UNIVERSITY OF AMSTERDAM

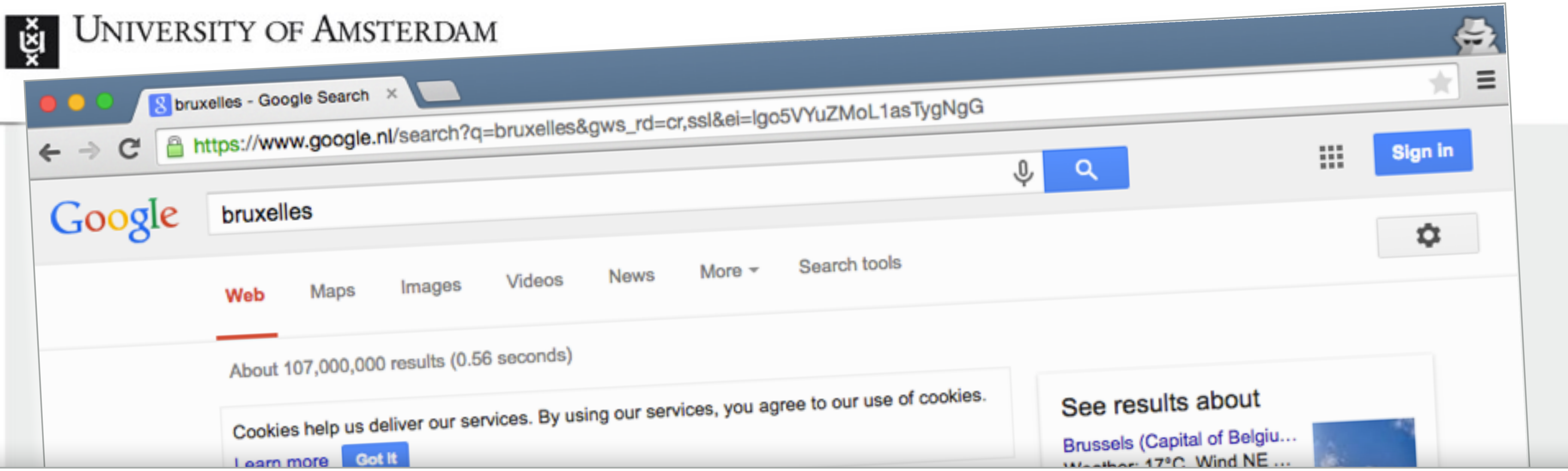
Predicting Search Satisfaction Metrics with Interleaved Comparisons

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

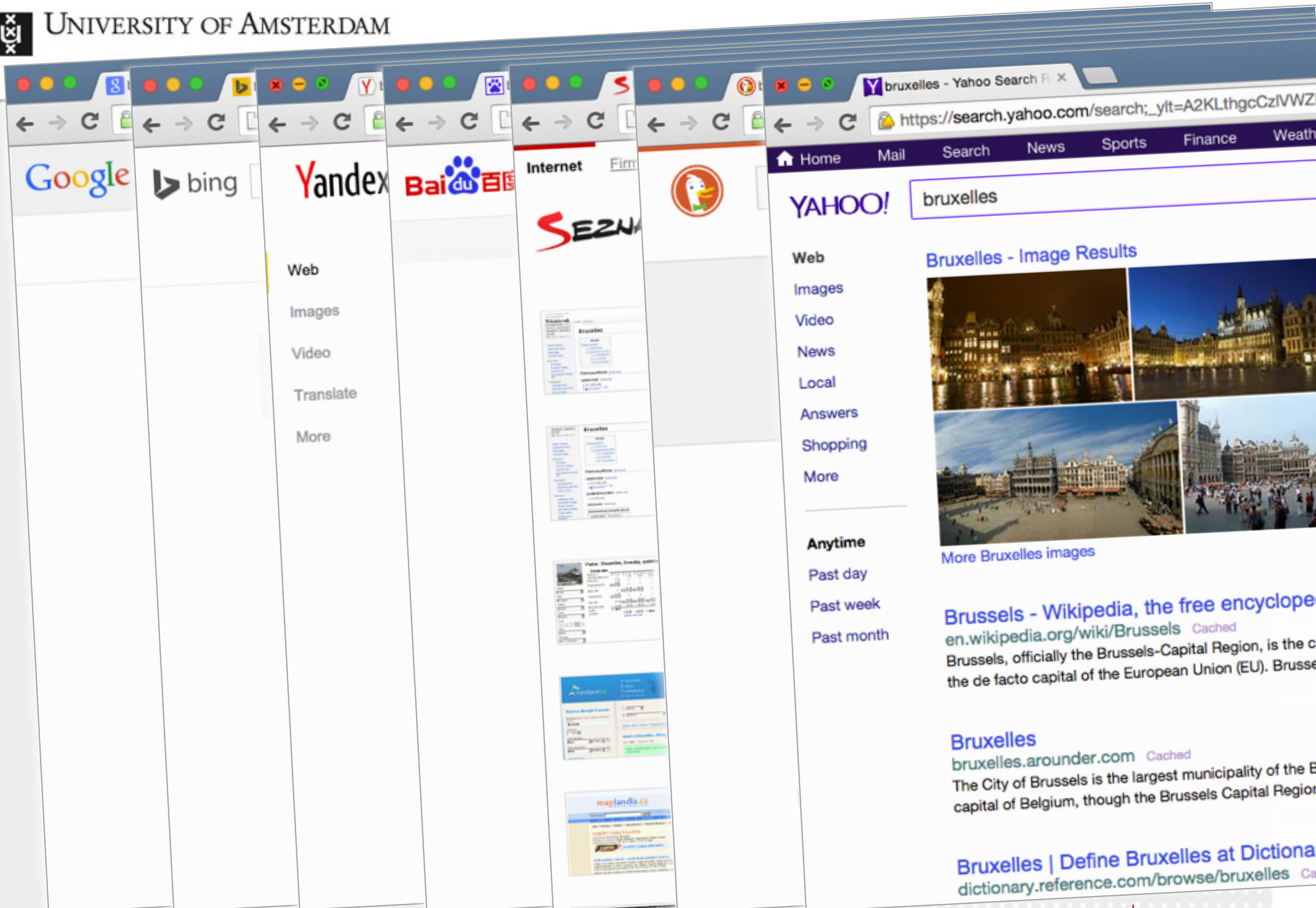
Katja Hofmann
Microsoft
katja.hofmann@microsoft.com

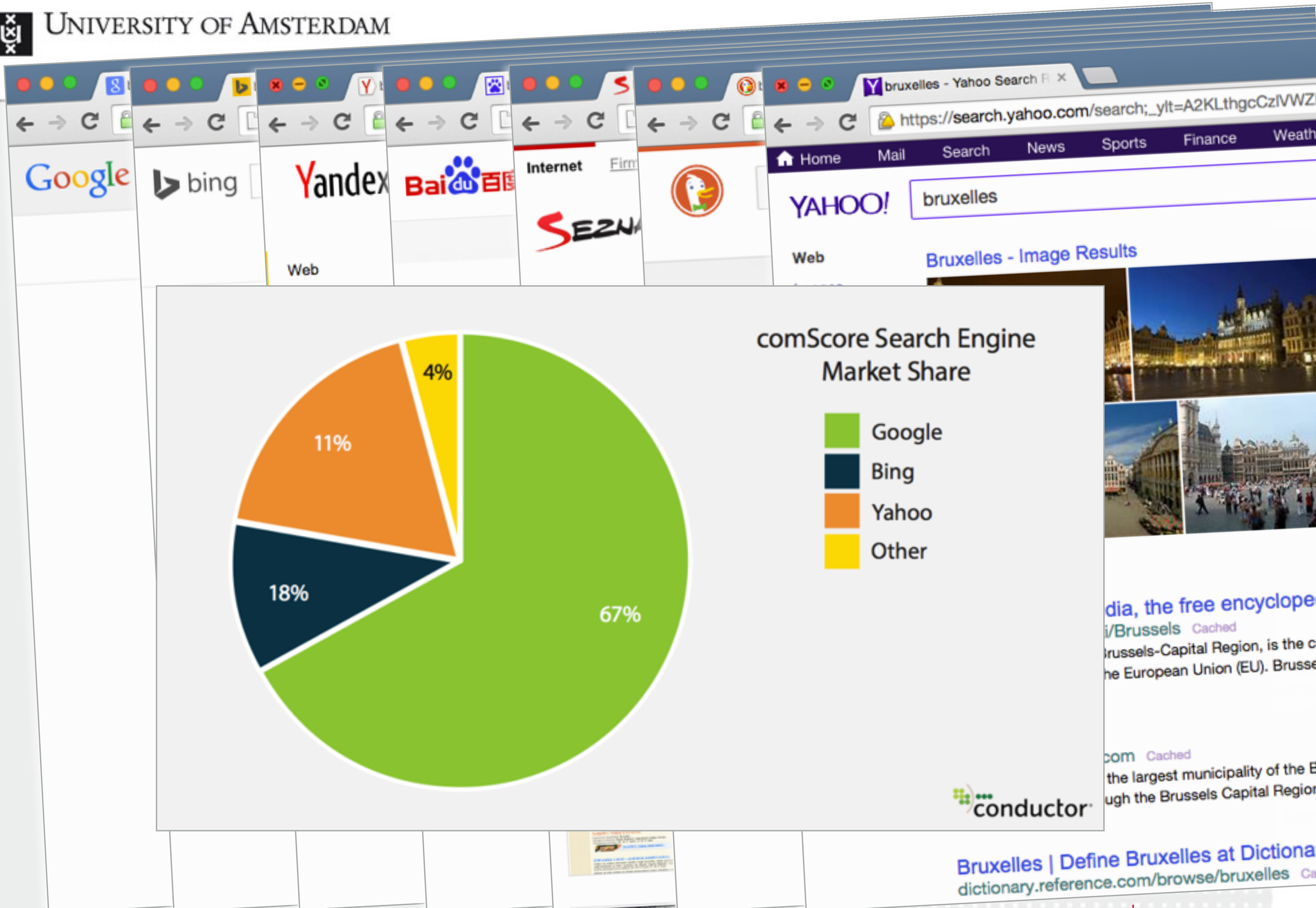
Filip Radlinski
Microsoft
filiprad@microsoft.com

