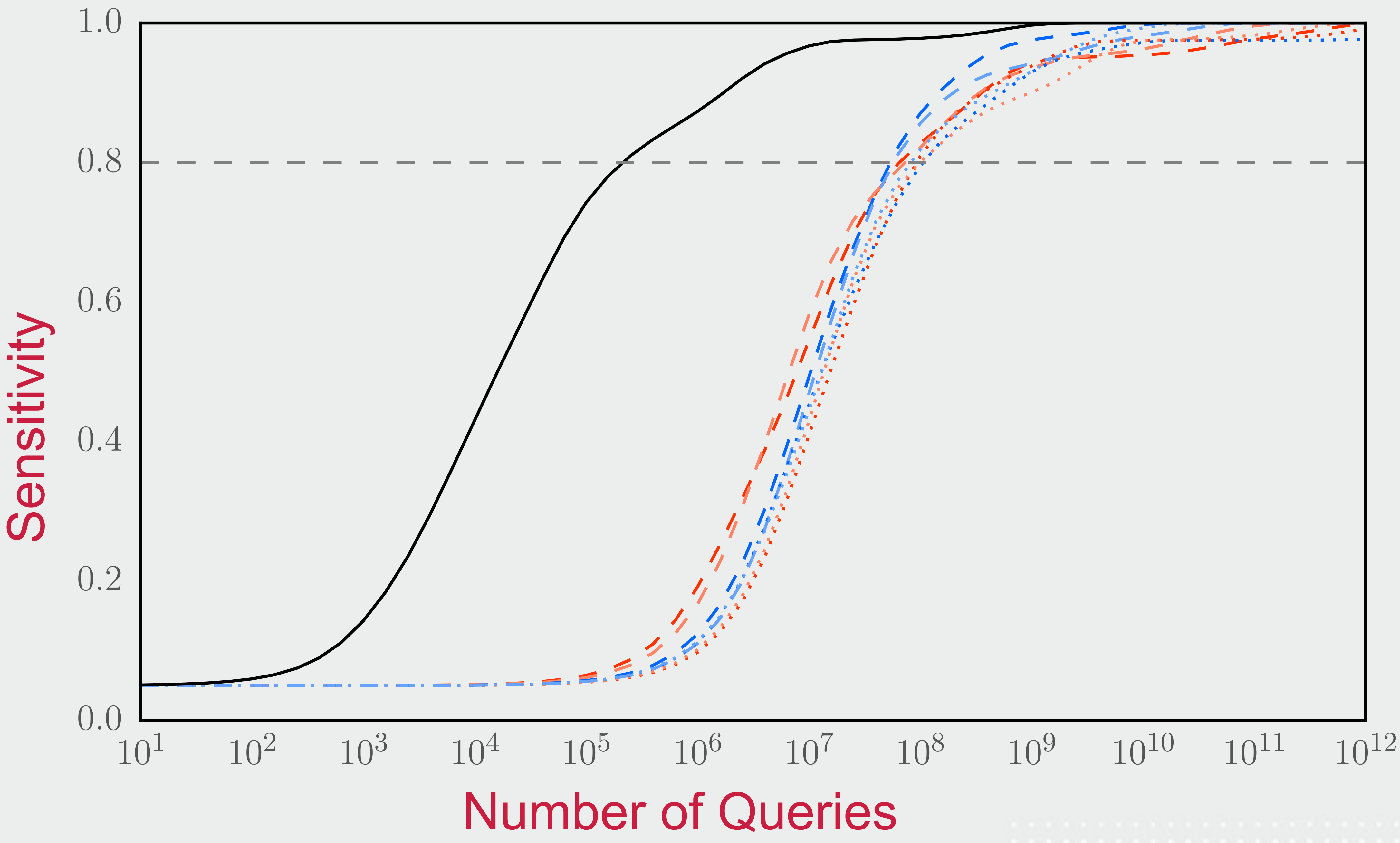
Data - Analysis - **Sensitivity (Power)**

- **How many queries** are required for statistically significant conclusions?

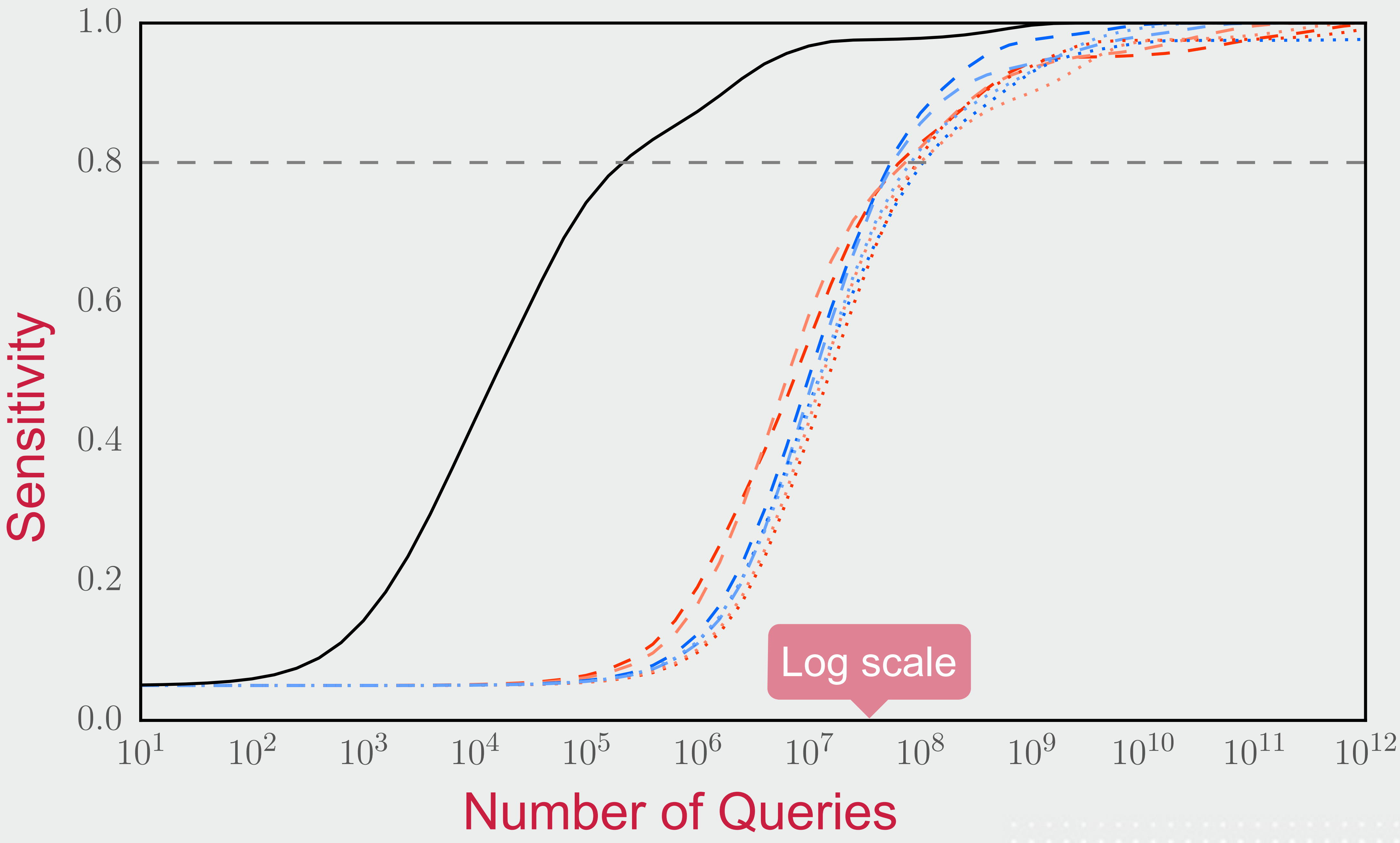
Data - Analysis - **Sensitivity (Power)**

- **How many queries** are required for statistically significant conclusions?
- Sensitivity (power) analysis
  - alpha=0.05, two sided
  - AB Testing: **independent** t-test
  - Interleaving (TDI): **paired** t-test