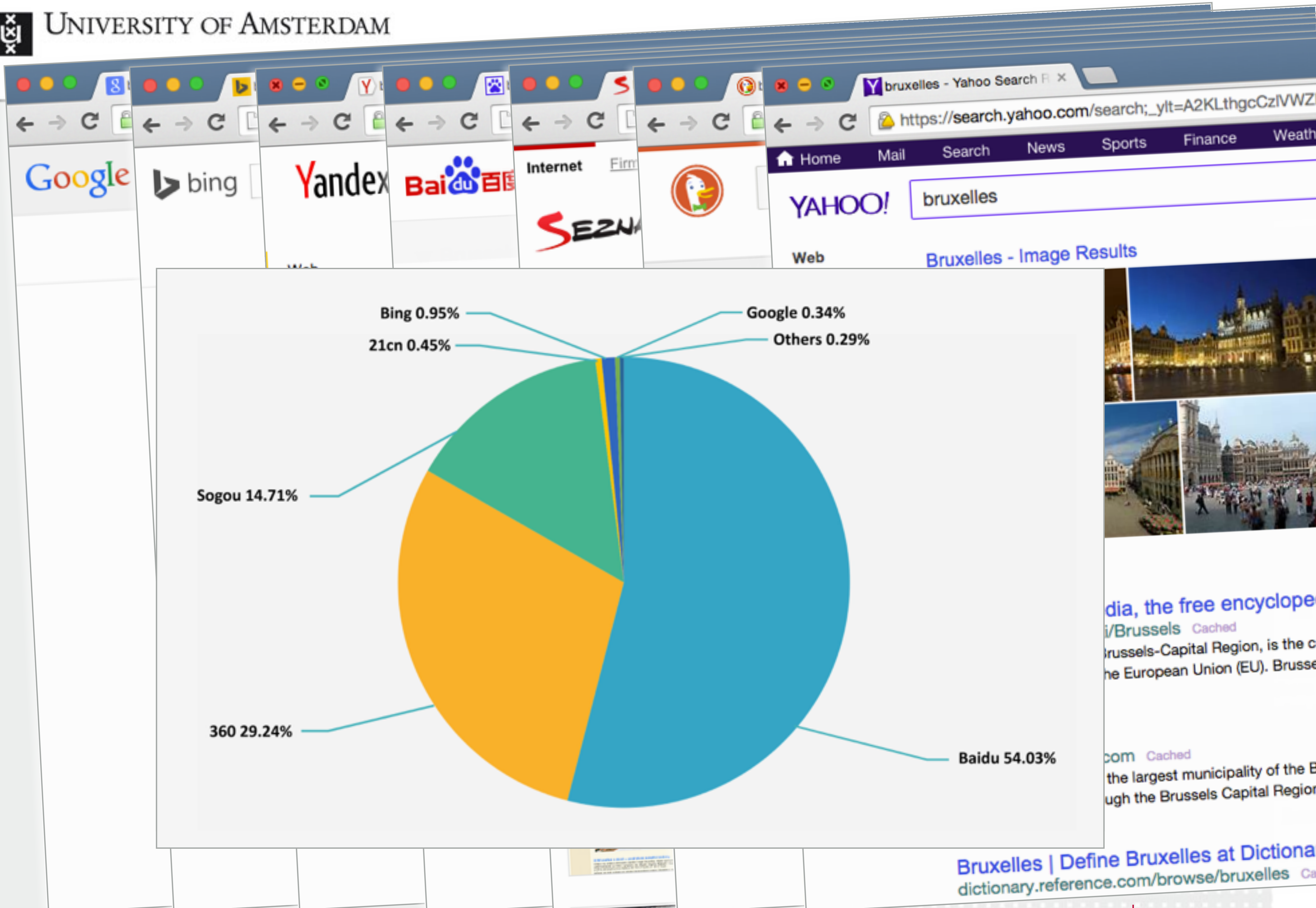
Beer&Tech, Criteo, October 28, 2015

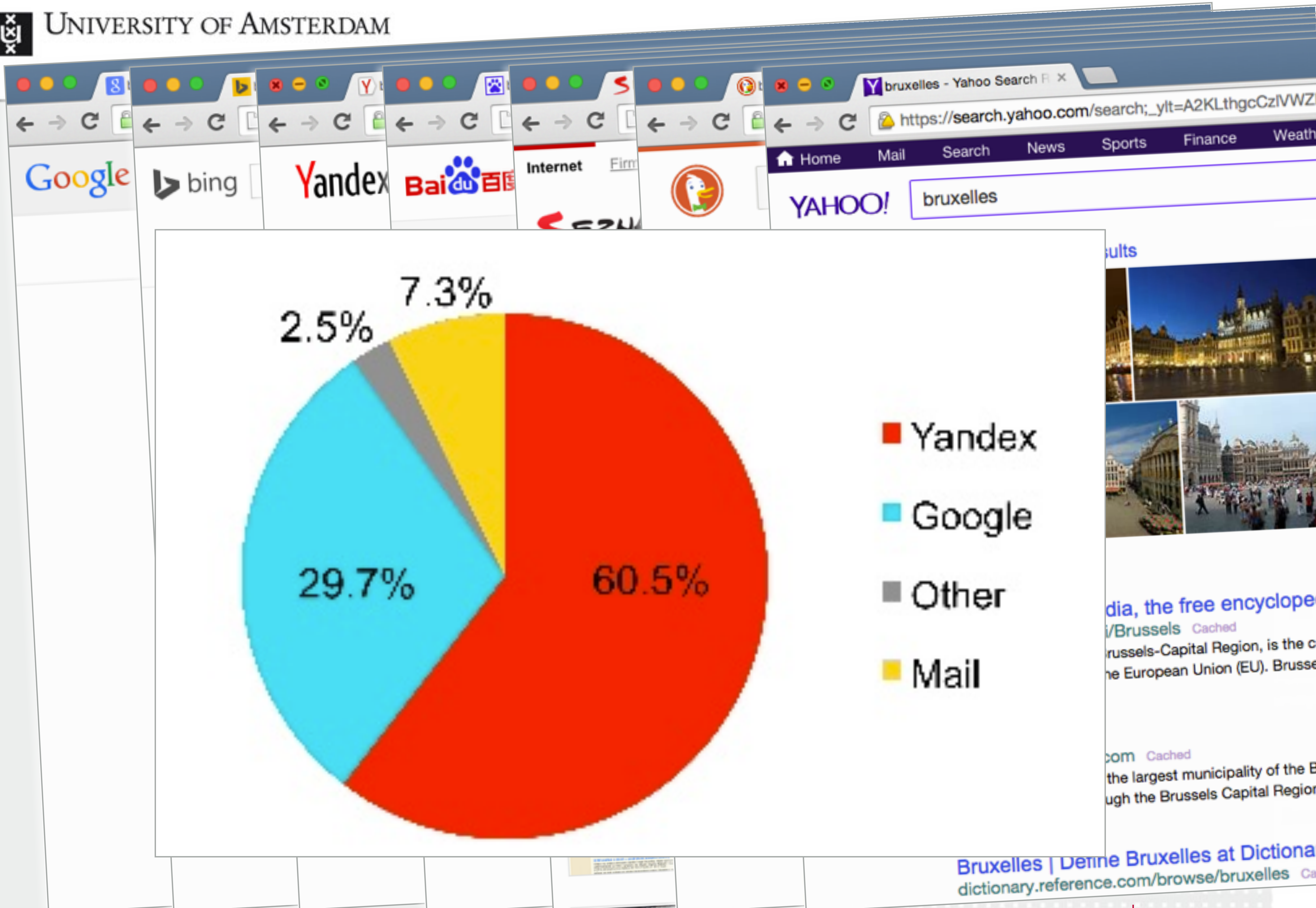


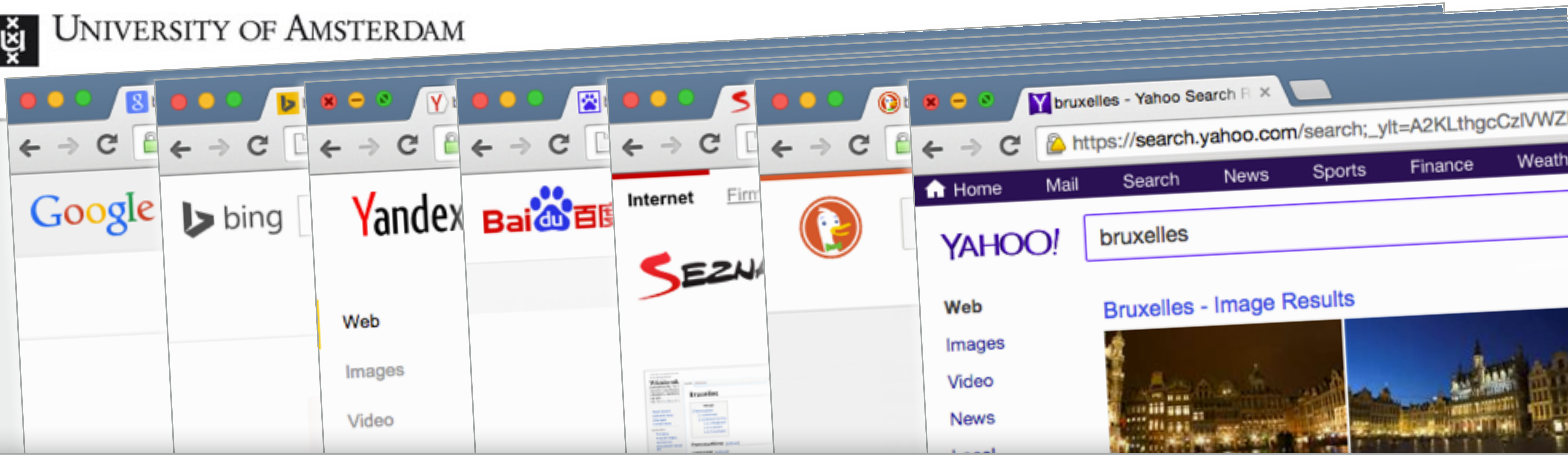
Search is not just Google


Google
bing
Yandex
Web
Images
Video
Translate
More
Internet
bruxelles - Yahoo Search
https://search.yahoo.com/search;_ylt=A2KLthgcCzlVWZ
Home
Mail
Search
News
Sports
Finance
YAHOO!
bruxelles
Web
Images
Video
News
Local
Answers
Shopping
More
Anytime
Past day
Past week
Past month
Bruxelles - Image Results
More Bruxelles images
Brussels - Wikipedia, the free encyclope
en.wikipedia.org/wiki/Brussels Cached
Brussels, officially the Brussels-Capital Region, is the c
the de facto capital of the European Union (EU). Bruss
Bruxelles
bruxelles.arounder.com Cached
The City of Brussels is the largest municipality of the B
capital of Belgium, though the Brussels Capital Regio
Bruxelles | Define Bruxelles at Dictiona
dictionary.reference.com/browse/bruxelles

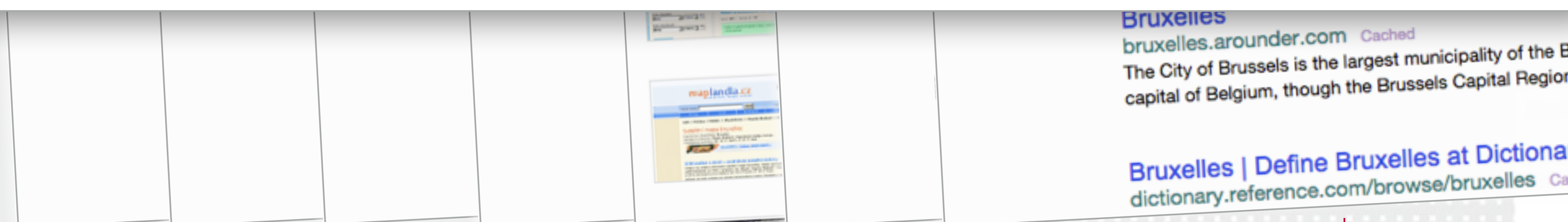

comScore Search Engine Market Share
Google
Bing
Yahoo
Other
67%
18%
11%
4%
conductor
bruxelles - Yahoo Search
https://search.yahoo.com/search;_ylt=A2KLthgcCzIVWZ
Home
Mail
Search
News
Sports
Finance
YAHOO!
bruxelles
Web
Bruxelles - Image Results
Bruxelles | Define Bruxelles at Dictiona
dictionary.reference.com/browse/bruxelles

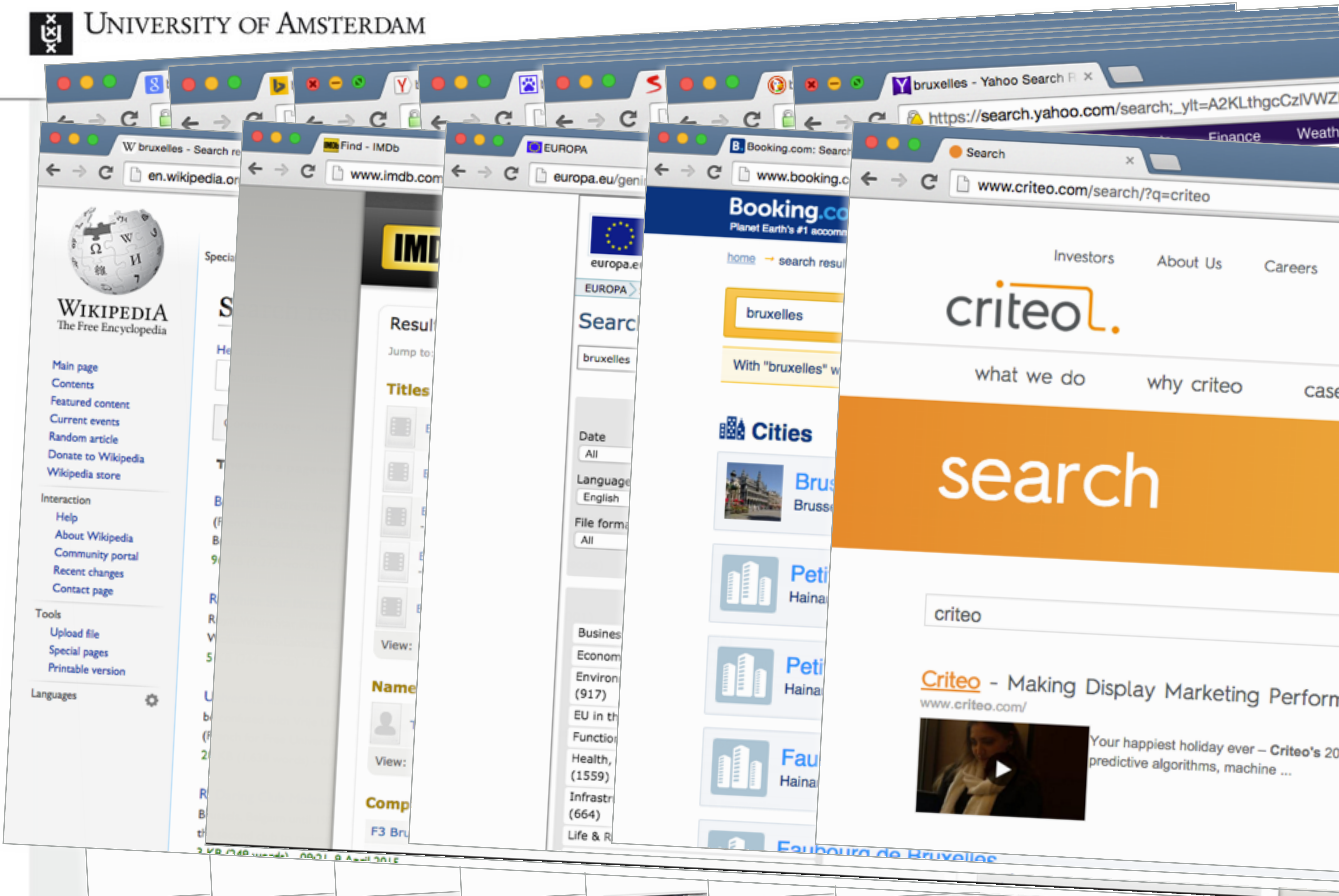

Google
bing
Yandex
Baidu
Internet
SEZNAM
bruxelles - Yahoo Search
https://search.yahoo.com/search;_ylt=A2KLthgcCzIVWZ
Home
Mail
Search
News
Sports
Finance
YAHOO!
bruxelles
Web
Bruxelles - Image Results
Bing 0.95%
21cn 0.45%
Google 0.34%
Others 0.29%
Sogou 14.71%
360 29.24%
Baidu 54.03%
Bruxelles | Define Bruxelles at Dictiona
dictionary.reference.com/browse/bruxelles

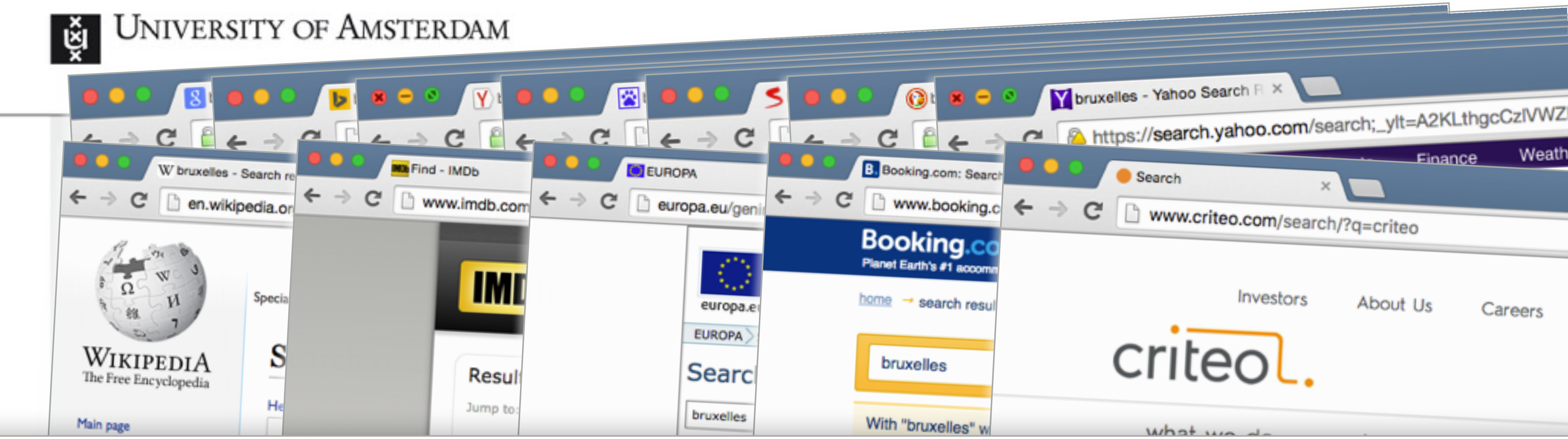

7.3%
2.5%
29.7%
60.5%
Yandex
Google
Other
Mail
bruxelles - Yahoo Search
https://search.yahoo.com/search;_ylt=A2KLthgcCzlVWZ
Home
Mail
Search
News
Sports
Finance
YAHOO!
bruxelles
Google
bing
Yandex
Internet
Bruxelles | Define Bruxelles at Dictiona
dictionary.reference.com/browse/bruxelles



Search is not just web search

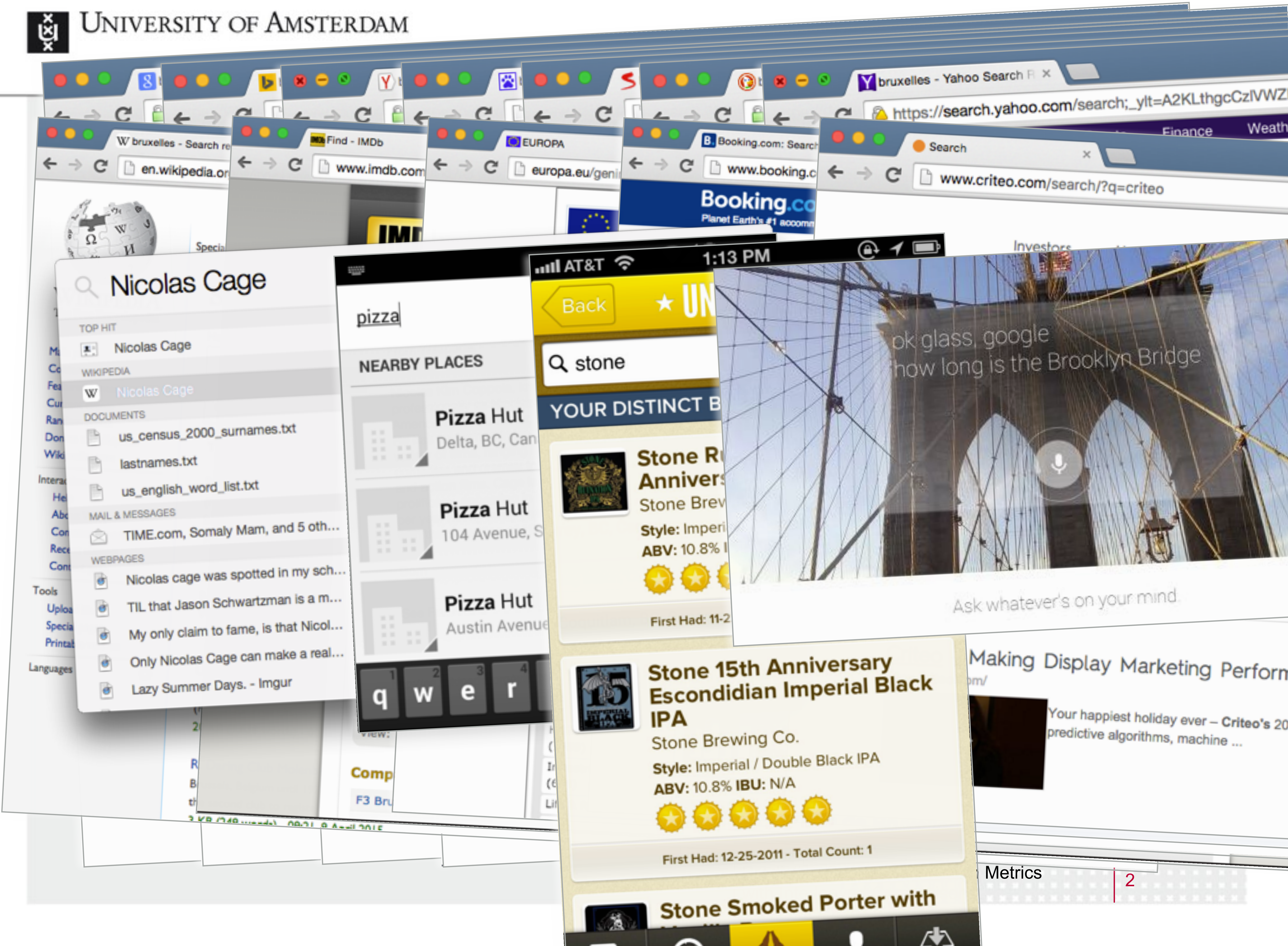


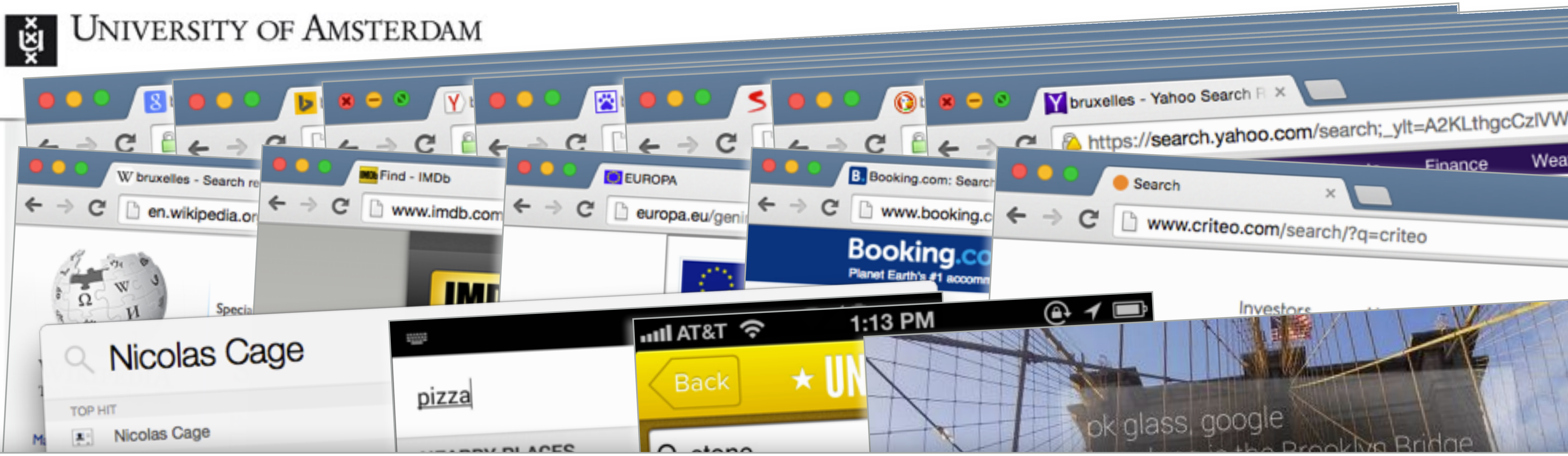

bruxelles - Yahoo Search
https://search.yahoo.com/search;_ylt=A2KLthgcCzlVWZ
Finance
W bruxelles - Search
en.wikipedia.or
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
Upload file
Special pages
Printable version
Languages
Find - IMDb
www.imdb.com
Titles
View:
EUROPA
EUROPA
Date
All
English
All
Booking.com: Search
www.booking.c
Planet Earth's #1 accomm
home
bruxelles
Cities
Search
www.criteo.com/search/?q=criteo
Investors
About Us
Careers
criteo
what we do
why criteo
search
criteo
Criteo - Making Display Marketing Perform
www.criteo.com/
Your happiest holiday ever – Criteo's 20
predictive algorithms, machine ...