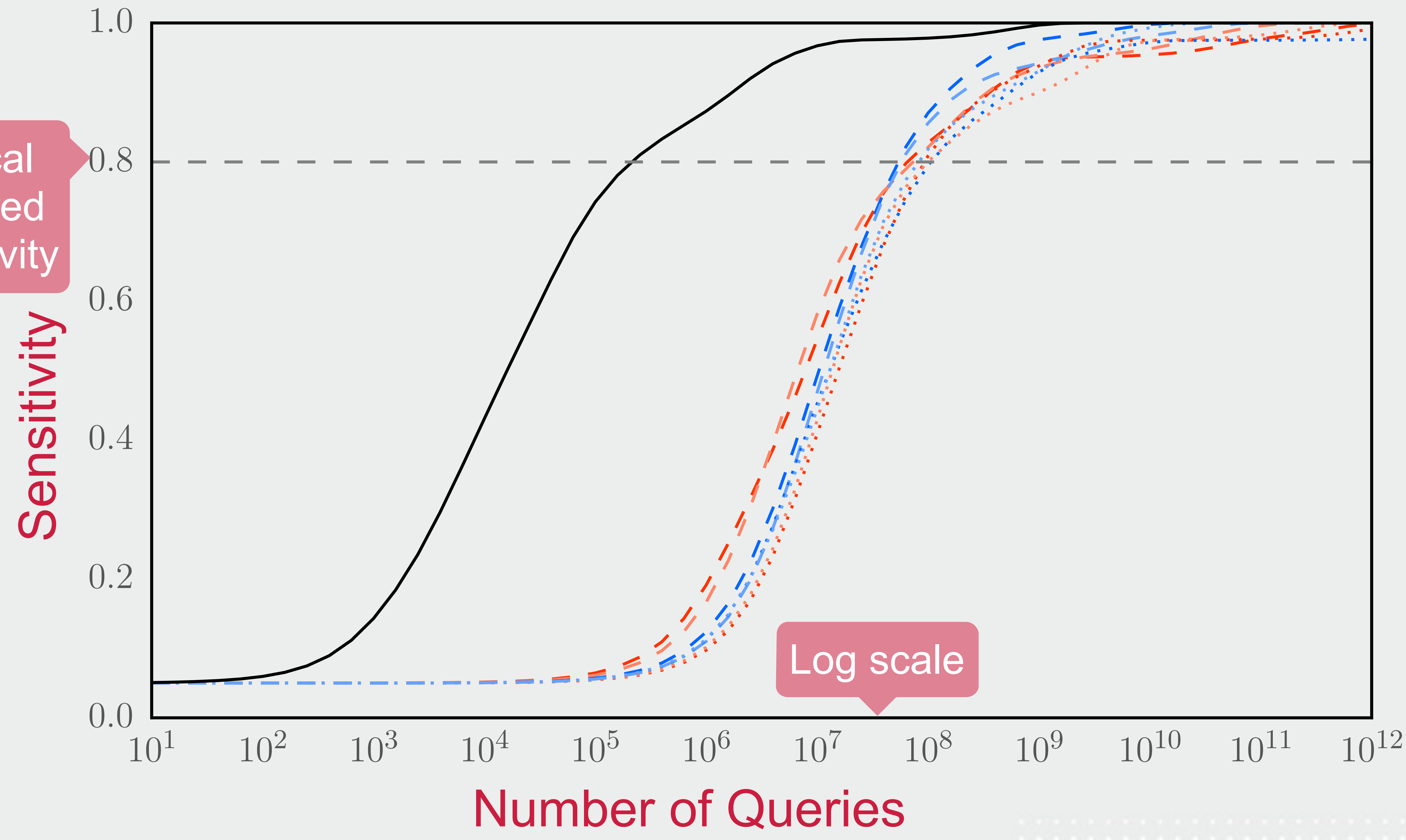
Data - Analysis - **Sensitivity**



Data - Analysis - **Sensitivity**



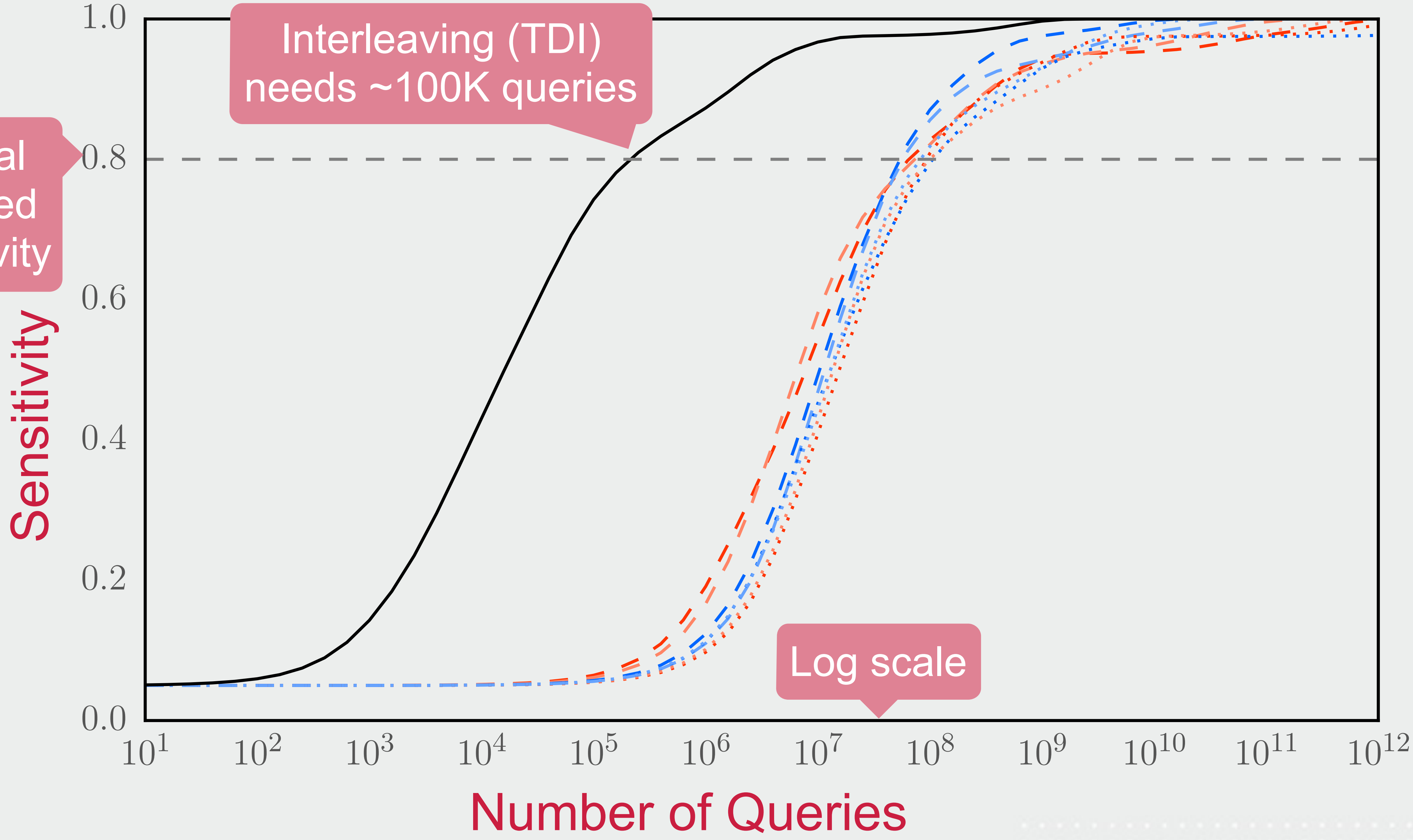
Data - Analysis - **Sensitivity**



Typical required sensitivity

Log scale

Data - Analysis - **Sensitivity**

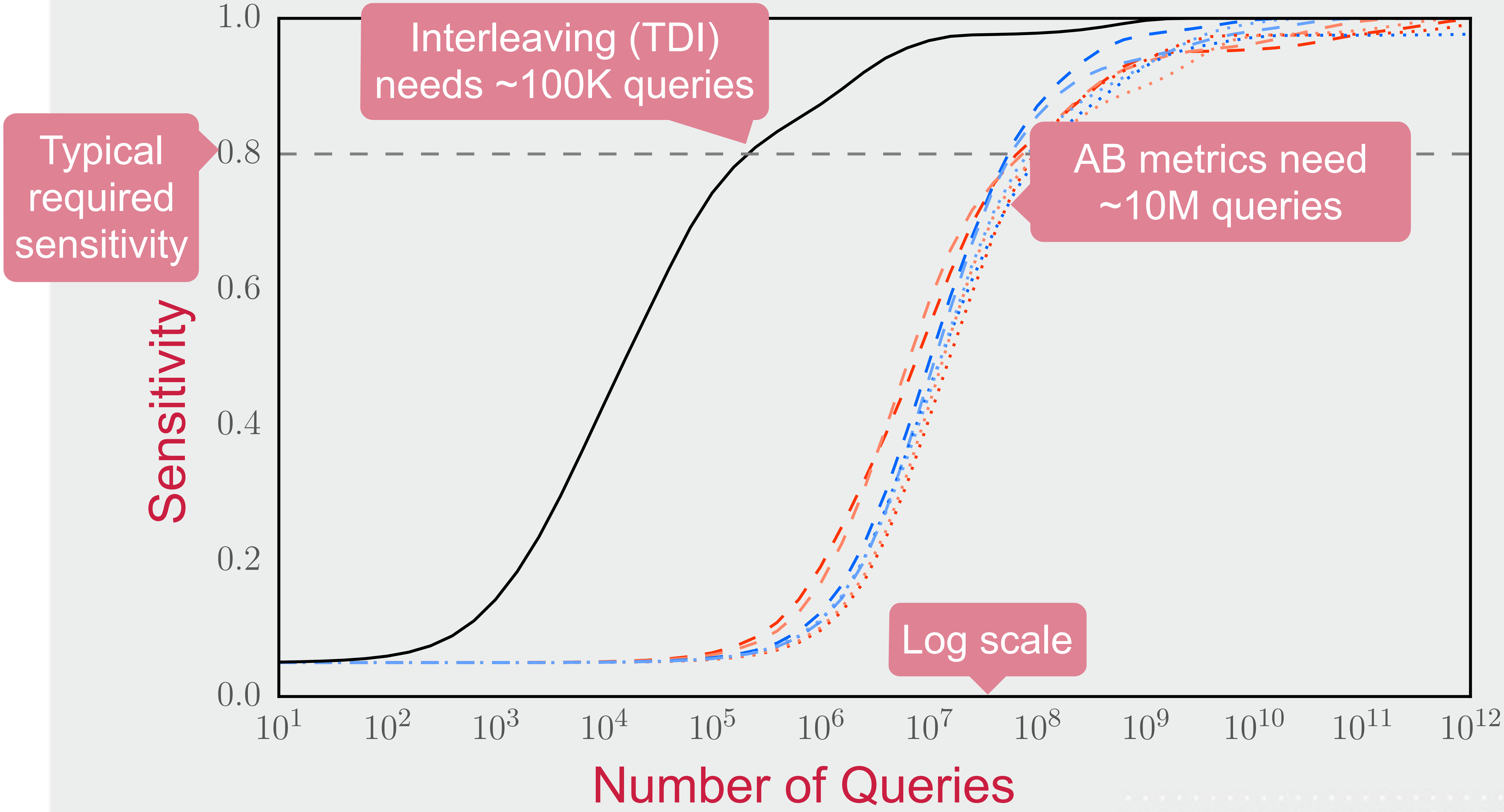


Interleaving (TDI) needs ~100K queries

Typical required sensitivity

Log scale

Data - Analysis - **Sensitivity**



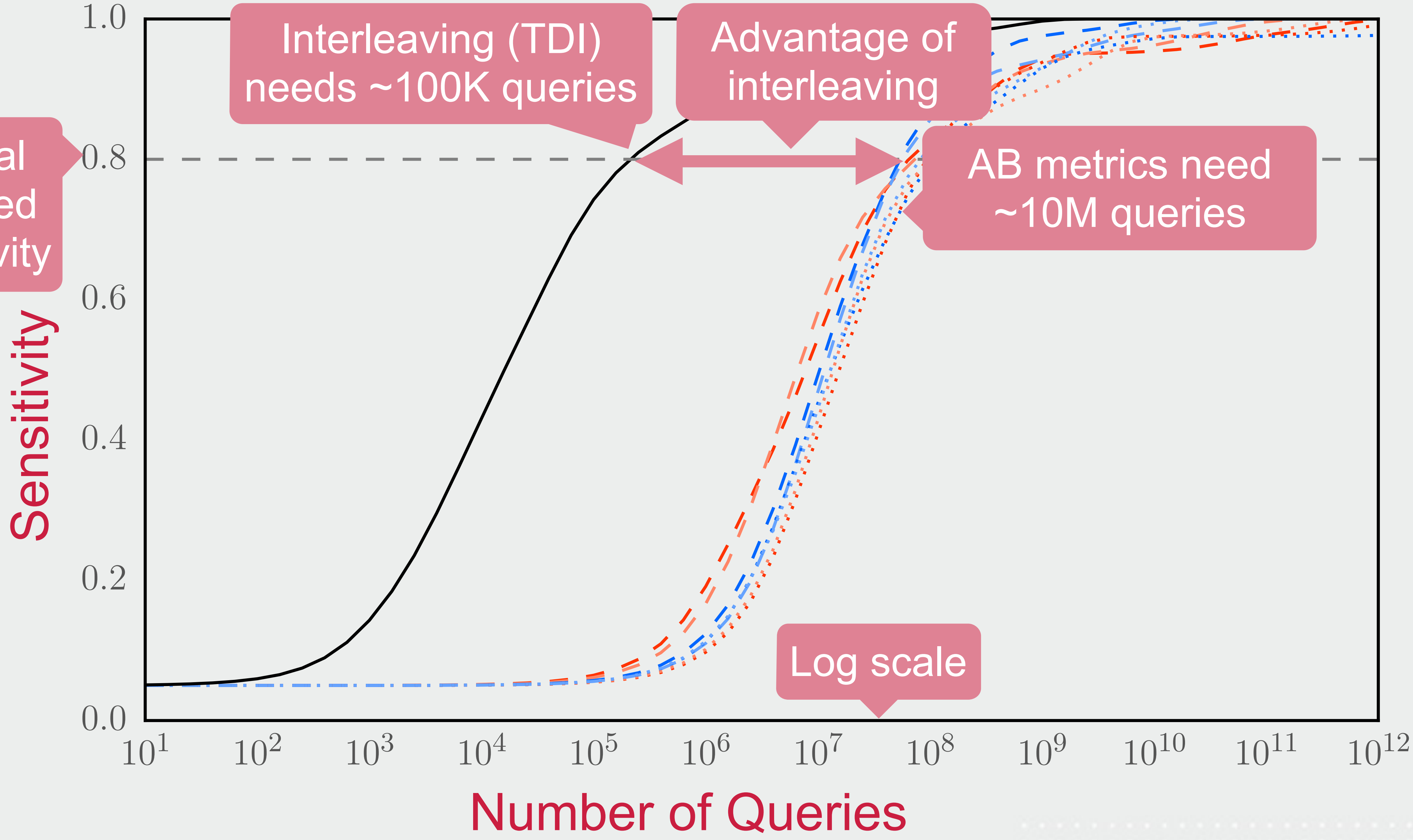
Interleaving (TDI) needs ~100K queries

Typical required sensitivity

AB metrics need ~10M queries

Log scale

Data - Analysis - **Sensitivity**



Typical required sensitivity

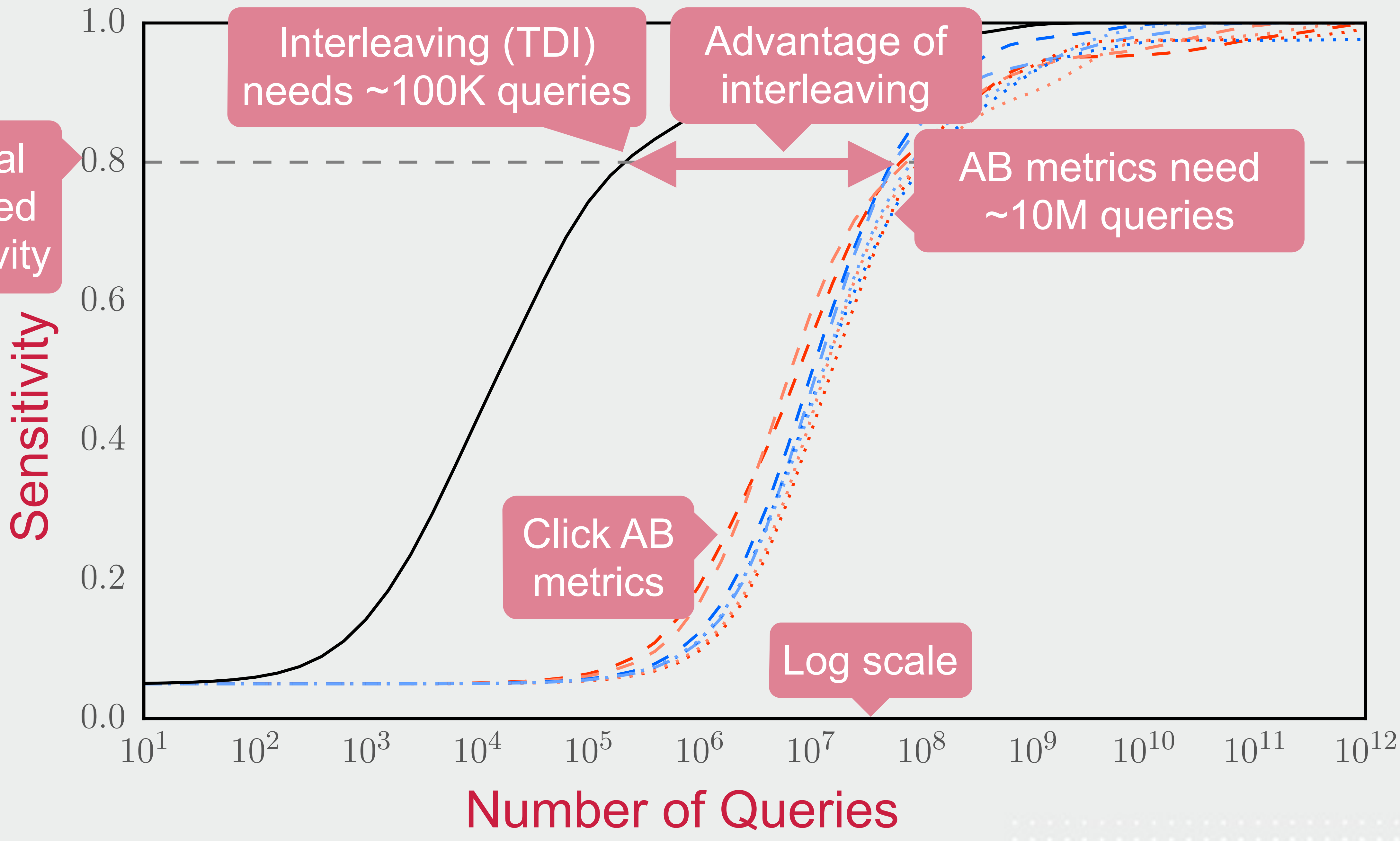
Interleaving (TDI) needs ~100K queries

Advantage of interleaving

AB metrics need ~10M queries

Log scale

Data - Analysis - **Sensitivity**



Typical required sensitivity