Search is not just in a browser

University of Amsterdam





Search is everywhere

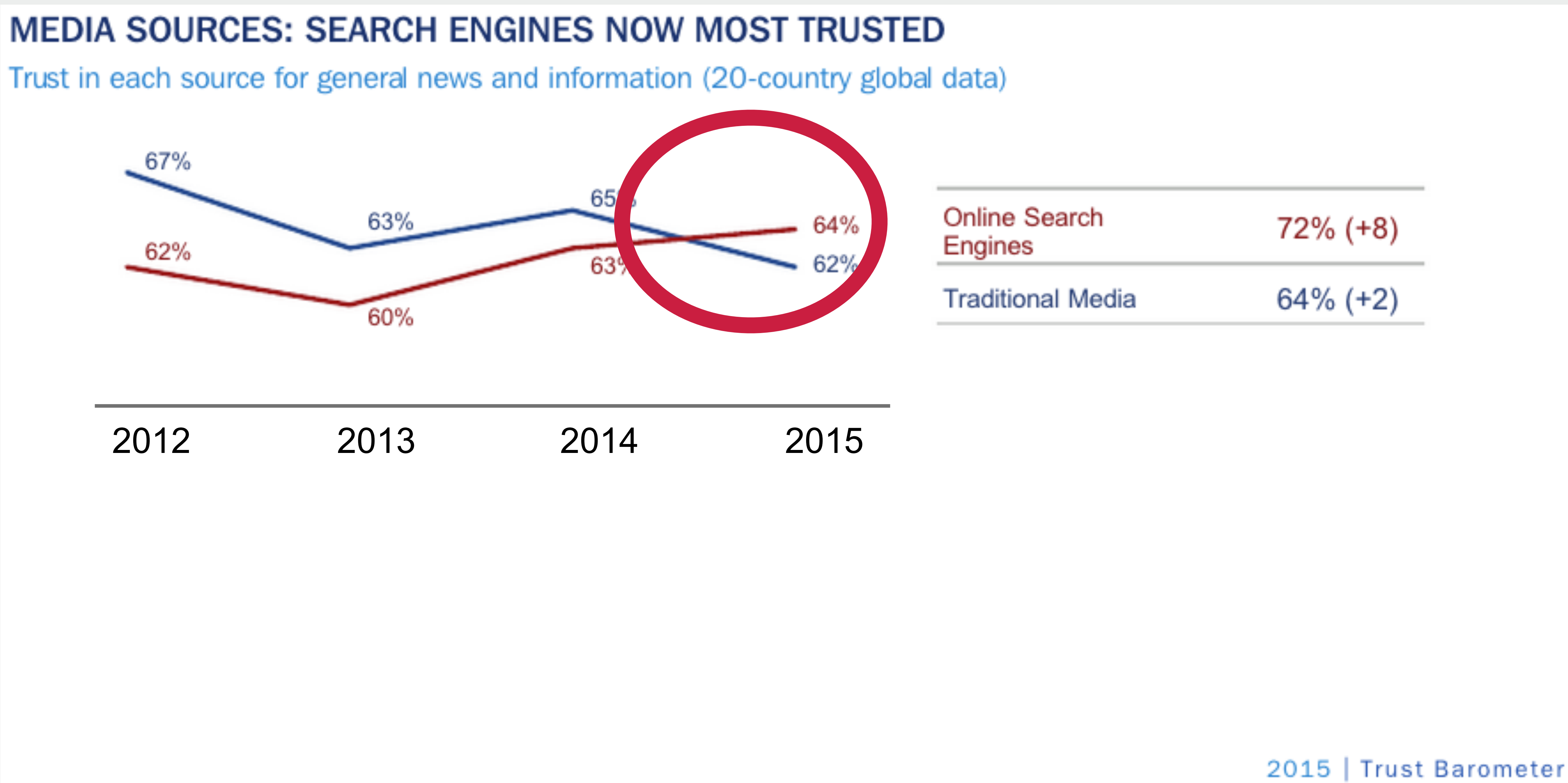
Motivation - **Search**

- **Half the world**'s population uses web search

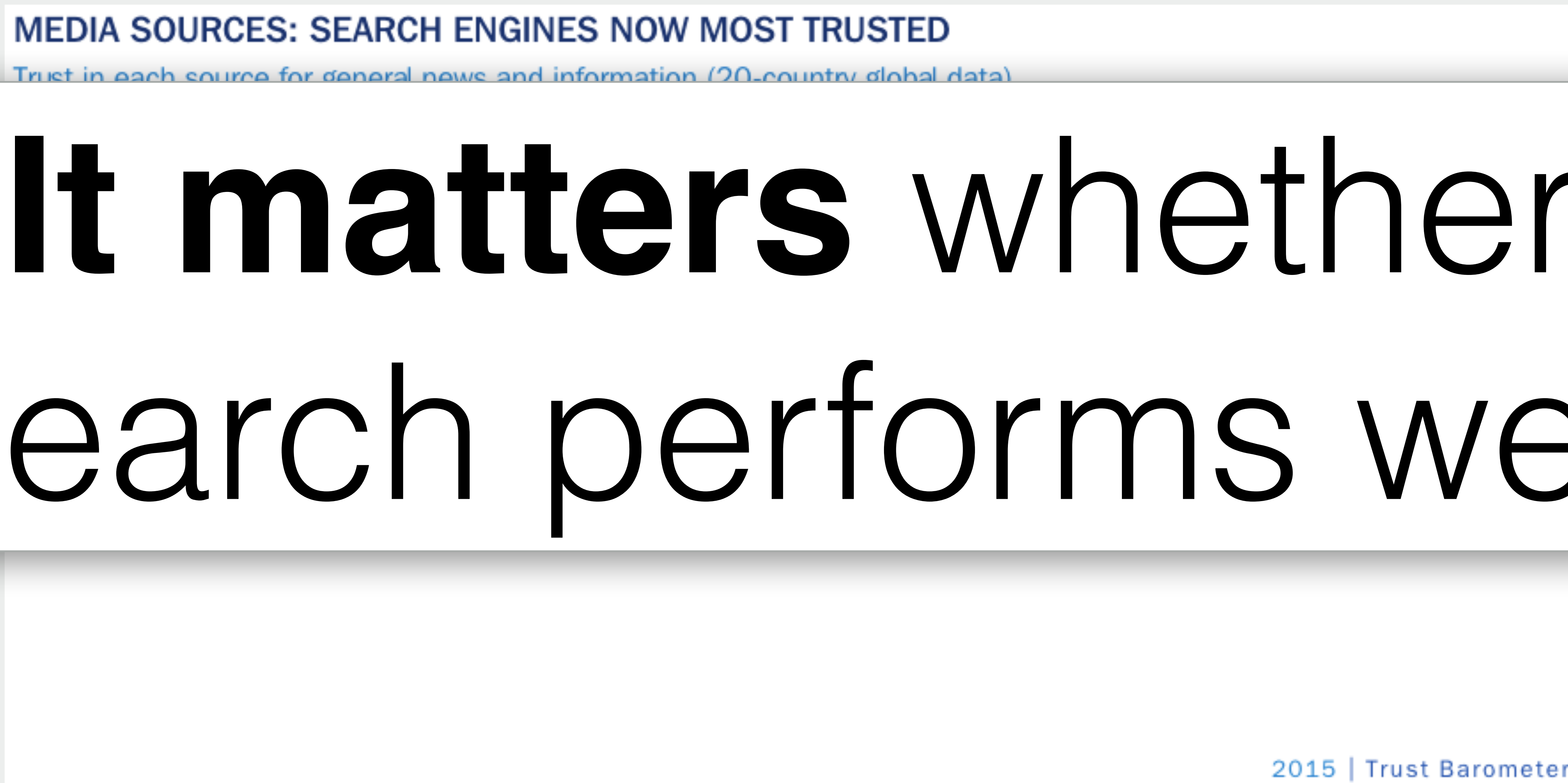
Motivation - **Search**

- **Half the world**'s population uses web search
- Web **search is trusted** more than traditional media

Motivation - **Search**

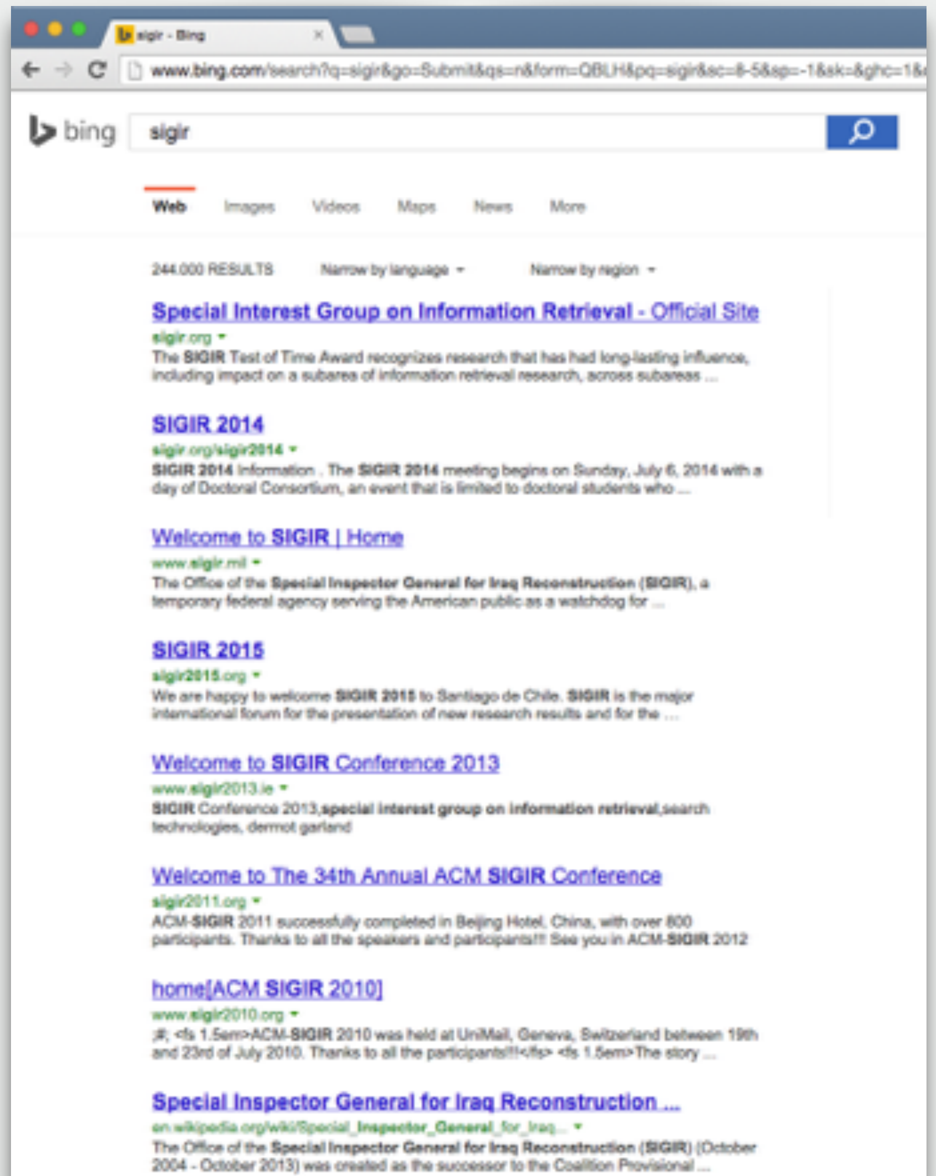


Motivation - **Search**

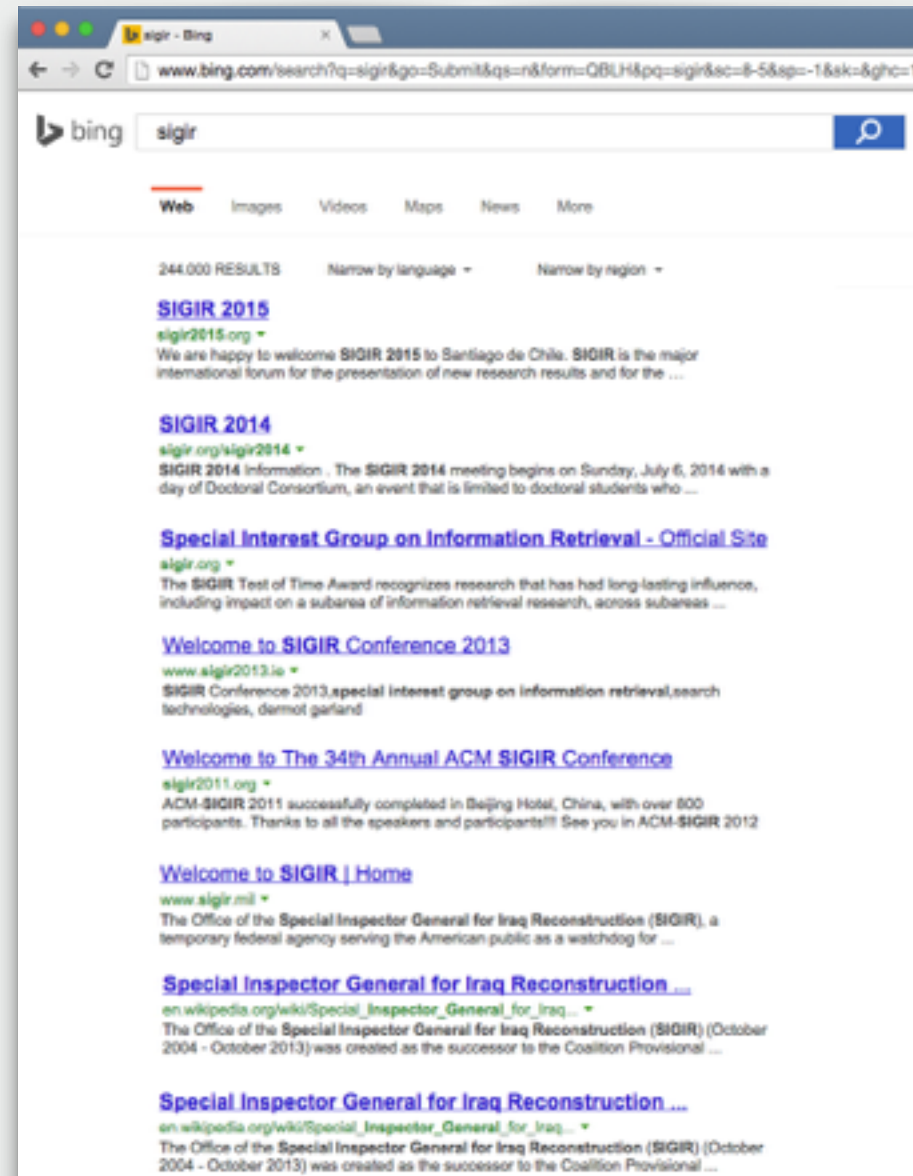


It matters whether search performs well

Motivation - **Evaluation**



or



University of Amsterdam

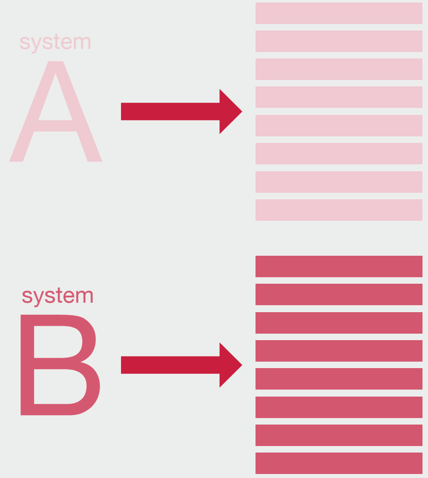
Motivation - **Evaluation**

system
A

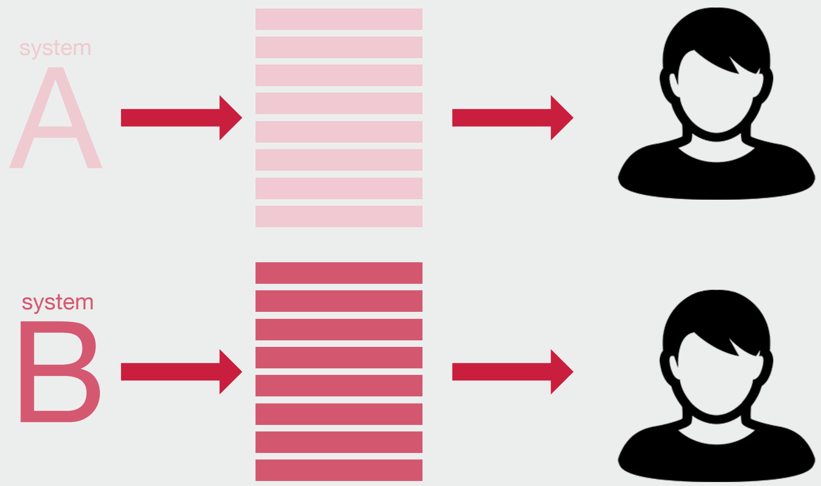
or

system
B

Motivation - **AB Testing**

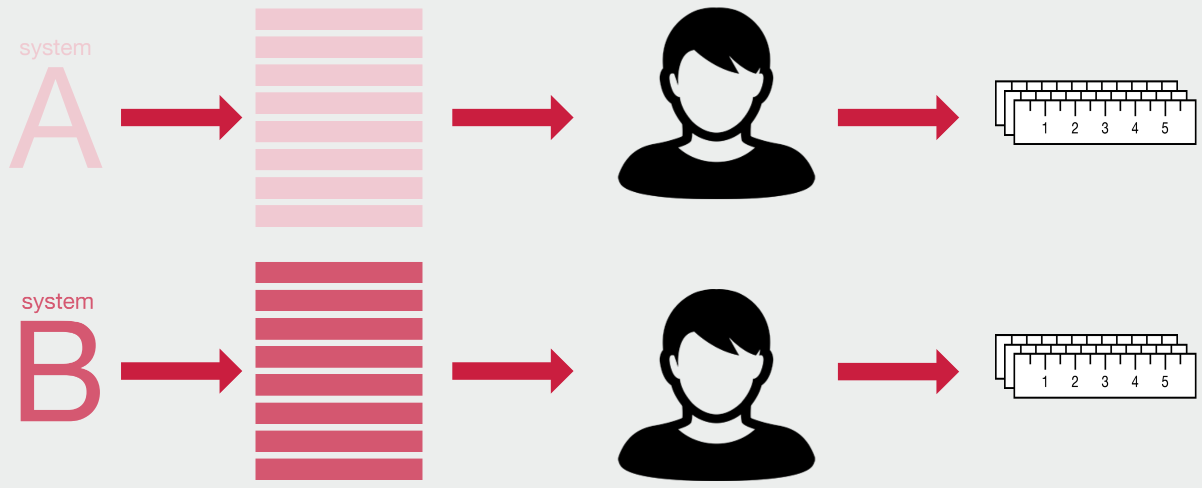


Motivation - **AB Testing**



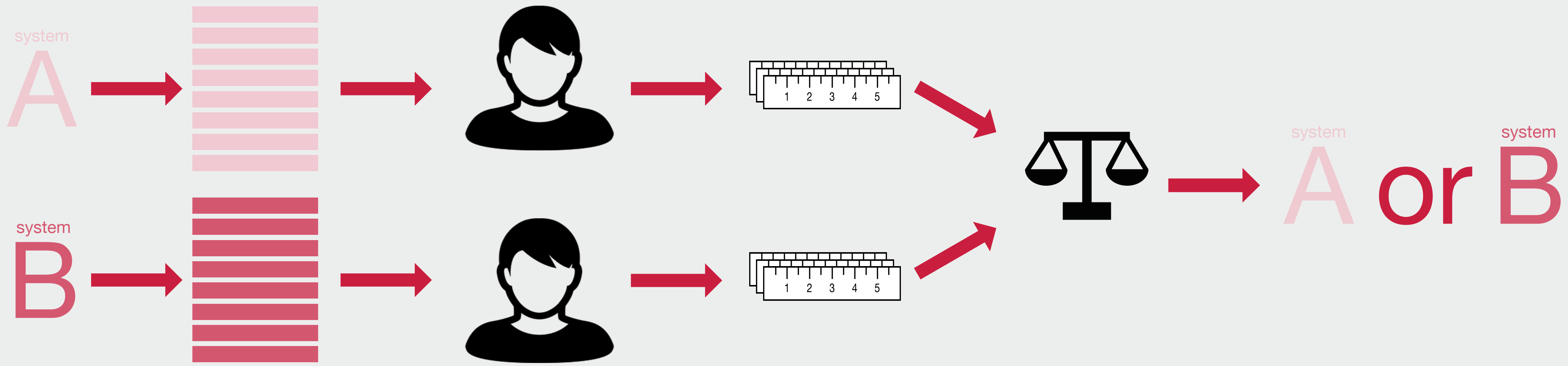
- User population **divided** into two groups