Interleaving (TDI) needs ~100K queries

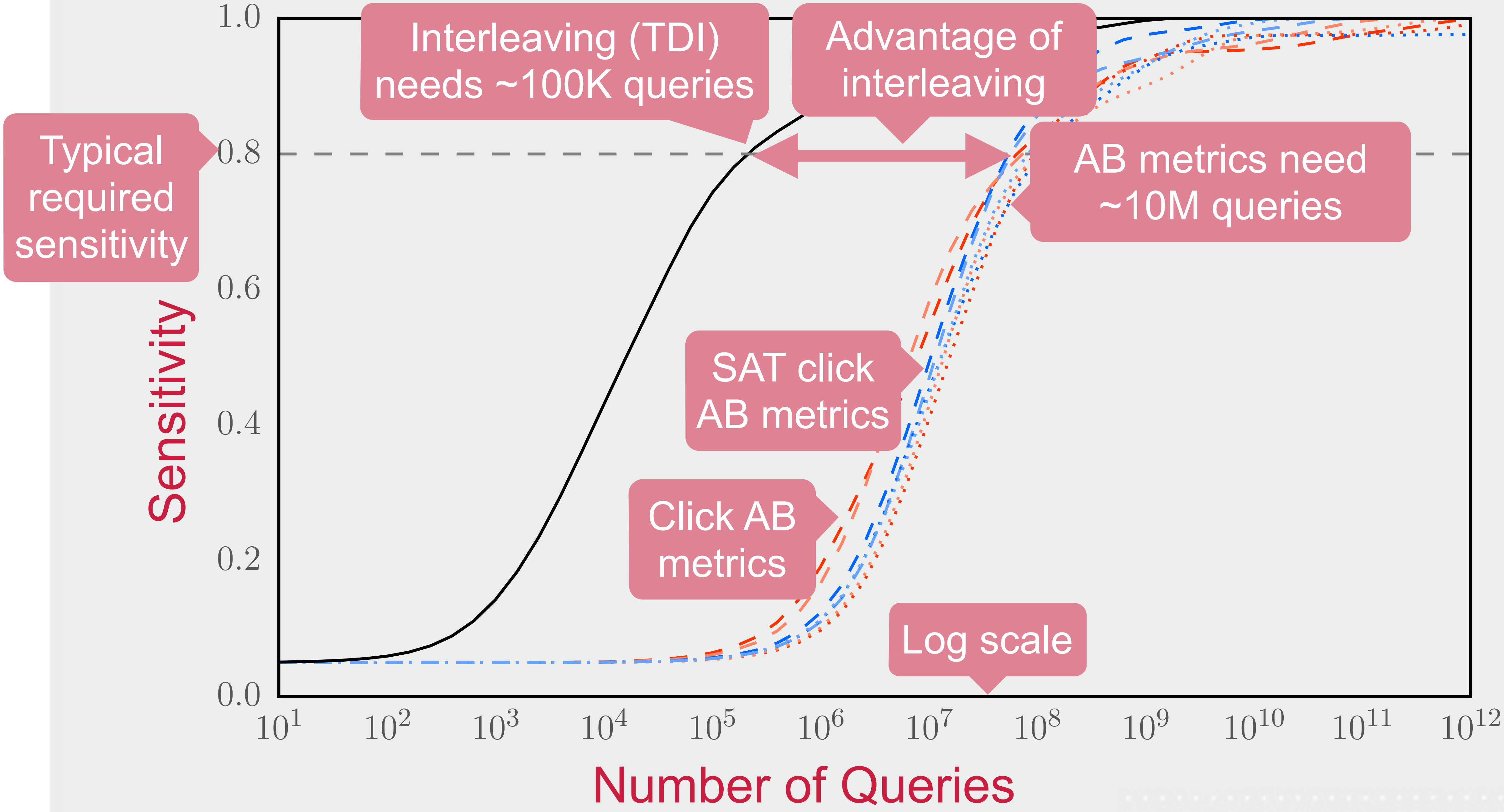
Advantage of interleaving

AB metrics need ~10M queries

Click AB metrics

Log scale

Data - Analysis - **Sensitivity**



Typical required sensitivity

Interleaving (TDI) needs ~100K queries

Advantage of interleaving

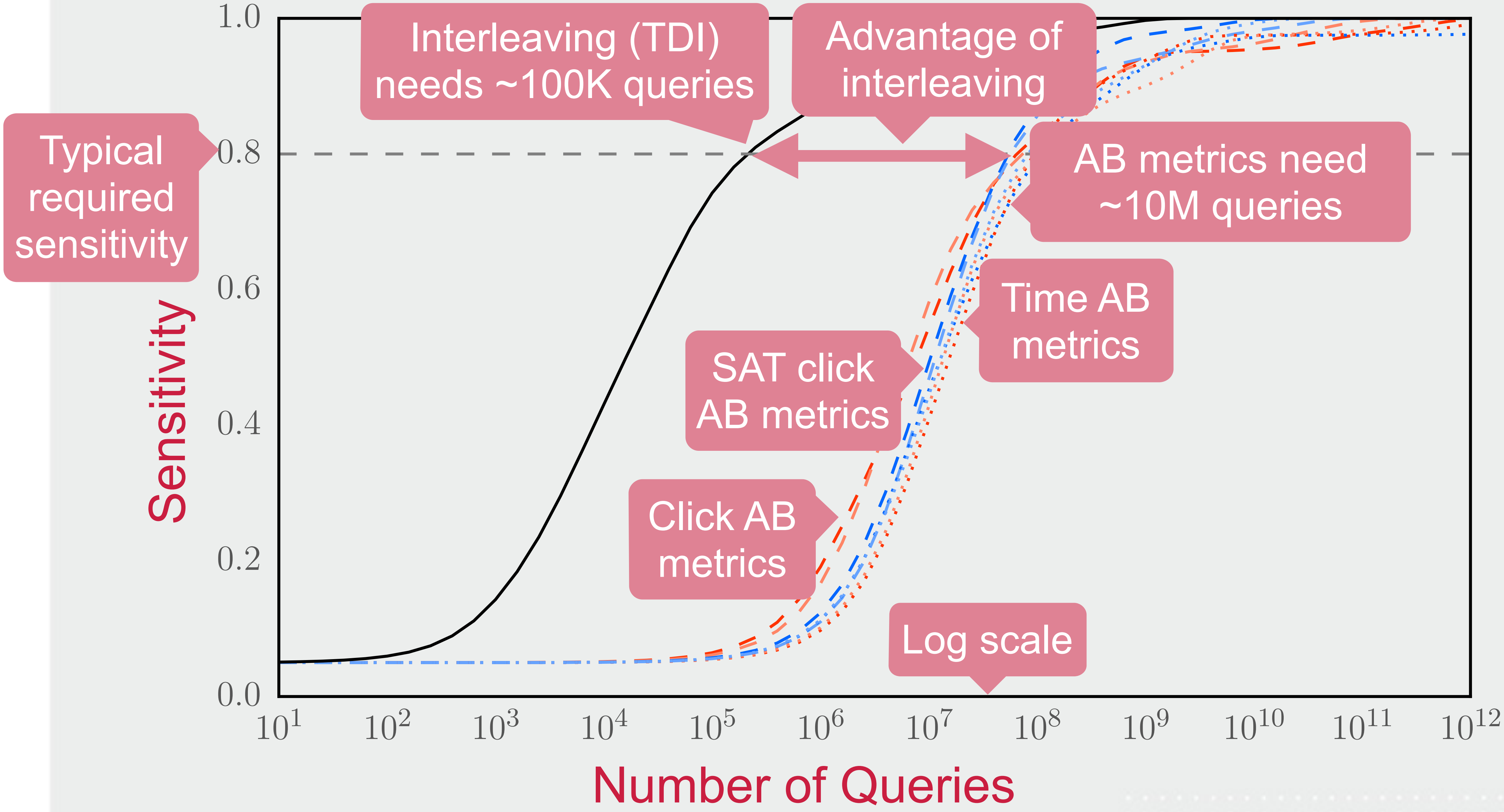
AB metrics need ~10M queries

SAT click AB metrics

Click AB metrics

Log scale

Data - Analysis - **Sensitivity**



Typical required sensitivity

Interleaving (TDI) needs ~100K queries

Advantage of interleaving

AB metrics need ~10M queries

Time AB metrics

SAT click AB metrics

Click AB metrics

Log scale

Data - Analysis - **Summary**

Data - Analysis - **Summary**

- **AB Testing** has **low sensitivity**

Data - Analysis - **Summary**

- **AB Testing** has **low sensitivity**
- **Interleaving** (TDI) has **high sensitivity** (10-100x AB)