Motivation - **AB Testing**



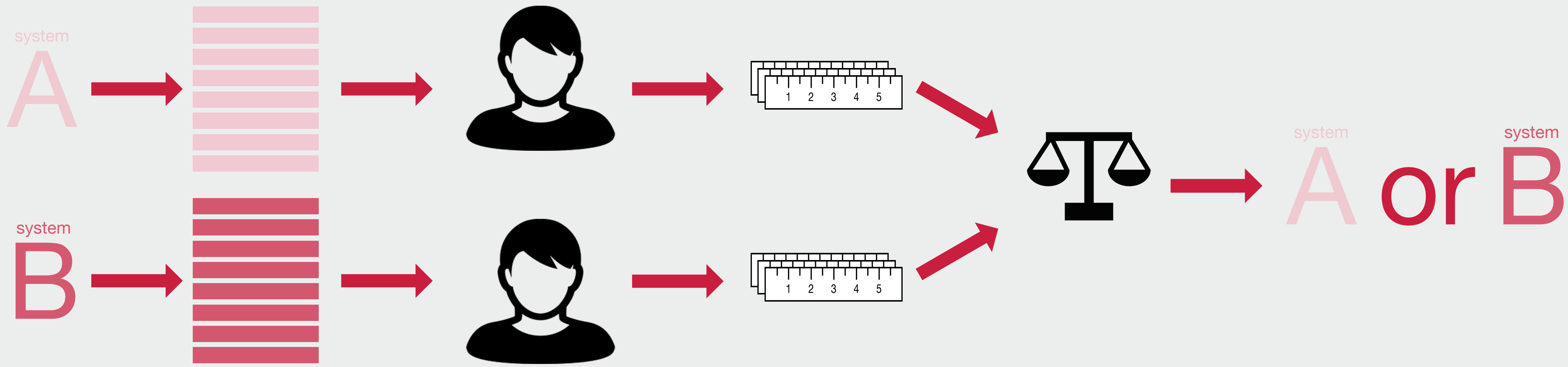
- User population **divided** into two groups
- Trusted and **sophisticated metrics**

Motivation - **AB Testing**



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

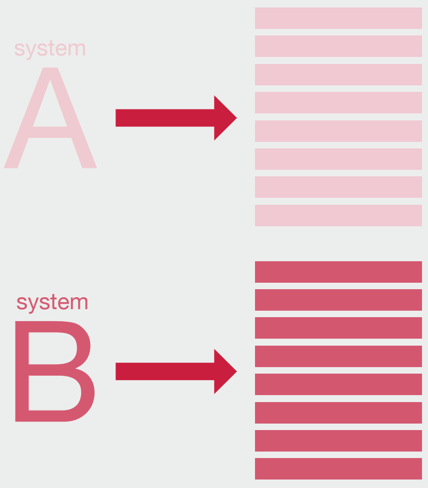
Motivation - **AB Testing**



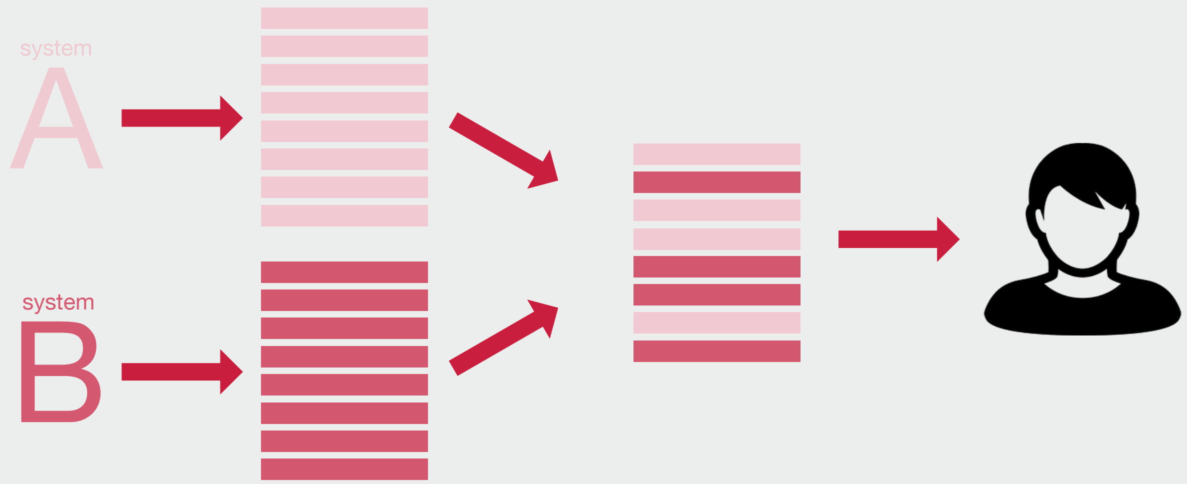
- User population **divided** into two groups
- Trusted and **sophisticated metrics**
- **Difference in metric value** indicates the winner
- **Between subject** design
  - Differences between users and their queries
  - **Low sensitivity**, millions of queries

University of Amsterdam

Motivation - **Interleaving**



Motivation - **Interleaving**



- All users see **both** systems

Motivation - **Interleaving**

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

Motivation - **Interleaving**

- All 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)

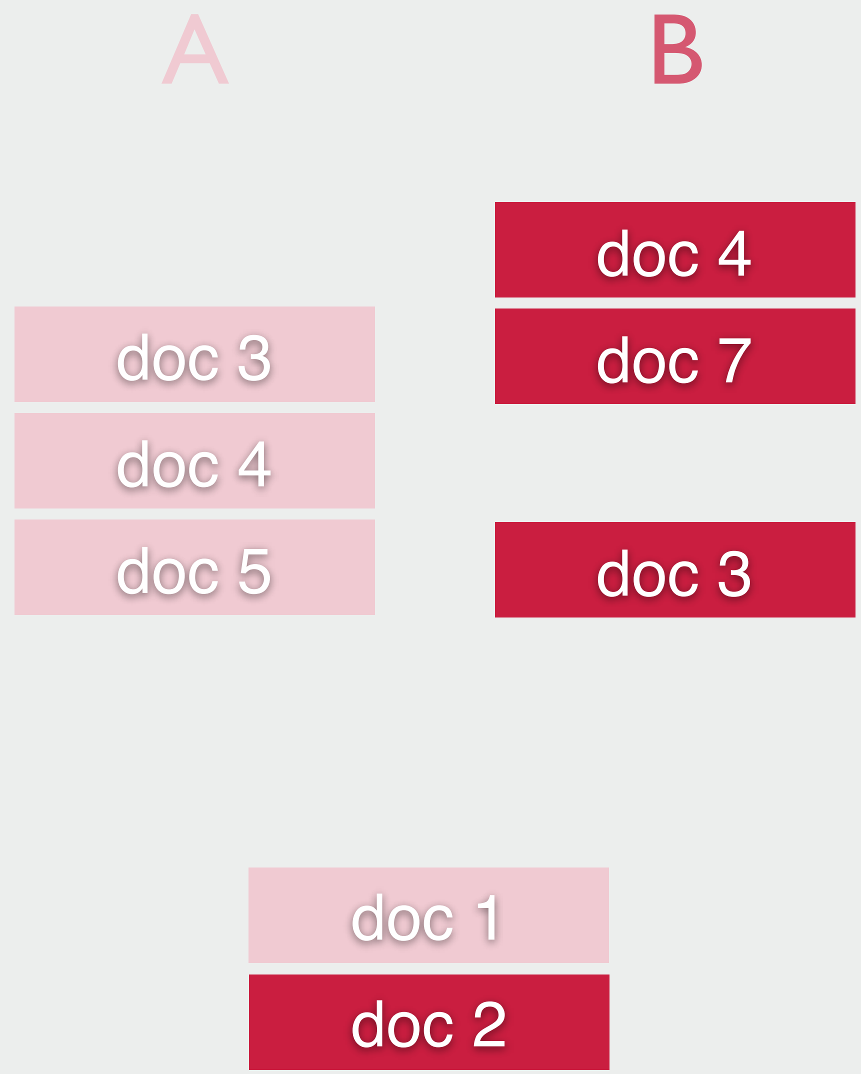
O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. In ACM Transactions on Information Systems. 2012

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

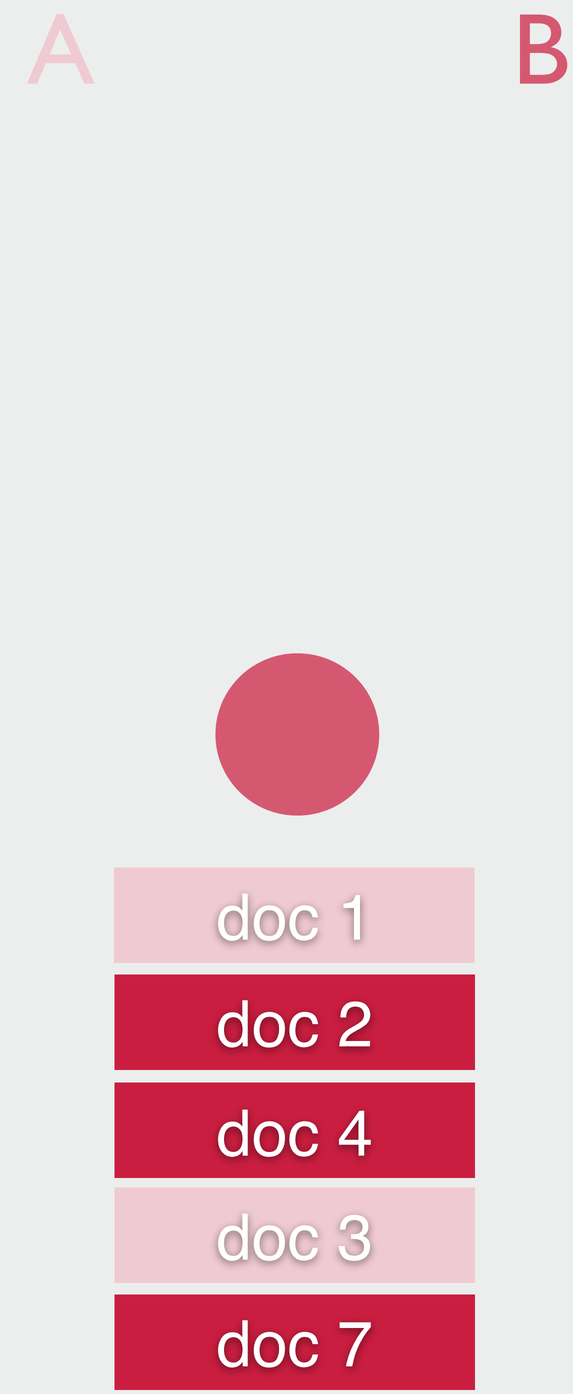
F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM'08. 2008

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



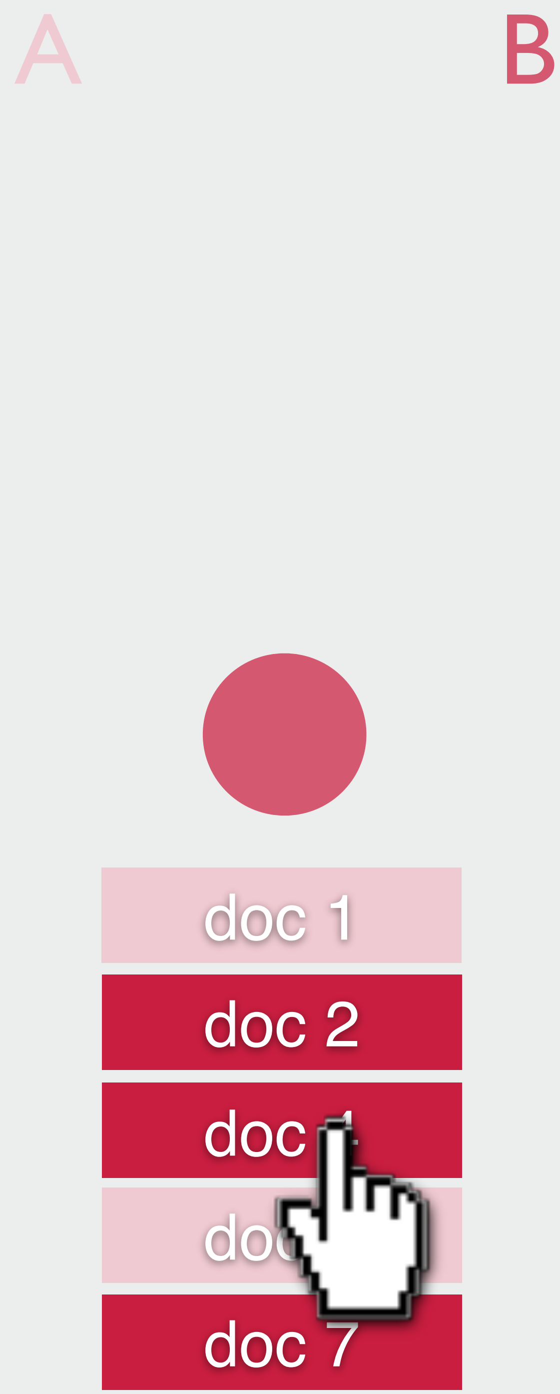
F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM'08. 2008