Data - Analysis - **Summary**

- **AB Testing** has **low sensitivity**
- **Interleaving** (TDI) has **high sensitivity** (10-100x AB)
- **Interleaving** (TDI) has **low agreement** with AB metrics

Data - Analysis - **Summary**

- **AB Testing** has **low sensitivity**
- **Interleaving** (TDI) has **high sensitivity** (10-100x AB)
- **Interleaving** (TDI) has **low agreement** with AB metrics

We aim to

Improve interleaving (TDI) to **increase agreement** with a given AB metric while **maintaining sensitivity**

Data - Analysis - **Aim**

	Sensitivity (required #queries)	Agreement with AB (prefer same ranker)
AB Testing	~10M ⊗	~90% ✓

Data - Analysis - **Aim**

	Sensitivity (required #queries)	Agreement with AB (prefer same ranker)
AB Testing	~10M ✗	~90% ✓
Interleaving (TDI)	~100K ✓	~60% ✗

Data - Analysis - **Aim**

	Sensitivity (required #queries)	Agreement with AB (prefer same ranker)
AB Testing	~10M ✗	~90% ✓
Interleaving (TDI)	~100K ✓	~60% ✗
Improved Interleaving (TDI)	**~100K ?** ✓	**~90% ?** ✓

Outline

Motivation

Data + analysis

Methods + results

Conclusions

Methods

1. **Matching AB Metrics**
2. Parameterized Credit Functions
3. Combined Credit Functions

Methods - **Matching AB Metric**

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …
- … we propose to **match AB metrics**

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …
- … we propose to **match AB metrics**
  - Count only **certain** clicks

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …
- … we propose to **match AB metrics**
  - Count only **certain** clicks
    - @1

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …
- … we propose to **match AB metrics**
  - Count only **certain** clicks
    - @1
    - SAT

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …
- … we propose to **match AB metrics**
  - Count only **certain** clicks
    - @1
    - SAT

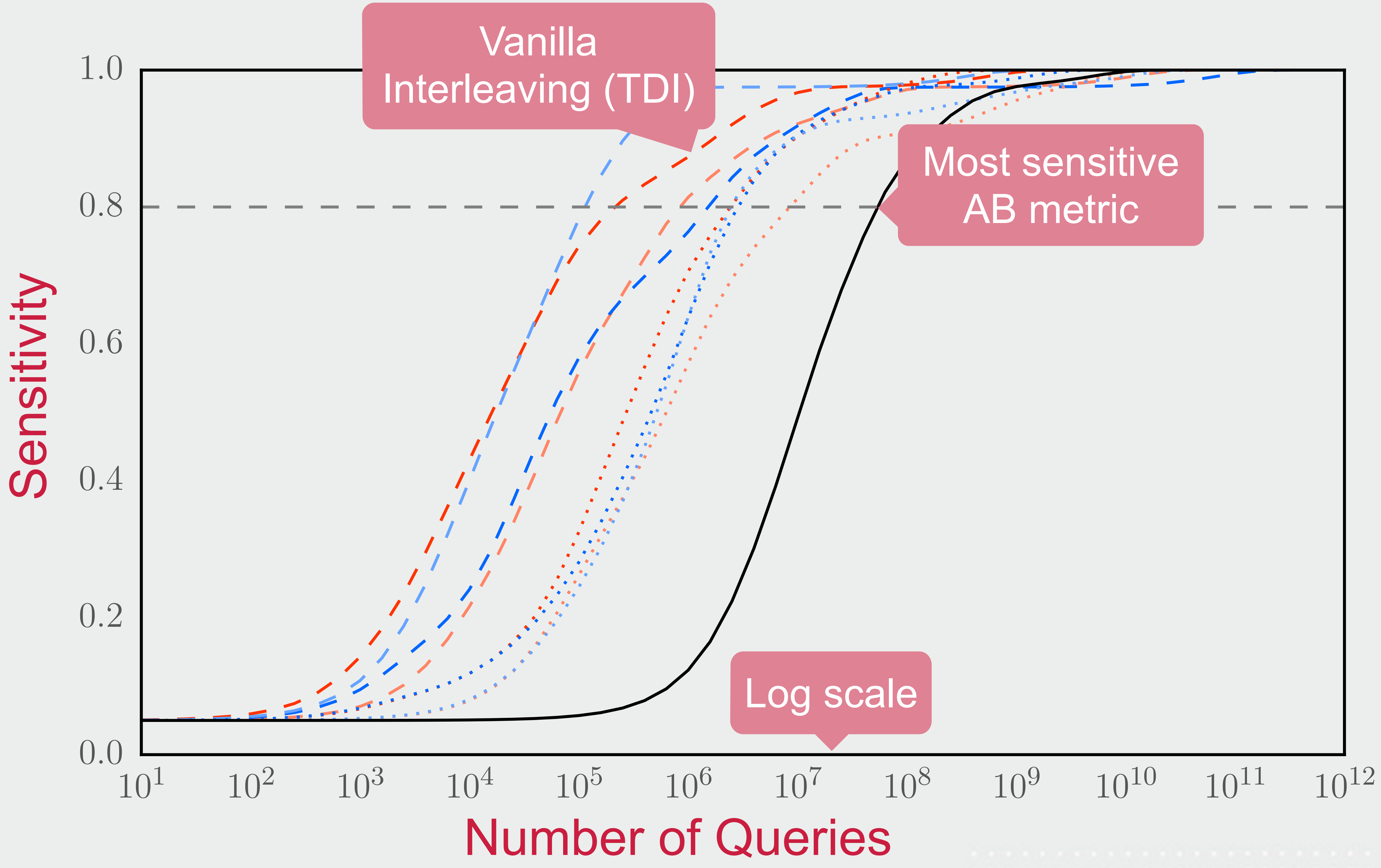
Filter out clicks,
can reduce sensitivity

Methods - **Matching AB Metric**

- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …
- … we propose to **match AB metrics**
  - Count only **certain** clicks
    - @1
    - SAT
  - Measure **time** to click

Filter out clicks,
can reduce sensitivity

Methods - Matching AB Metric - **Sensitivity**

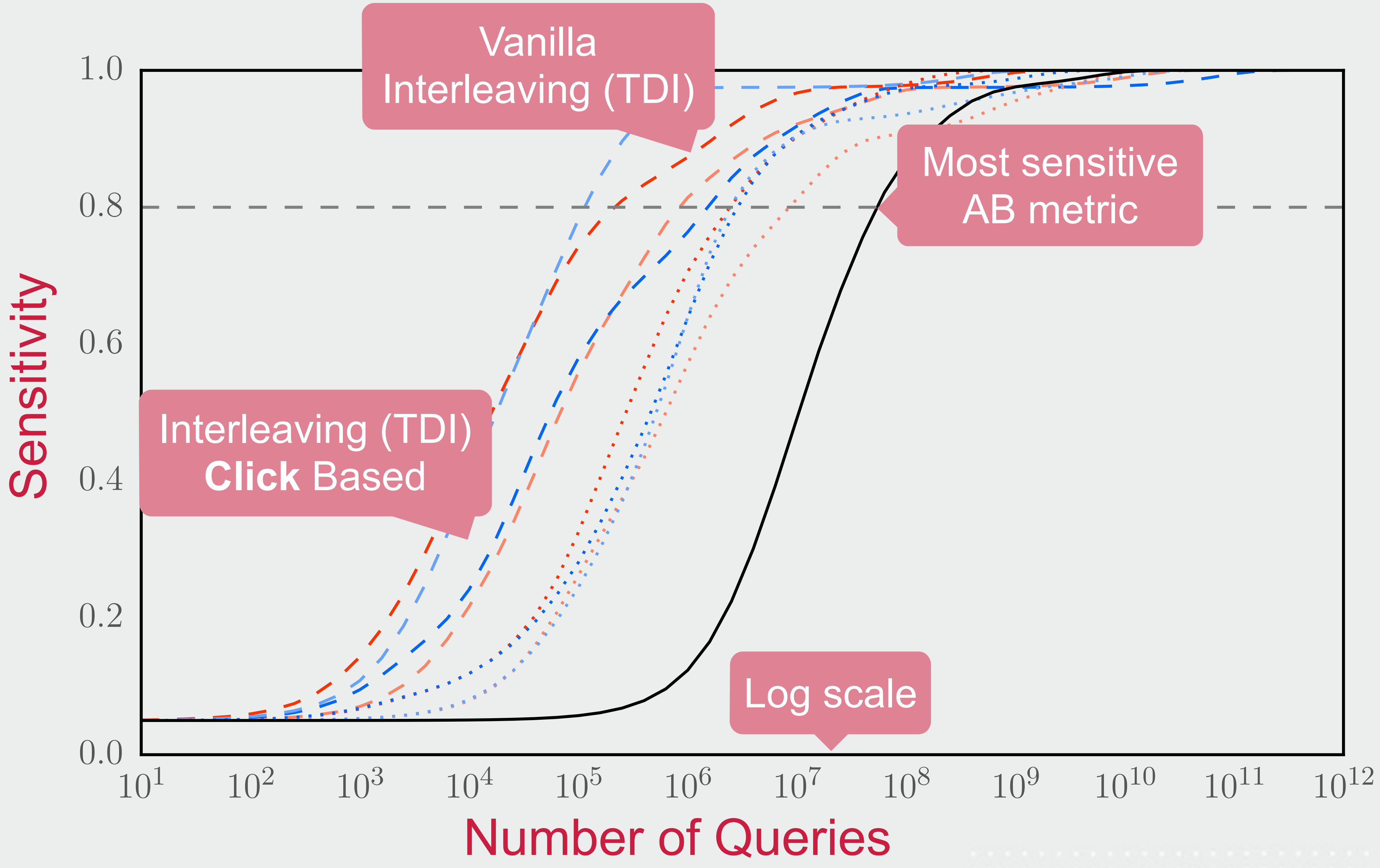


Vanilla Interleaving (TDI)