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



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

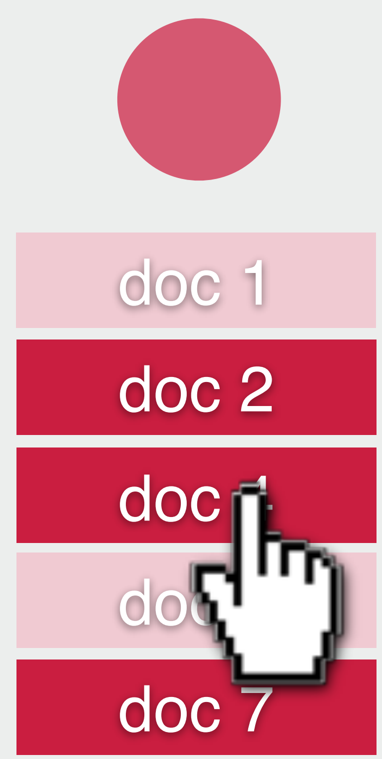
Motivation - **Team Draft Interleaving (TDI)**



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

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

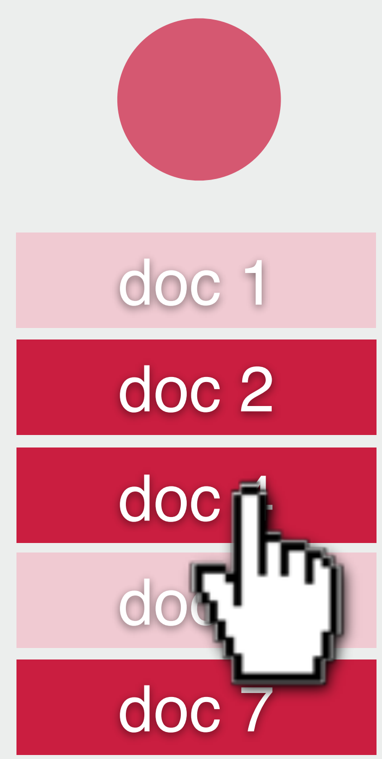
- Infer winner: **B > A**



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

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

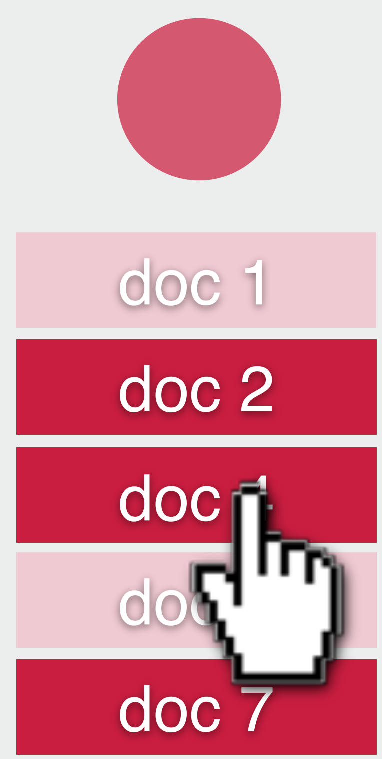
- Infer winner: **B > A**
- Count **fraction of wins** over many queries



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: **B > A**
- Count **fraction of wins** over many queries
- Well tested in practice
  - Used at Bing, Yandex, Seznam



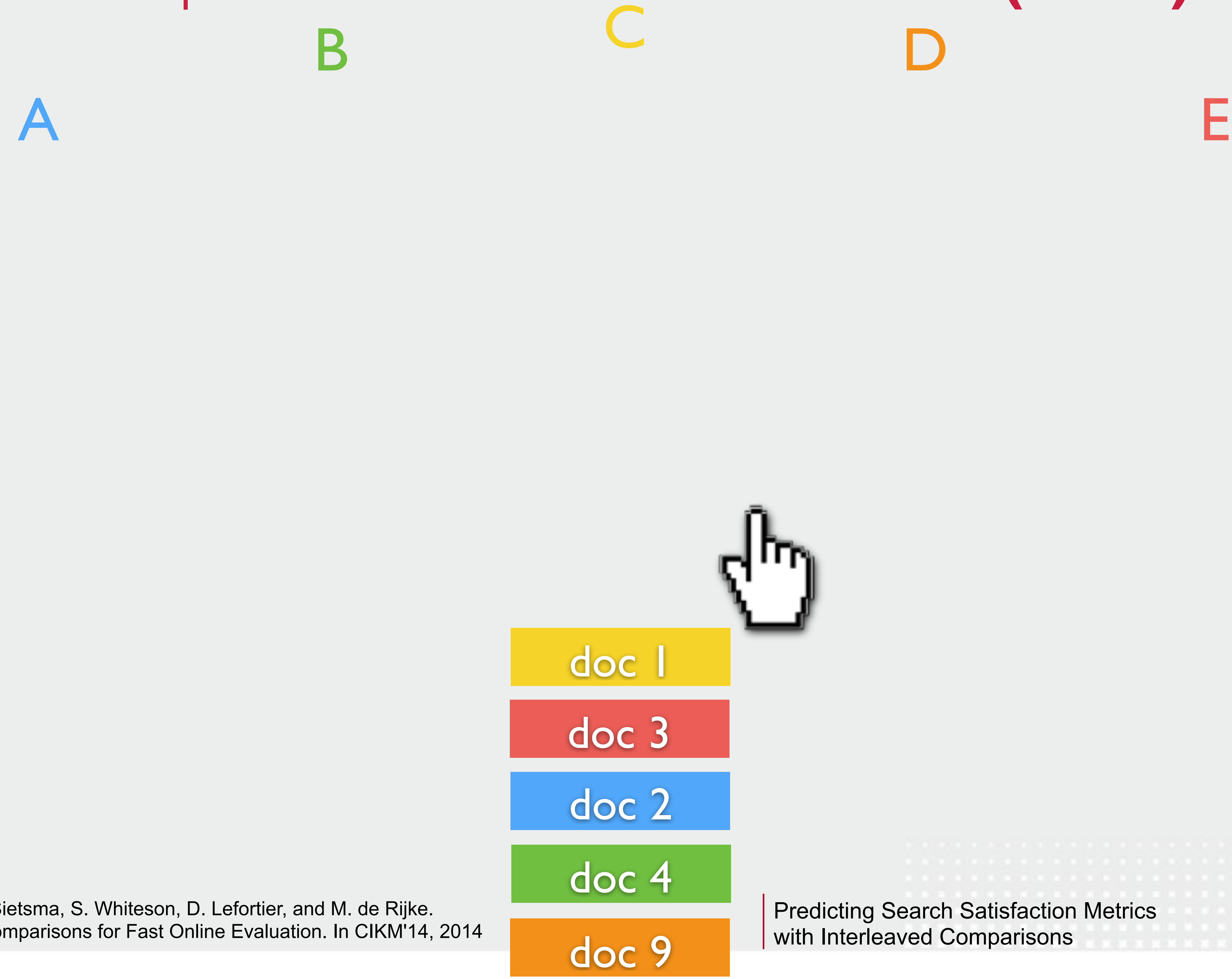
F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM'08. 2008

Side step - **Team Draft Multileave (TDM)**

A	B	C	D	E
doc 1	doc 2	doc 1	doc 4	doc 3
doc 2	doc 4	doc 2	doc 2	doc 1
doc 3	doc 7	doc 8	doc 1	doc 2
doc 4	doc 1	doc 3	doc 9	doc 5
doc 5	doc 3	doc 9	doc 5	doc 7

A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke.
Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

Side step - **Team Draft Multileave (TDM)**



A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke.
Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

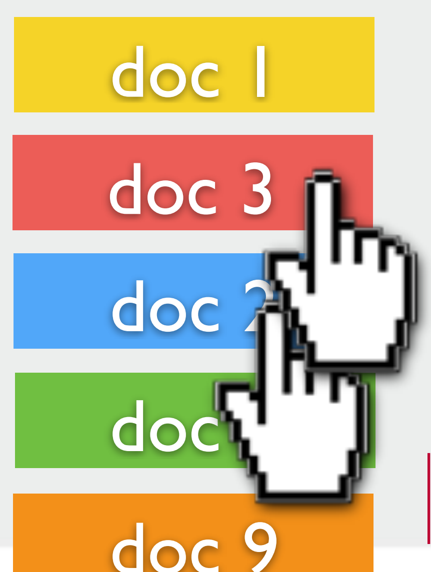
Side step - **Team Draft Multileave (TDM)**

C

B D

A E

- Infer ranking over systems: **A & E > B & C & D**



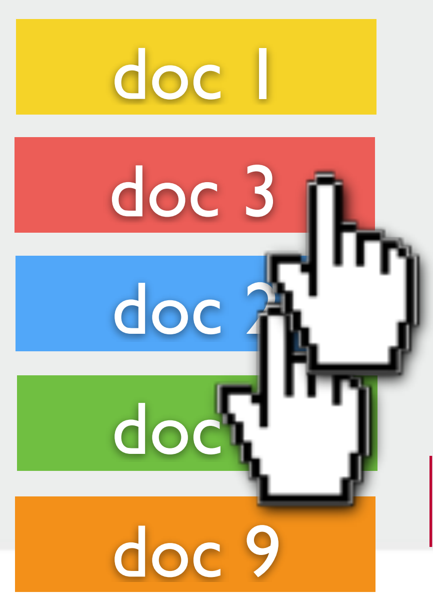
A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

Side step - **Team Draft Multileave (TDM)**

B C D

A E

- Infer ranking over systems: **A & E > B & C & D**
- Aggregate **rankings** over many queries



A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

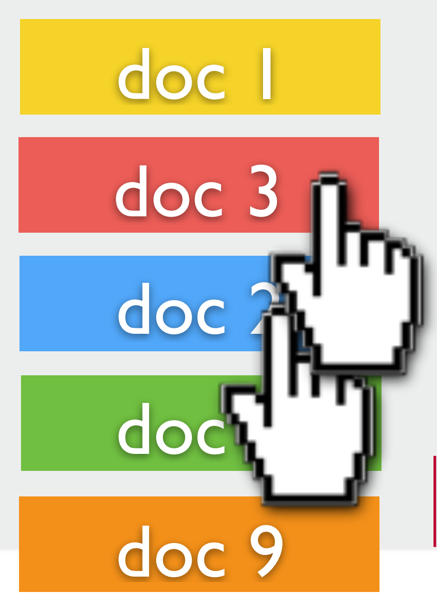
Side step - **Team Draft Multileave (TDM)**

C

B D

A E

- Infer ranking over systems: **A & E > B & C & D**
- Aggregate **rankings** over many queries
- Many less queries required



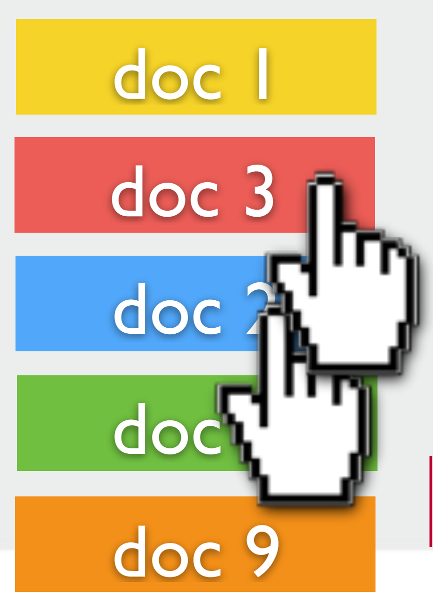
A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

Side step - **Team Draft Multileave (TDM)**

B C D

A E

- Infer ranking over systems: **A & E > B & C & D**
- Aggregate **rankings** over many queries
- Many less queries required
  - Relative to when all systems would be compared pairwise



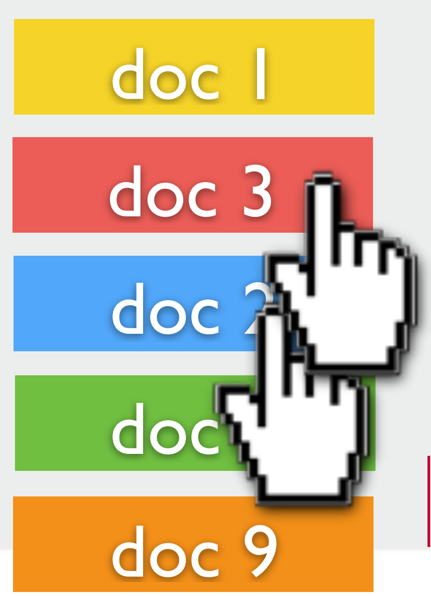
A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

Side step - **Team Draft Multileave (TDM)**

B C D

A E

- Infer ranking over systems: **A & E > B & C & D**
- Aggregate **rankings** over many queries
- Many less queries required
  - Relative to when all systems would be compared pairwise
- But not tested in practice (yet)