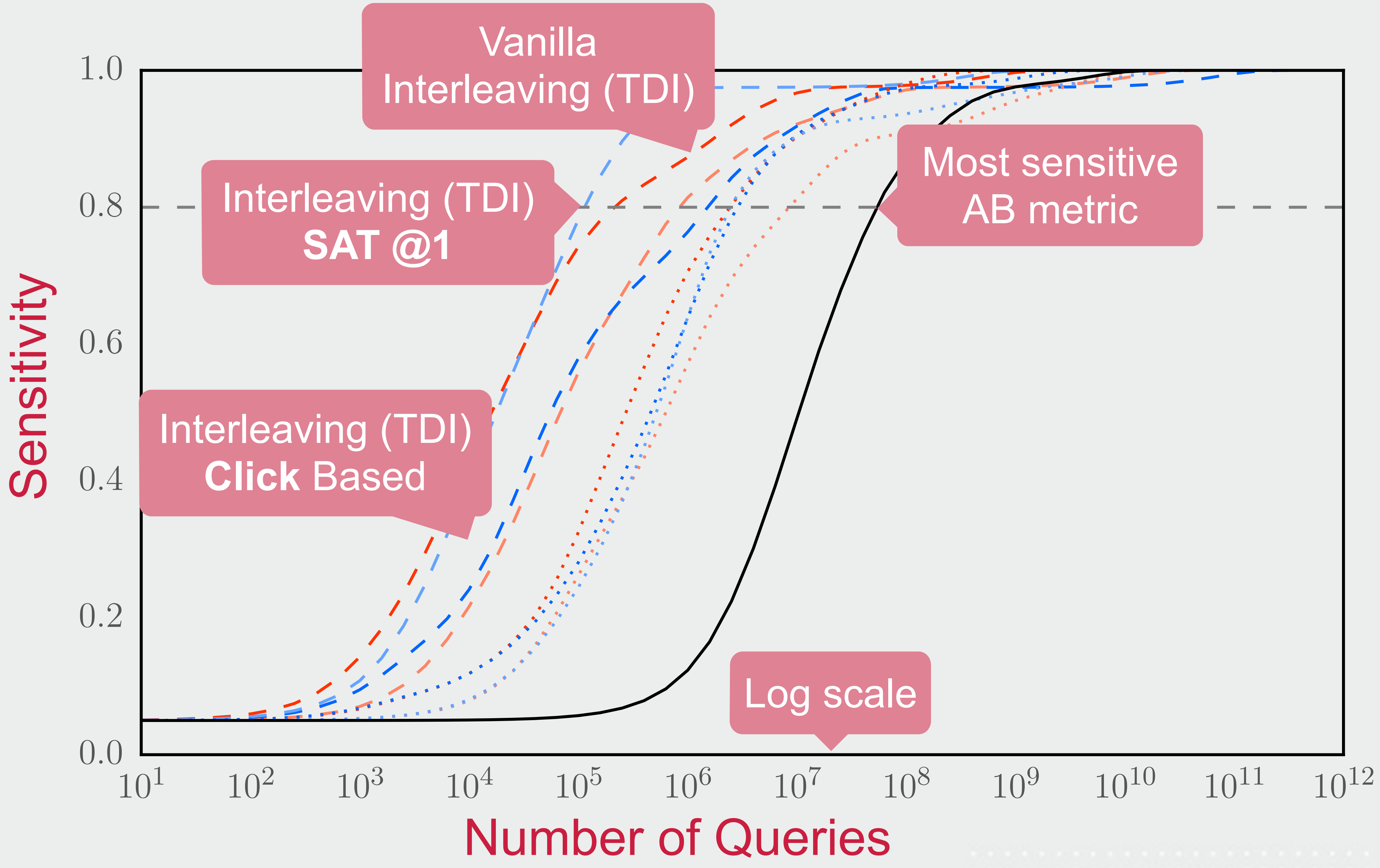
Most sensitive AB metric

Log scale

Methods - Matching AB Metric - **Sensitivity**



Methods - Matching AB Metric - **Sensitivity**



Vanilla Interleaving (TDI)

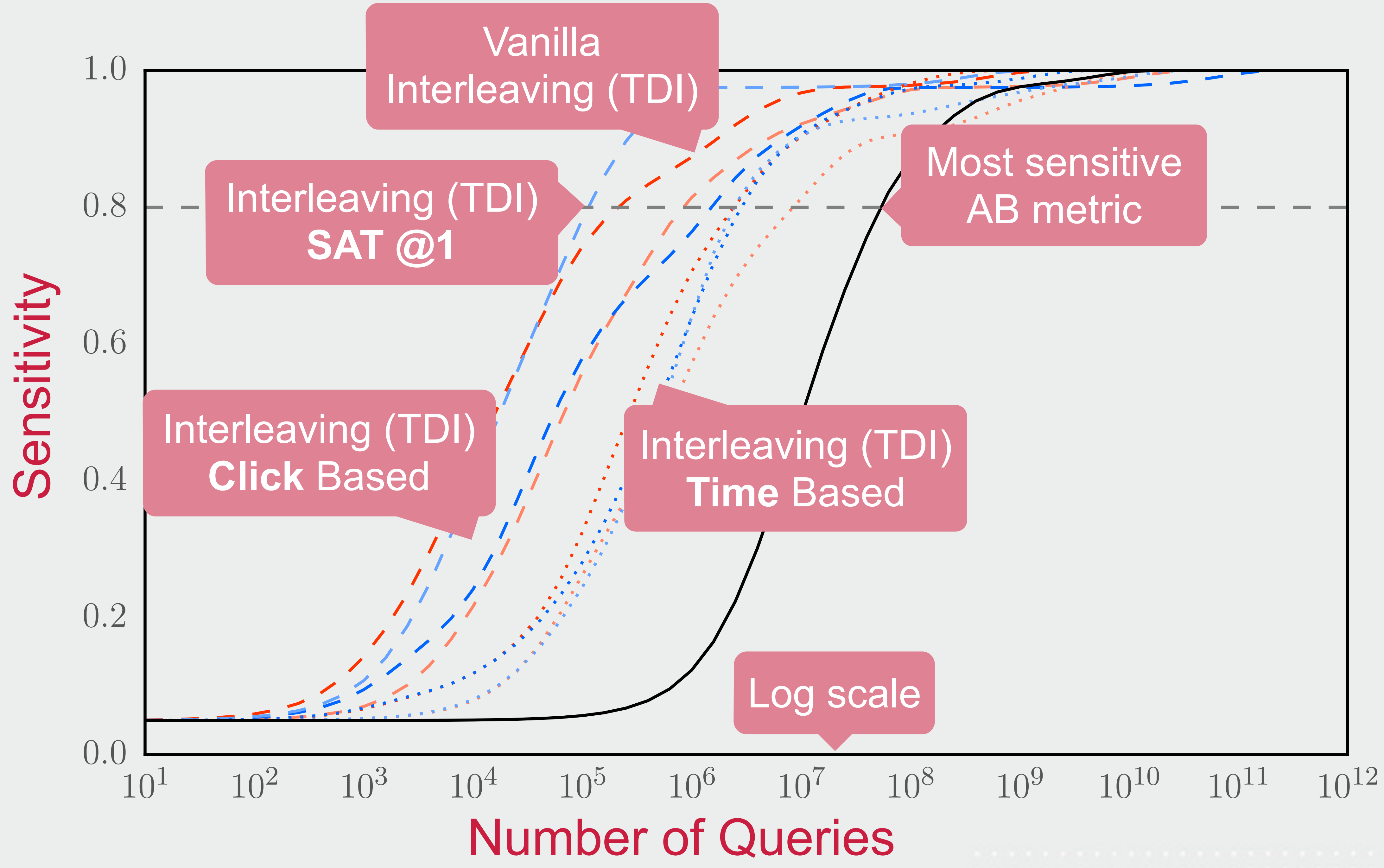
Interleaving (TDI) **SAT @1**

Interleaving (TDI) **Click** Based

Most sensitive AB metric

Log scale

Methods - Matching AB Metric - **Sensitivity**



Vanilla Interleaving (TDI)

Interleaving (TDI) **SAT @1**

Most sensitive AB metric

Interleaving (TDI) **Click** Based

Interleaving (TDI) **Time** Based

Log scale

Methods - Matching AB metric - **Agreement**

Vanilla interleaving

	TDI
AB	0.63
AB@1	0.71
AB_S	0.71
AB_S@1	0.76
AB_T	0.53
AB_T@1	0.45
$AB_{T,S}$	0.47
$AB_{T,S}$@1	0.42

matching AB metric

Methods - Matching AB metric - **Agreement**

Vanilla interleaving

	TDI	TDI@1	TDI$_S$	TDI$_S$@1	TDI$_T$	TDI$_T$@1	TDI$_{T,S}$	TDI$_{T,S}$@1
AB	0.63							
AB@1	0.71	0.68						
AB$_S$	0.71		**0.87**					
AB$_S$@1	0.76			0.63				
AB$_T$	0.53				**0.71**			
AB$_T$@1	0.45					0.58		
AB$_{T,S}$	0.47						0.58	
AB$_{T,S}$@1	0.42							0.58

Methods - Matching AB metric - **Agreement**

Vanilla interleaving

	TDI	TDI@1	TDI$_S$	TDI$_S$@1	TDI$_T$	TDI$_T$@1	TDI$_{T,S}$	TDI$_{T,S}$@1
AB	0.63	0.66	**0.84**	0.66	0.61	0.61	0.58	0.53
AB@1	0.71	0.68	**0.76**	0.63	0.63	0.47	0.55	0.55
AB$_S$	0.71	0.68	**0.87**	0.68	0.68	0.58	0.61	0.55
AB$_S$@1	0.76	0.68	**0.82**	0.63	0.74	0.53	0.61	0.50
AB$_T$	0.53	0.55	0.47	0.55	**0.71**	0.55	0.68	0.58
AB$_T$@1	0.45	0.47	0.45	0.58	**0.63**	0.58	0.61	0.62
AB$_{T,S}$	0.47	0.55	0.53	**0.71**	0.66	0.66	0.58	0.53
AB$_{T,S}$@1	0.42	0.50	0.53	**0.66**	0.61	**0.66**	0.58	0.58

Methods - Matching AB metric - **Agreement**

Vanilla interleaving

	TDI	TDI@1	TDI$_{S}$	TDI$_{S}$@1	TDI$_{T}$	TDI$_{T}$@1	TDI$_{T,S}$	TDI$_{T,S}$@1
AB	0.63	0.66	**0.84**	0.66	0.61	0.61	0.58	0.53
AB@1	0.71	0.68	**0.76**	0.63	0.63	0.47	0.55	0.55
AB$_{S}$	0.71	0.68	**0.87**	0.68	0.68	0.58	0.61	0.55
AB$_{S}$@1	0.76	0.68	**0.82**	0.63	0.74	0.53	0.61	0.50
AB$_{T}$	0.53	0.55	0.47	0.55	**0.71**	0.55	0.68	0.58
AB$_{T}$@1	0.45	0.47	0.45	0.58	**0.63**	0.58	0.61	0.62
AB$_{T,S}$	0.47	0.55	0.53	**0.71**	0.66	0.66	0.58	0.53
AB$_{T,S}$@1	0.42	0.50	0.53	**0.66**	0.61	**0.66**	0.58	0.58

Highest agreement not on diagonal

Methods

1. Matching AB Metrics
2. **Parameterized Credit Functions**
3. Combined Credit Functions

Methods - **Parametrized Credit**

Methods - **Parametrized Credit**

- We aim to increase agreement

Methods - **Parametrized Credit**

Remember, we have a model that predicts **SAT probability**

- We aim to increase agreement
- **Parameterize TDI** with a SAT threshold t_S
    - TDI_S^{ts} and $TDI_{T,S}^{ts}$

Methods - **Parametrized Credit**

Remember, we have a model that predicts **SAT probability**

- We aim to increase agreement
- **Parameterize TDI** with a SAT threshold t_s
  - TDI_S^{ts} and $TDI_{T,S}^{ts}$