A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

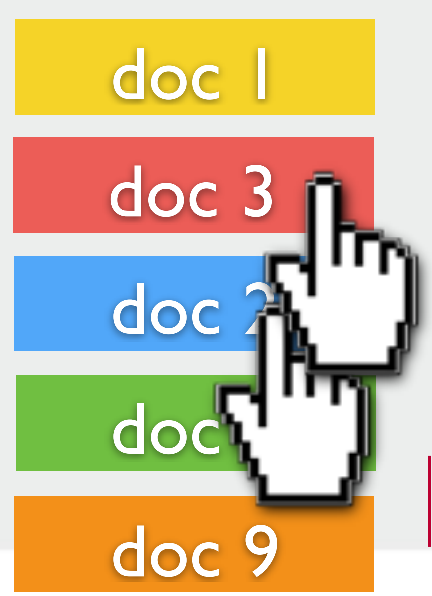
Not used in the rest of this work

Side step - **Team Draft Multileave (TDM)**

B C D

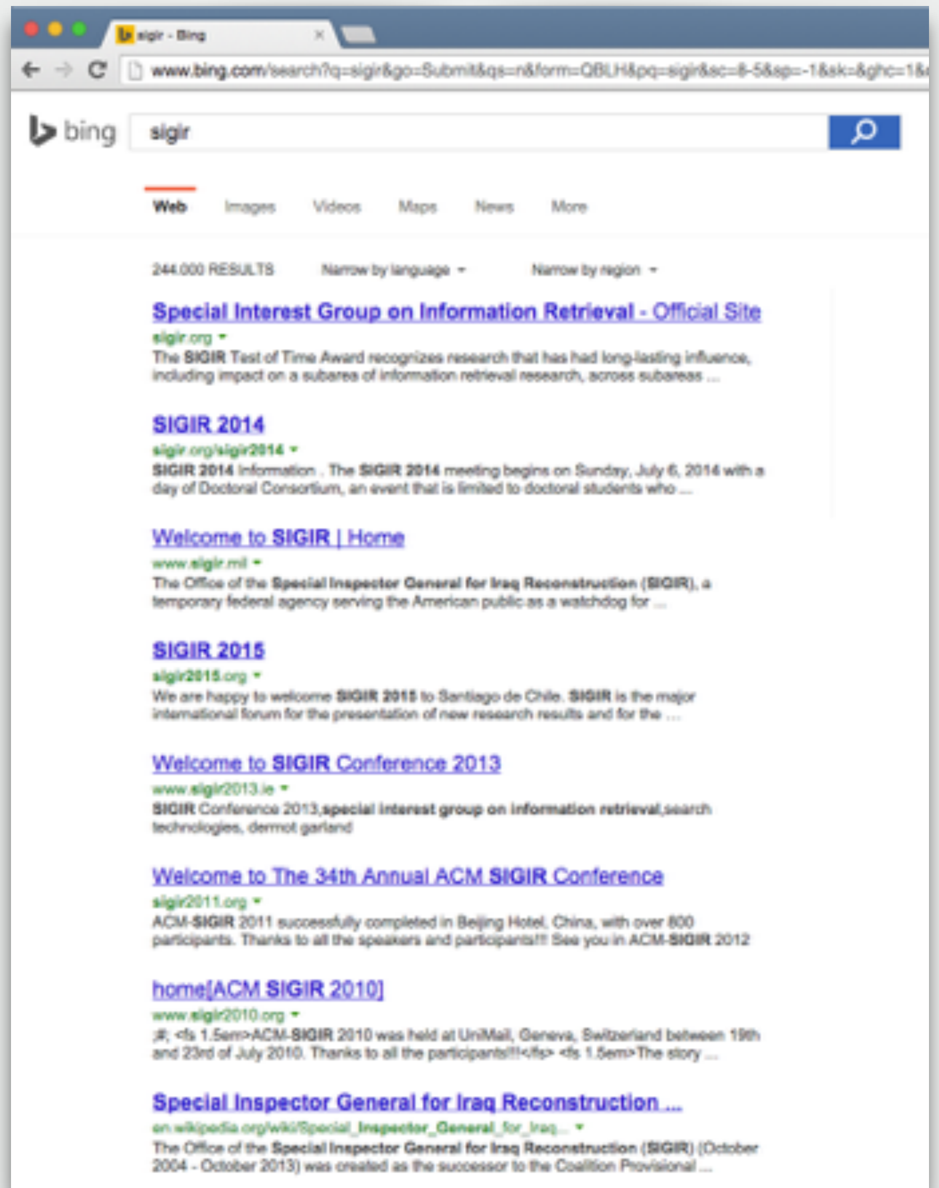
A E

- Infer ranking over systems: A & E > B & C & D
- Aggregate **rankings** over many queries
- Many less queries required
  - Relative to when all systems would be compared pairwise
- But not tested in practice (yet)

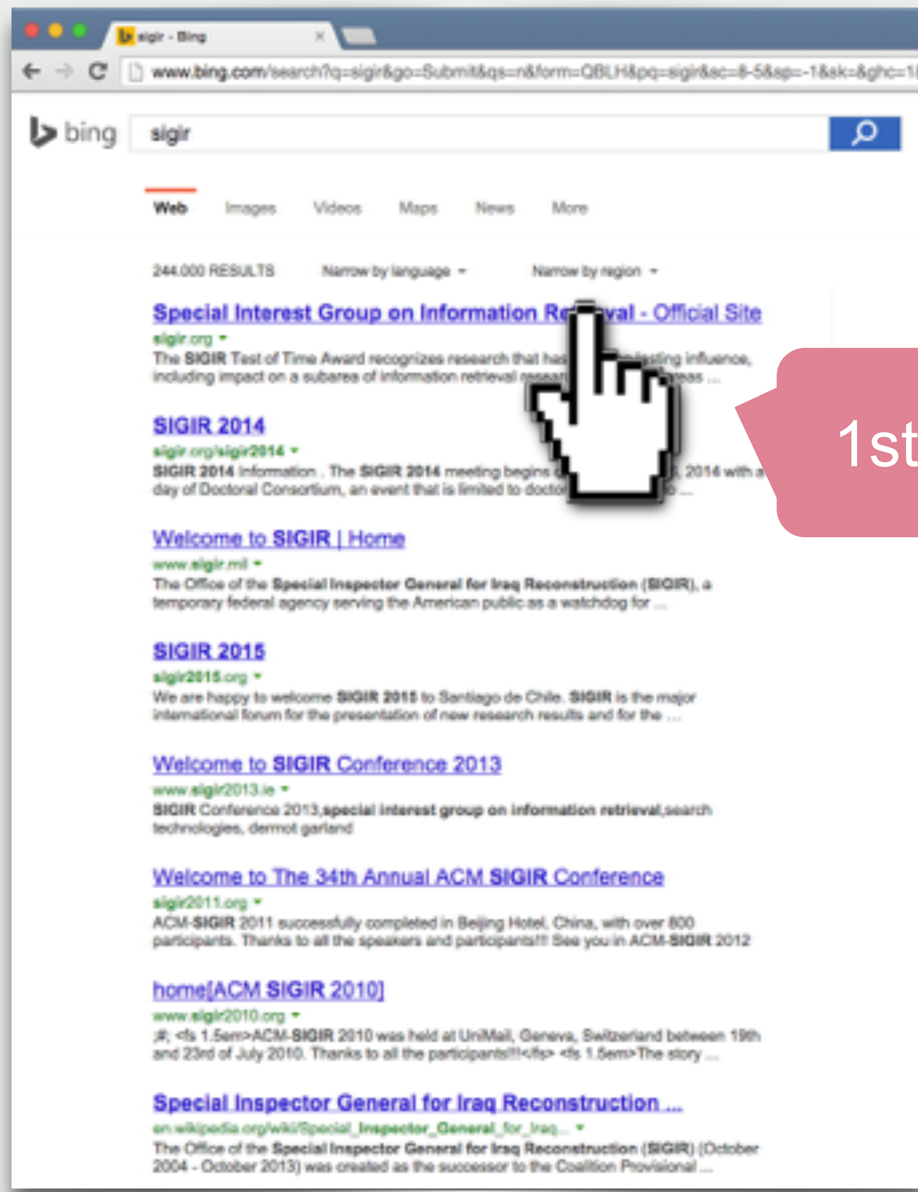


A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved Comparisons for Fast Online Evaluation. In CIKM'14, 2014

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

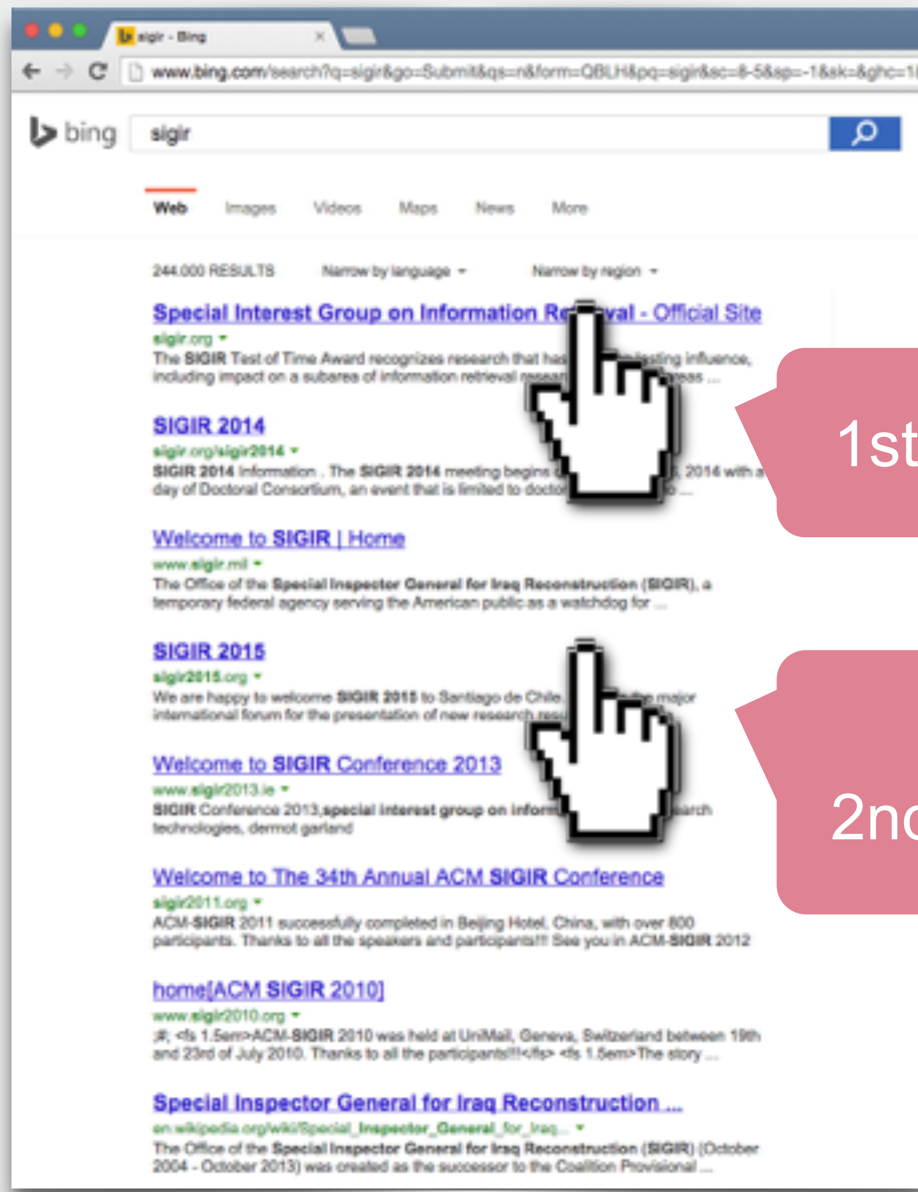


Motivation - AB Testing - **As a Gold Standard**



1st click, 5sec dwell time

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



1st click, 5sec dwell time

“SAT” click:
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**

UNIVERSITY OF AMSTERDAM

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**

Y. Kim, A. Hassan, R. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In WSDM'14. 2014.

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[n]**
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**

Y. Kim, A. Hassan, R. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In WSDM'14. 2014.

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**

Y. Kim, A. Hassan, R. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In WSDM'14. 2014.

Motivation - AB Testing - **Metrics**

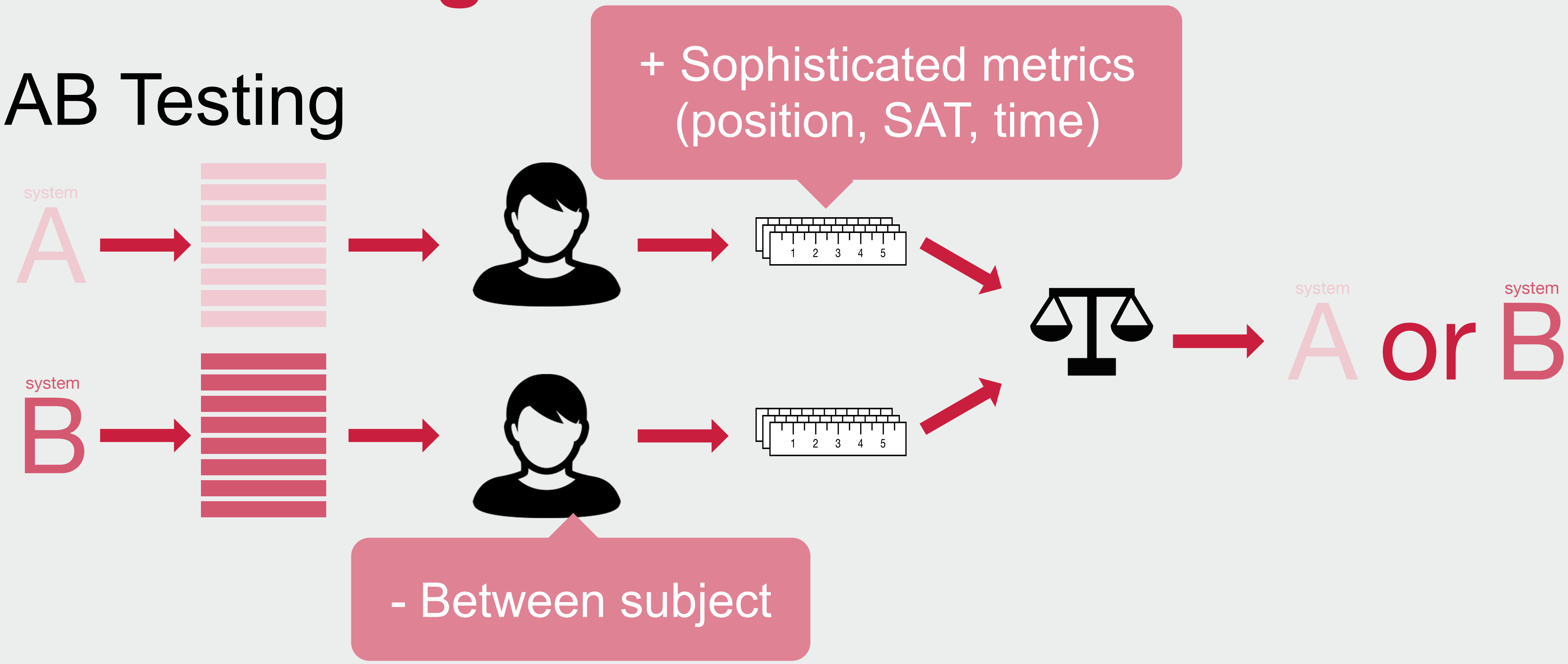
AB Metric	Description
AB	Fraction queries with at least one **click**
AB@1	Fraction queries with at least one **click on 1st positio**
AB_S	Fraction queries with at least one **SAT click**
AB_S@1	Fraction queries with at least one **SAT click on 1st position**
AB_T	**Time** from the query issue **until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