Click based

Time based

Methods - **Parametrized Credit**

Remember, we have a model that predicts **SAT probability**

- We aim to increase agreement
- **Parameterize TDI** with a SAT threshold t_S
  - TDI_S^{ts} and $TDI_{T,S}^{ts}$

Click based

Time based

Filter out non SAT clicks, **can reduce** sensitivity

Methods - **Parametrized Credit**

Remember, we have a model that predicts **SAT probability**

- We aim to increase agreement
- **Parameterize TDI** with a SAT threshold t_s
  - TDI_S^{ts} and $TDI_{T,S}^{ts}$

Click based

Time based

Filter out non SAT clicks, **can reduce** sensitivity

- Find **optimal threshold t_s**
  - Maximize agreement for **each** AB metric

Methods - **Parametrized Credit**

Remember, we have a model that predicts **SAT probability**

- We aim to increase agreement
- **Parameterize TDI** with a SAT threshold t_s
  - TDI_S^{ts} and $TDI_{T,S}^{ts}$

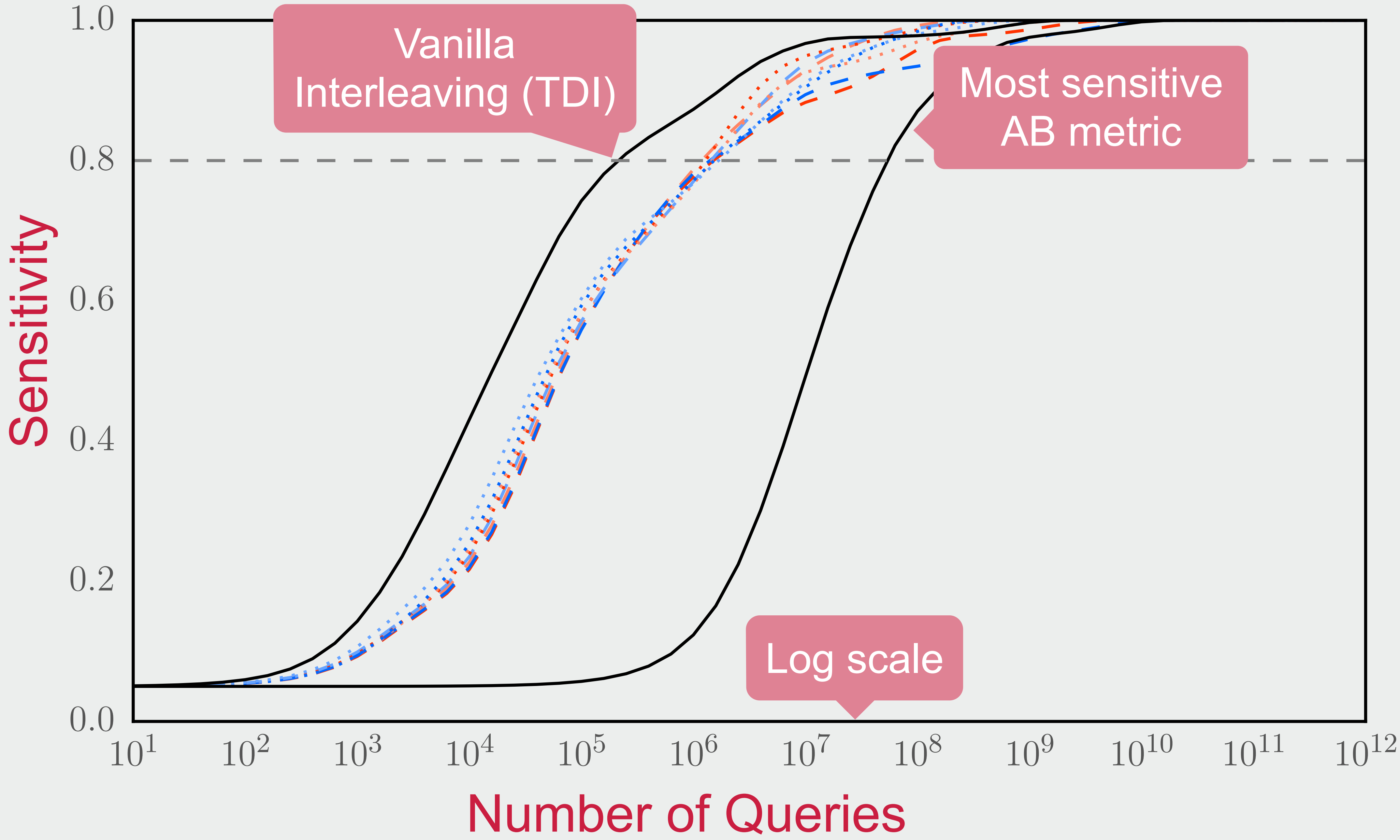
Click based

Time based

Filter out non SAT clicks, **can reduce** sensitivity

- Find **optimal threshold t_s**
  - Maximize agreement for **each** AB metric
- Repeat n=100 times:
  - Take bootstrap sample
  - Grid search to find t_s that maximizes agreement
  - Report performance on “out of bag” sample

Methods - Parametrized Credit - **Sensitivity**

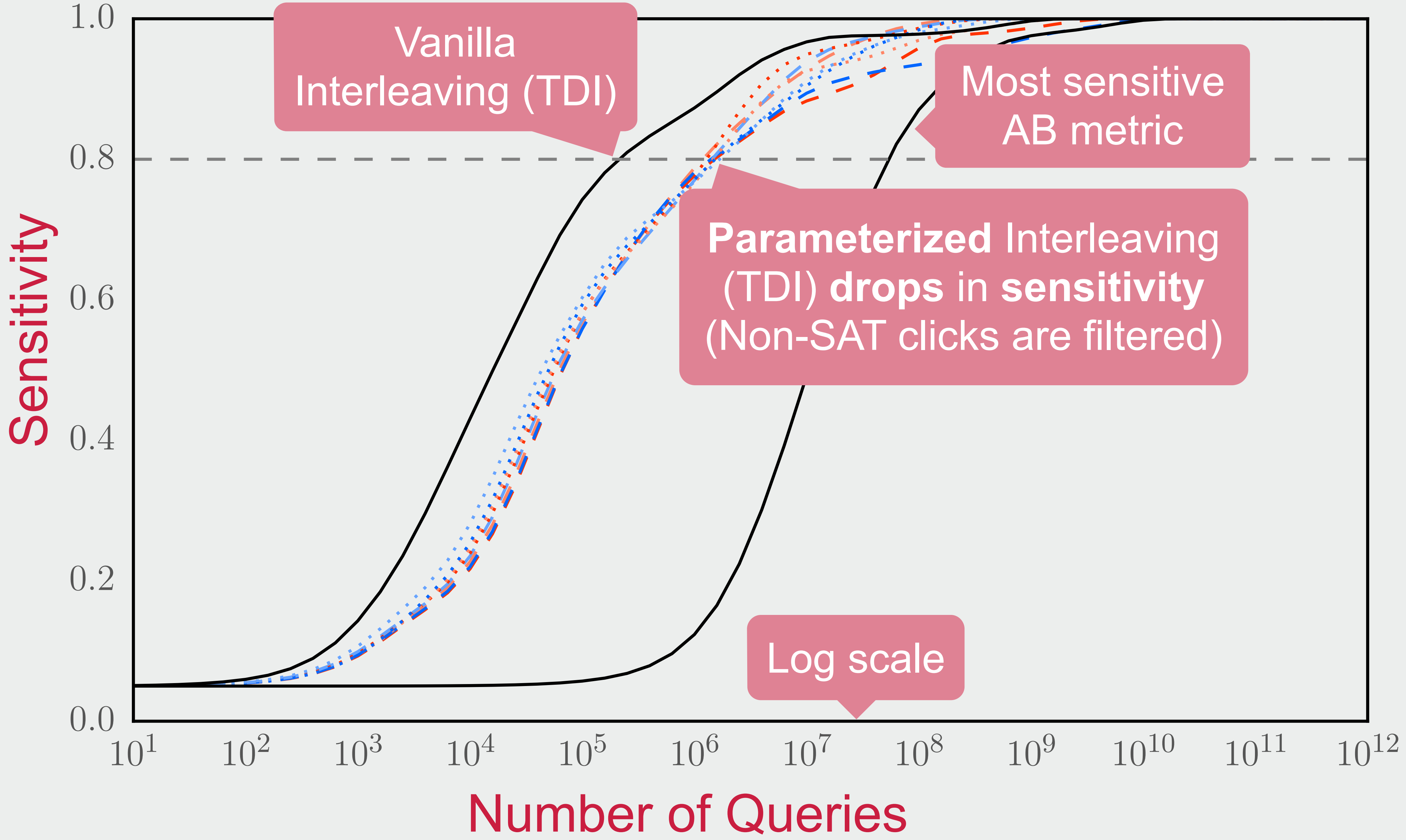


Vanilla Interleaving (TDI)

Most sensitive AB metric

Log scale

Methods - Parametrized Credit - **Sensitivity**



Vanilla Interleaving (TDI)

Most sensitive AB metric

Parameterized Interleaving (TDI) **drops** in **sensitivity** (Non-SAT clicks are filtered)

Log scale

Methods - Parametrized Credit - **Agreement**

	Vanilla
AB Metric	**TDI**
AB	0.63
AB@1	0.71
AB$_S$	0.71
AB$_S$@1	0.76
AB$_T$	0.53
AB$_T$@1	0.45
AB$_{T,S}$	0.47
AB$_{T,S}$@1	0.42

Methods - Parametrized Credit - **Agreement**

Vanilla

Click based

AB Metric	TDI	TDI_{S}^{ts}
AB	0.63	**0.82**
AB@1	0.71	
AB_{S}	0.71	
AB_{S}@1	0.76	
AB_{T}	0.53	
AB_{T}@1	0.45	
$\mathrm{AB}_{T,S}$	0.47	
$\mathrm{AB}_{T,S}$@1	0.42	

Methods - Parametrized Credit - **Agreement**

	Vanilla	Click based
AB Metric	**TDI**	**TDI_S^{ts}**
AB	0.63	**0.82**
AB@1	0.71	**0.79**
AB_S	0.71	**0.84**
AB_S@1	0.76	**0.84**
AB_T	0.53	0.47
AB_T@1	0.45	**0.49**
$\text{AB}_{T,S}$	0.47	0.46
$\text{AB}_{T,S}$@1	0.42	0.52

Methods - Parametrized Credit - **Agreement**

	Vanilla	Click based	Time based
AB Metric	**TDI**	**TDI_S^{ts}**	**$TDI_{T,S}^{ts}$**
AB	0.63	**0.82**	0.53
AB@1	0.71	**0.79**	0.54
AB_S	0.71	**0.84**	0.48
AB_S@1	0.76	**0.84**	0.48
AB_T	0.53	0.47	**0.67**
AB_T@1	0.45	**0.49**	**0.62**
$AB_{T,S}$	0.47	0.46	**0.61**
$AB_{T,S}$@1	0.42	0.52	**0.62**

Methods - Parametrized Credit - **Agreement**

	Vanilla	Click based	Time based
AB Metric	**TDI**	**TDI_S^{ts}**	**$TDI_{T,S}^{ts}$**
AB	0.63	**0.82**	0.53
AB@1	0.71	**0.79**	0.54
AB_S	0.71	**0.84**	0.48
AB_S@1	0.76	**0.84**	0.48
AB_T	0.53	0.47	**0.67**
AB_T@1	0.45	**0.49**	**0.62**
$AB_{T,S}$	0.47	0.46	**0.61**
$AB_{T,S}$@1	0.42	0.52	**0.62**

Methods

1. Matching AB Metrics
2. Parameterized Credit Functions
3. **Combined Credit Functions**

Methods - **Combined Credit**

Methods - **Combined Credit**

- **Combine** parameterized **credit functions**
  - $w_S \cdot \mathrm{TDI}_S^{ts} + w_T \cdot \mathrm{TDI}_{T,S}^{ts}$

Click weight

Time weight

Methods - **Combined Credit**

- **Combine** parameterized **credit functions**
  - $w_S \cdot TDI_S^{ts} + w_T \cdot TDI_{T,S}^{ts}$

Click weight

Time weight

- Find optimal weights
  - Maximizing agreement

Methods - **Combined Credit**

- **Combine** parameterized **credit functions**
  - $w_S \cdot TDI_S^{ts} + w_T \cdot TDI_{T,S}^{ts}$

Click weight

Time weight

- Find optimal weights
  - Maximizing agreement
- Using the same maximization procedure
  - Bootstrap sample, parameter sweep

Methods - Combined Credit - **Agreement**

AB Metric	**TDI**
AB	0.63
AB@1	0.71
AB$_{S}$	0.71
AB$_{S}$@1	0.76
AB$_{T}$	0.53
AB$_{T}$@1	0.45
AB$_{T,S}$	0.47
AB$_{T,S}$@1	0.42

Methods - Combined Credit - **Agreement**

		$TDI_{T,S}^{W}$	Click weight	Time weight
AB Metric	**TDI**	**agreement**	w_S	w_T
AB	0.63	**0.84**	*1.00*	*0.00*
AB@1	0.71			
AB_S	0.71			
AB_S@1	0.76			
AB_T	0.53			
AB_T@1	0.45			
$AB_{T,S}$	0.47			
$AB_{T,S}$@1	0.42			

Methods - Combined Credit - **Agreement**

		$TDI_{T,S}^{W}$	Click weight	Time weight
AB Metric	**TDI**	**agreement**	w_S	w_T
AB	0.63	**0.84**	*1.00*	*0.00*
AB@1	0.71	**0.75**	*1.00*	*0.05*
AB_S	0.71	**0.85**	*1.00*	*0.00*
AB_S@1	0.76	**0.83**	*1.00*	*0.02*
AB_T	0.53	**0.68**	*0.99*	*0.90*
AB_T@1	0.45	**0.56**	*0.96*	*0.79*
$AB_{T,S}$	0.47	**0.63**	*0.91*	*0.88*
$AB_{T,S}$@1	0.42	**0.50**	*0.06*	*0.25*

Methods - Combined Credit - **Agreement**

		$TDI_{T,S}^{W}$	Click weight	Time weight
AB Metric	**TDI**	**agreement**	w_S	w_T
AB	0.63	**0.84**	*1.00*	*0.00*
AB@1	0.71	**0.75**	*1.00*	*0.05*
AB_S	0.71	**0.85**	*1.00*	*0.00*
AB_S@1	0.76	**0.83**	*1.00*	*0.02*
AB_T	0.53	**0.68**	*0.99*	*0.90*
AB_T@1	0.45	**0.56**	*0.96*	*0.79*
$AB_{T,S}$	0.47	**0.63**	*0.91*	*0.88*
$AB_{T,S}$@1	0.42	**0.50**	*0.06*	*0.25*

Methods - Combined Credit - **Agreement**

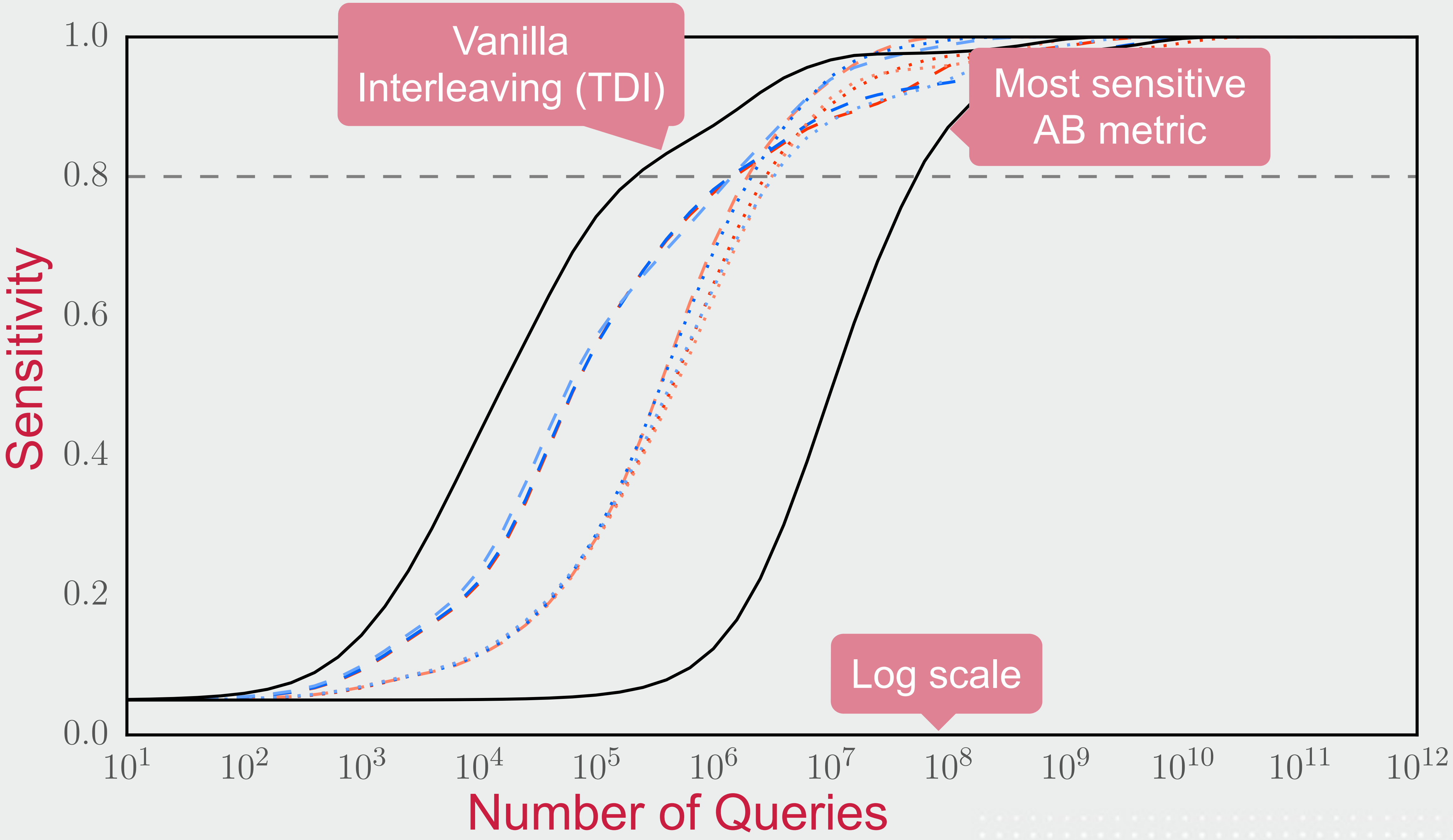
		$TDI_{T,S}^{W}$	Click weight	Time weight
AB Metric	TDI	agreement	w_S	w_T
AB	0.63	**0.84**	*1.00*	*0.00*
AB@1	0.71	**0.75**	*1.00*	*0.05*
AB_S	0.71	**0.85**	*1.00*	*0.00*
AB_S@1	0.76	**0.83**	*1.00*	*0.02*
AB_T	0.53	**0.68**	*0.99*	*0.90*
AB_T@1	0.45	**0.56**	*0.96*	*0.79*
$AB_{T,S}$	0.47	**0.63**	*0.91*	*0.88*
$AB_{T,S}$@1	0.42	**0.50**	*0.06*	*0.25*

Methods - Combined Credit - **Agreement**

		$\mathrm{TDI}_{T,S}^{W}$	Click weight	Time weight
AB Metric	TDI	agreement	w_S	w_T
AB	0.63	**0.84**	*1.00*	*0.00*
AB@1	0.71	**0.75**	*1.00*	*0.05*
AB_S	0.71	**0.85**	*1.00*	*0.00*
AB_S@1	0.76	**0.83**	*1.00*	*0.02*
AB_T	0.53	**0.68**	*0.99*	*0.90*
AB_T@1	0.45	**0.56**	*0.96*	*0.79*
$\mathrm{AB}_{T,S}$	0.47	**0.63**	*0.91*	*0.88*
$\mathrm{AB}_{T,S}$@1	0.42	**0.50**	*0.06*	*0.25*

All significantly better

Methods - Combined Credit - **Sensitivity**

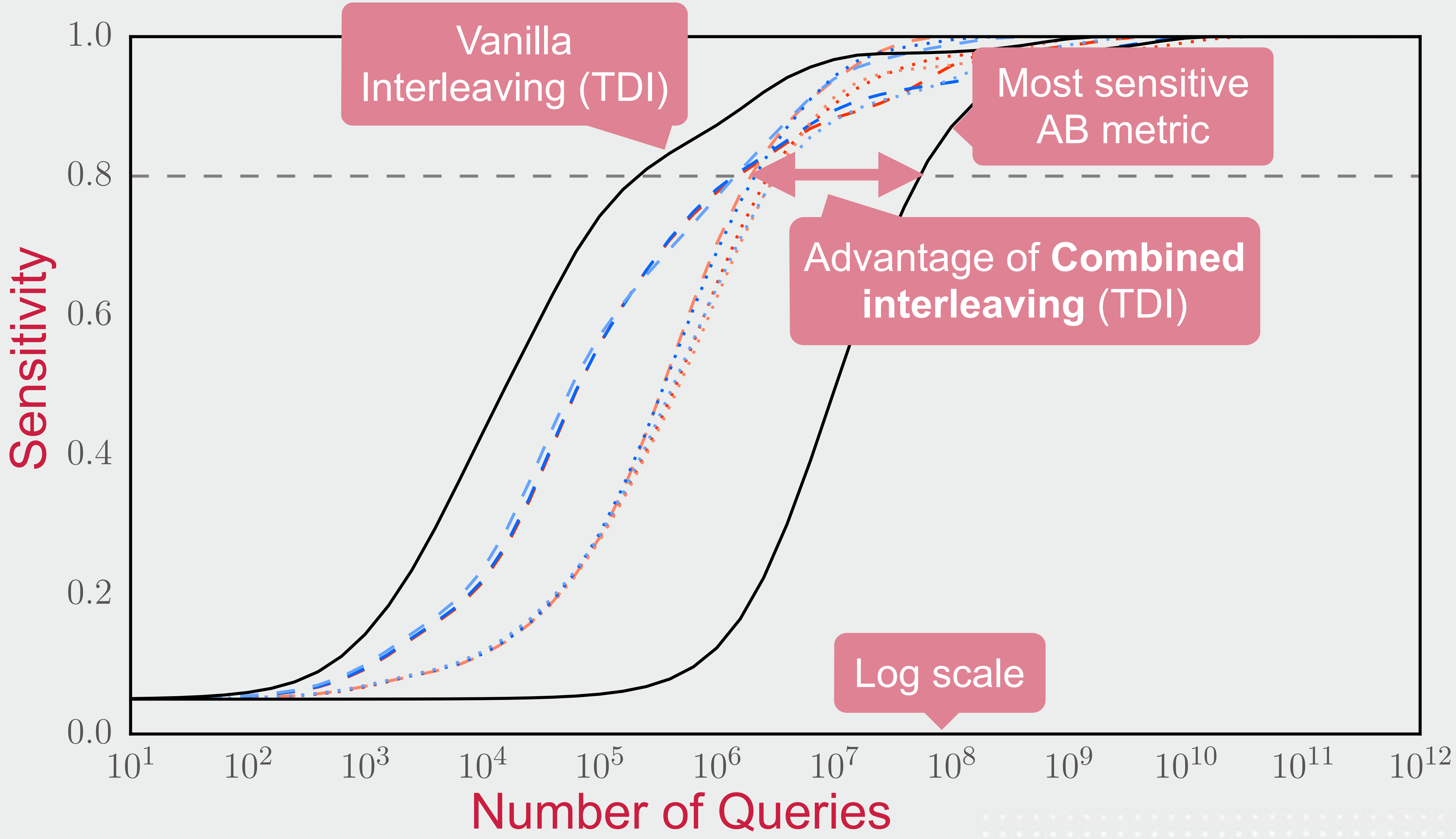


Vanilla Interleaving (TDI)

Most sensitive AB metric

Log scale

Methods - Combined Credit - **Sensitivity**



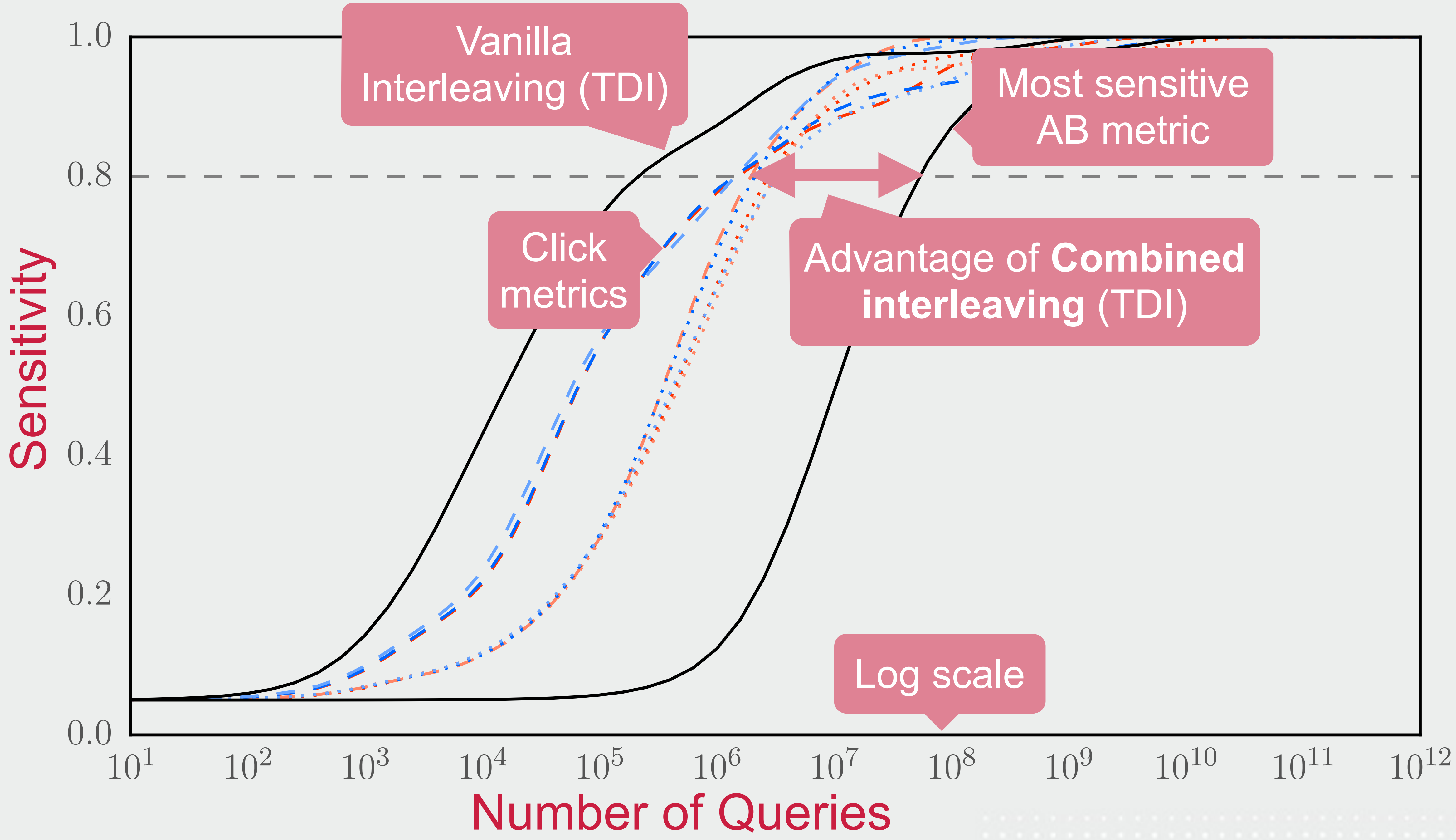
Vanilla Interleaving (TDI)

Most sensitive AB metric

Advantage of **Combined interleaving** (TDI)

Log scale

Methods - Combined Credit - **Sensitivity**



Vanilla Interleaving (TDI)

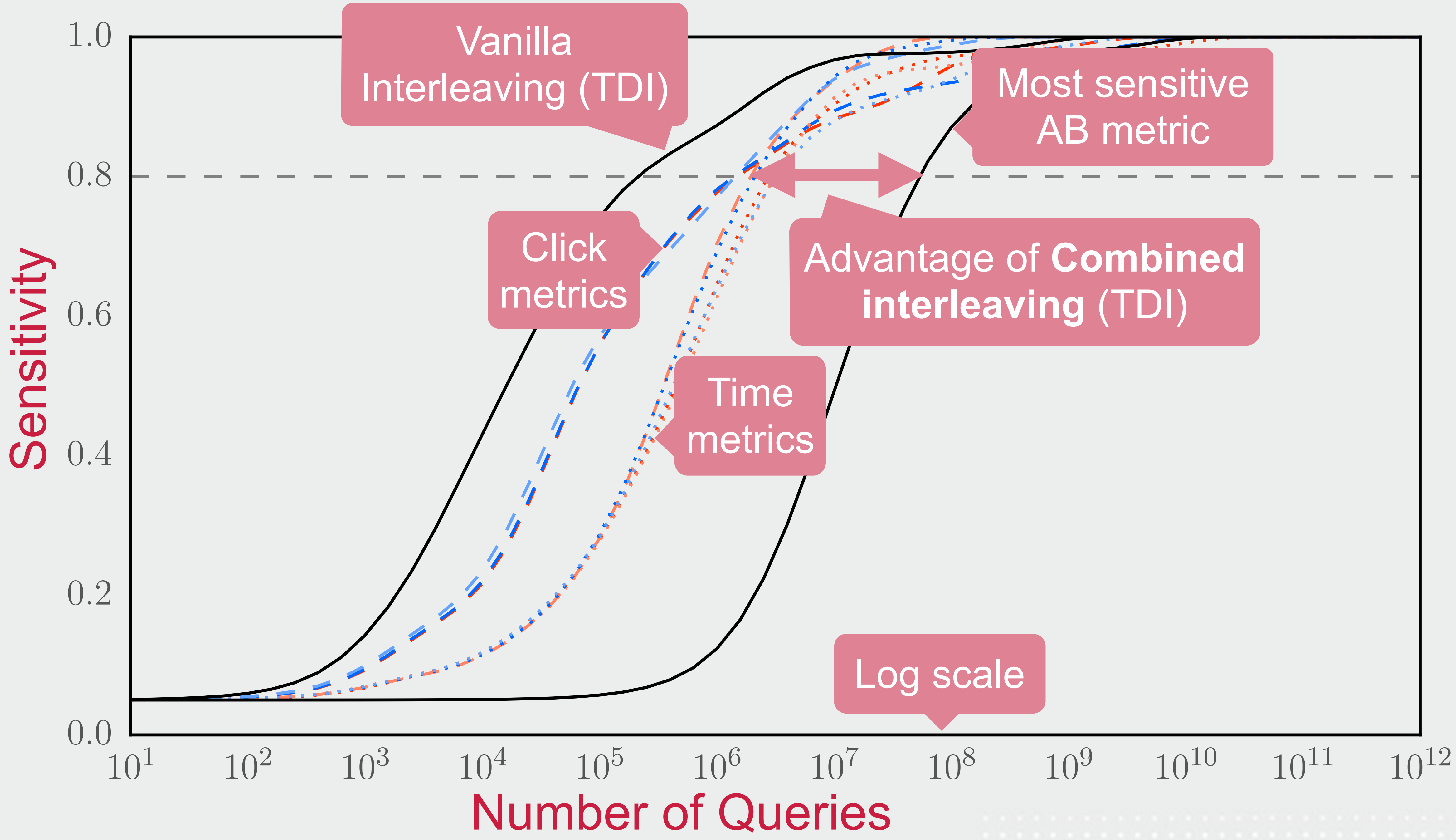
Most sensitive AB metric

Click metrics

Advantage of **Combined interleaving** (TDI)

Log scale

Methods - Combined Credit - **Sensitivity**



Vanilla Interleaving (TDI)

Most sensitive AB metric

Click metrics

Advantage of **Combined interleaving** (TDI)

Time metrics

Log scale

Outline

Motivation

Data + analysis

Methods + results

Conclusions

Conclusions - **Data Analysis**

Conclusions - **Data Analysis**

Confirming earlier findings

- Sensitivity:
  - **AB Testing** is 10-100x **less sensitive than Interleaving**

Conclusions - **Data Analysis**

Confirming earlier findings

- Sensitivity:
  - **AB Testing** is 10-100x **less sensitive than Interleaving**

New insight

- Agreement
  - **Between AB Testing** and **Interleaving** (TDI) is **low:** <76%

Conclusions - **Methods**

Conclusions - **Methods**

- Interleaving (TDI) with credit **matching** AB metrics
  - **Unpredictable**

Conclusions - **Methods**

- Interleaving (TDI) with credit **matching** AB metrics
  - **Unpredictable**

- Interleaving (TDI) with **parameterized** credit functions
  - Improvements for **some** AB metrics

Conclusions - **Methods**

- Interleaving (TDI) with credit **matching** AB metrics
  - **Unpredictable**
- Interleaving (TDI) with **parameterized** credit functions
  - Improvements for **some** AB metrics
- Interleaving (TDI) with **combined** credit functions
  - Improvements for **all** AB metrics

Conclusions - **Future Work**

Conclusions - **Future Work**

- Consider **even richer user signals** (sessions, task level features)

Conclusions - **Future Work**

- Consider **even richer user signals** (sessions, task level features)
- Take **magnitude** and **uncertainty** of AB metric differences into account

Conclusions - **Future Work**

- Consider **even richer user signals** (sessions, task level features)
- Take **magnitude** and **uncertainty** of AB metric differences into account
- Understanding of **where and why agreement is low or high**

Take Away

Take Away

- **Rich** user **signals** in **interleaving**

Take Away

- **Rich** user **signals** in **interleaving**
- **Agreement** of Interleaving with an AB metric can be made as high as **87%**

Take Away

- **Rich** user **signals** in **interleaving**
- **Agreement** of Interleaving with an AB metric can be made as high as **87%**
- While maintaining **high sensitivity** of Interleaving

Take Away

- **Rich** user **signals** in **interleaving**
- **Agreement** of Interleaving with an AB metric can be made as high as **87%**
- While maintaining **high sensitivity** of Interleaving

Supported by ACM SIGIR Travel Grant

- Microsoft Research
- UNIVERSITY OF AMSTERDAM
- http://anneschuth.nl
- @anneschuth