Y. Kim, A. Hassan, R. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In WSDM'14. 2014.

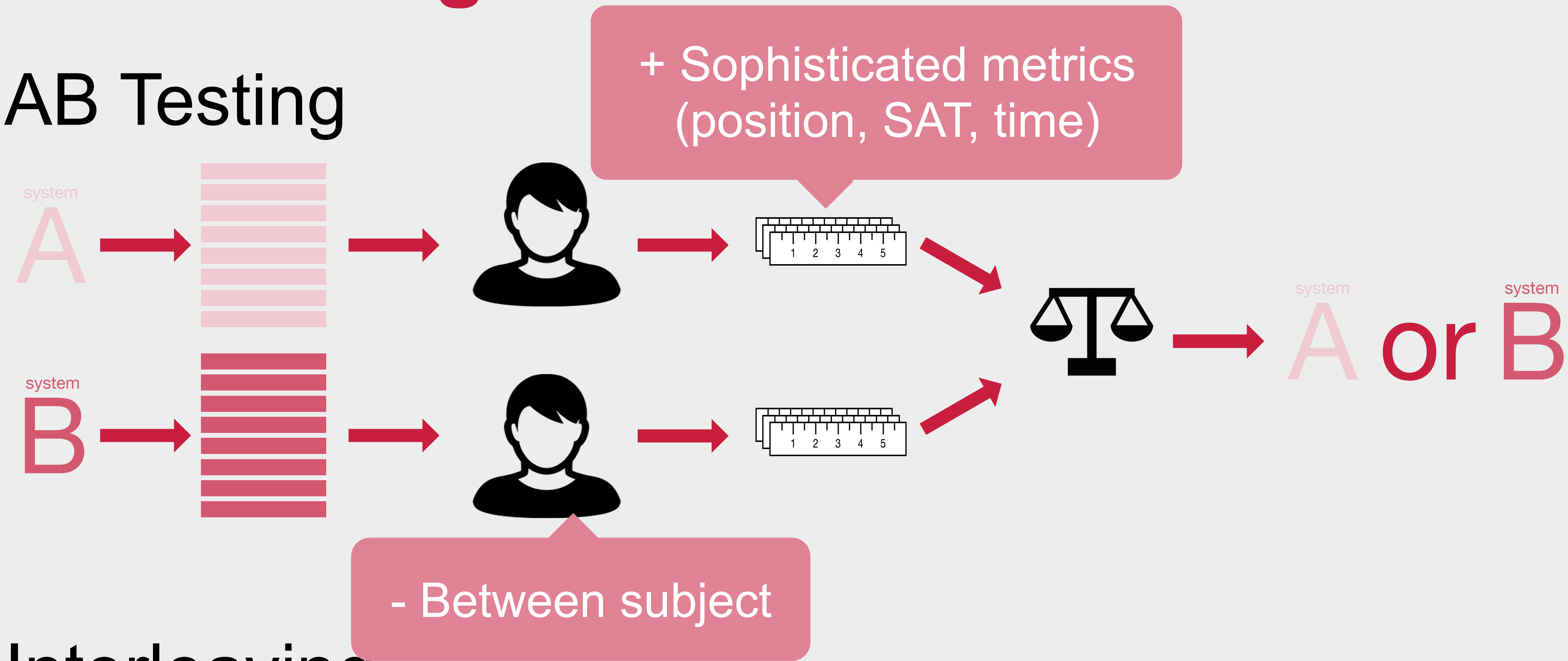
Motivation - **Agreement**

- AB Testing

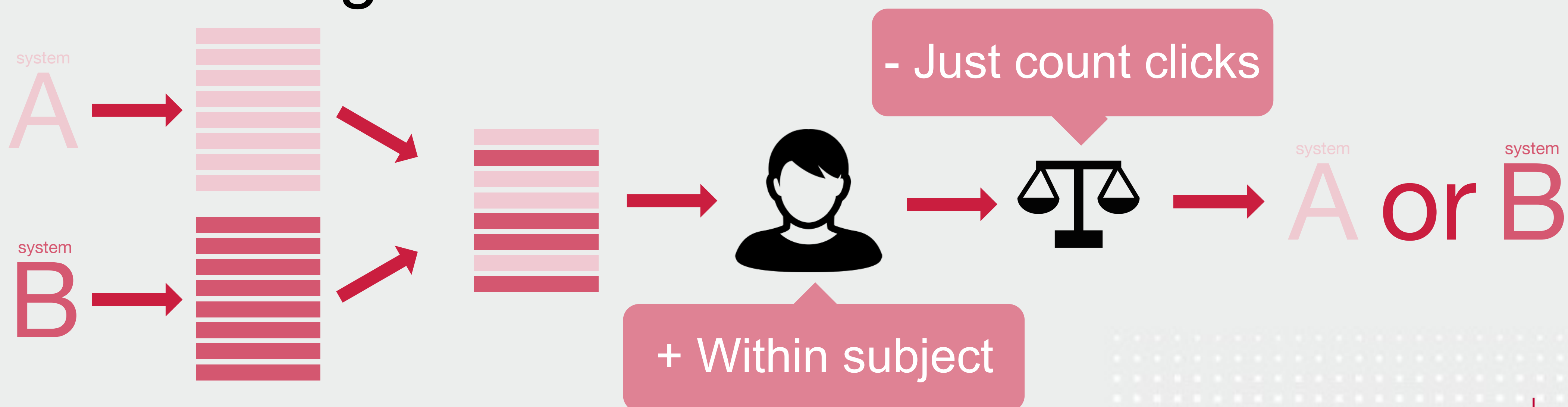


Motivation - **Agreement**

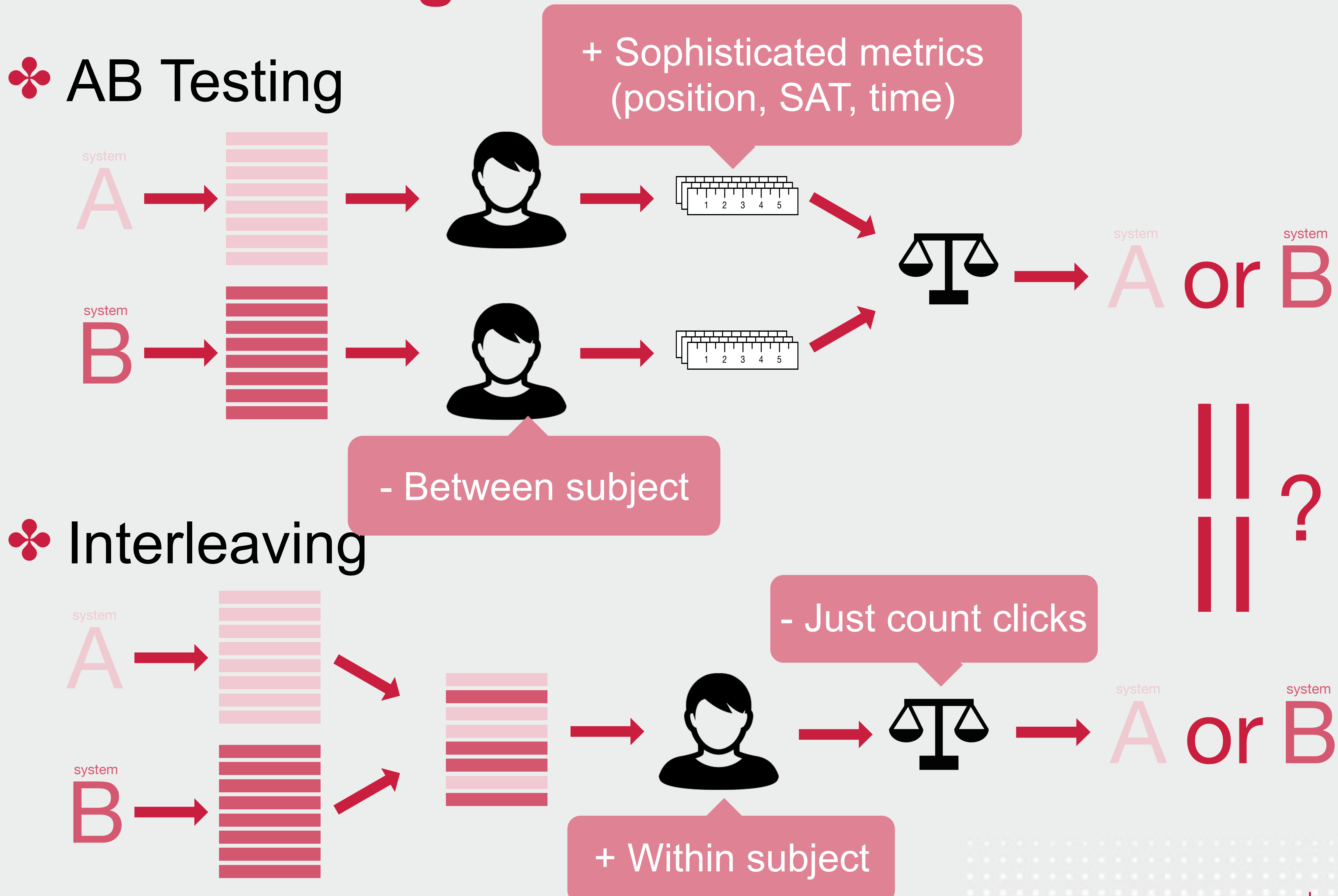
✤ AB Testing



✤ Interleaving



Motivation - **Agreement**



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 - **Properties**

These are our datapoints

- **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)**

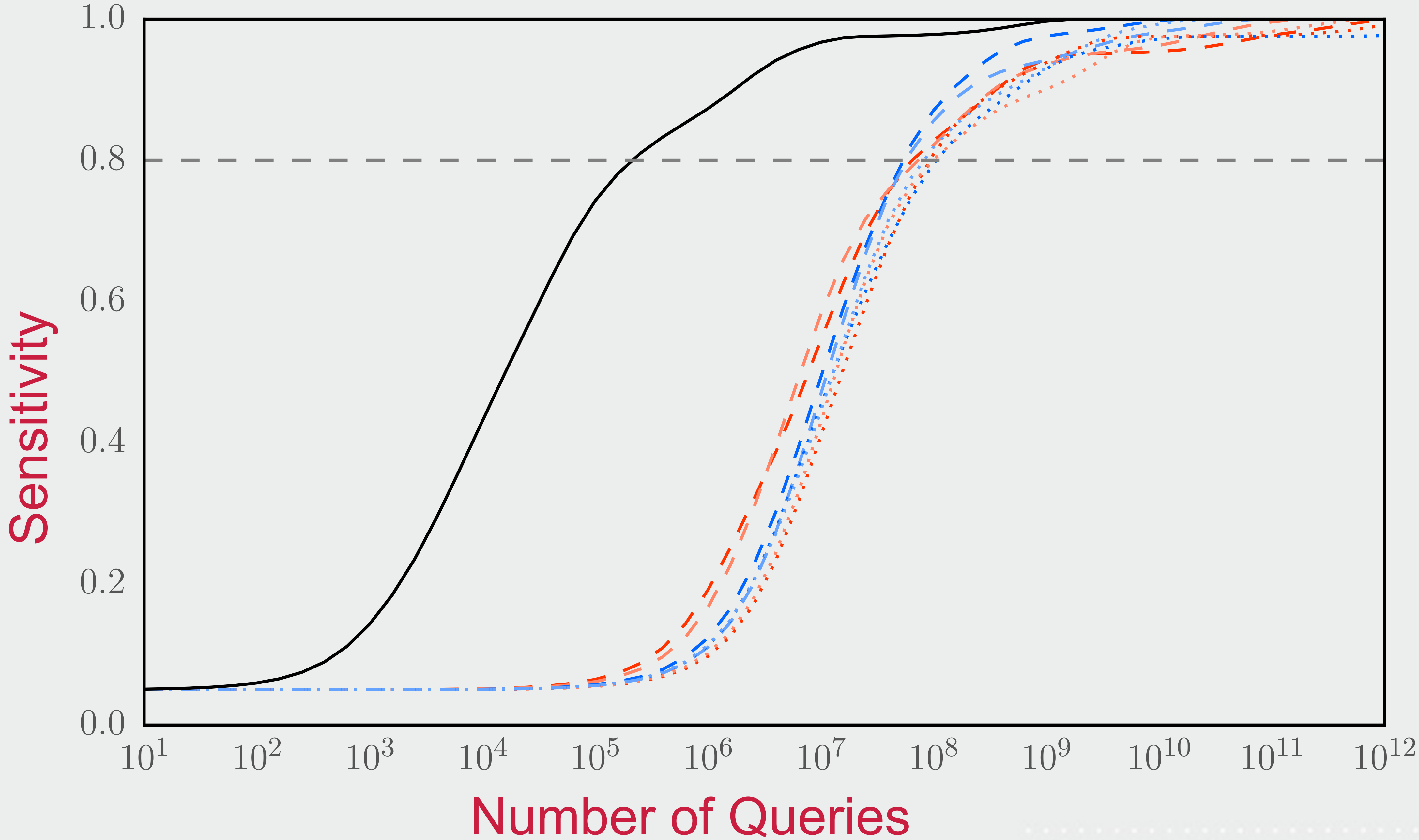
- **How many queries** are required for statistically significant conclusions?

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

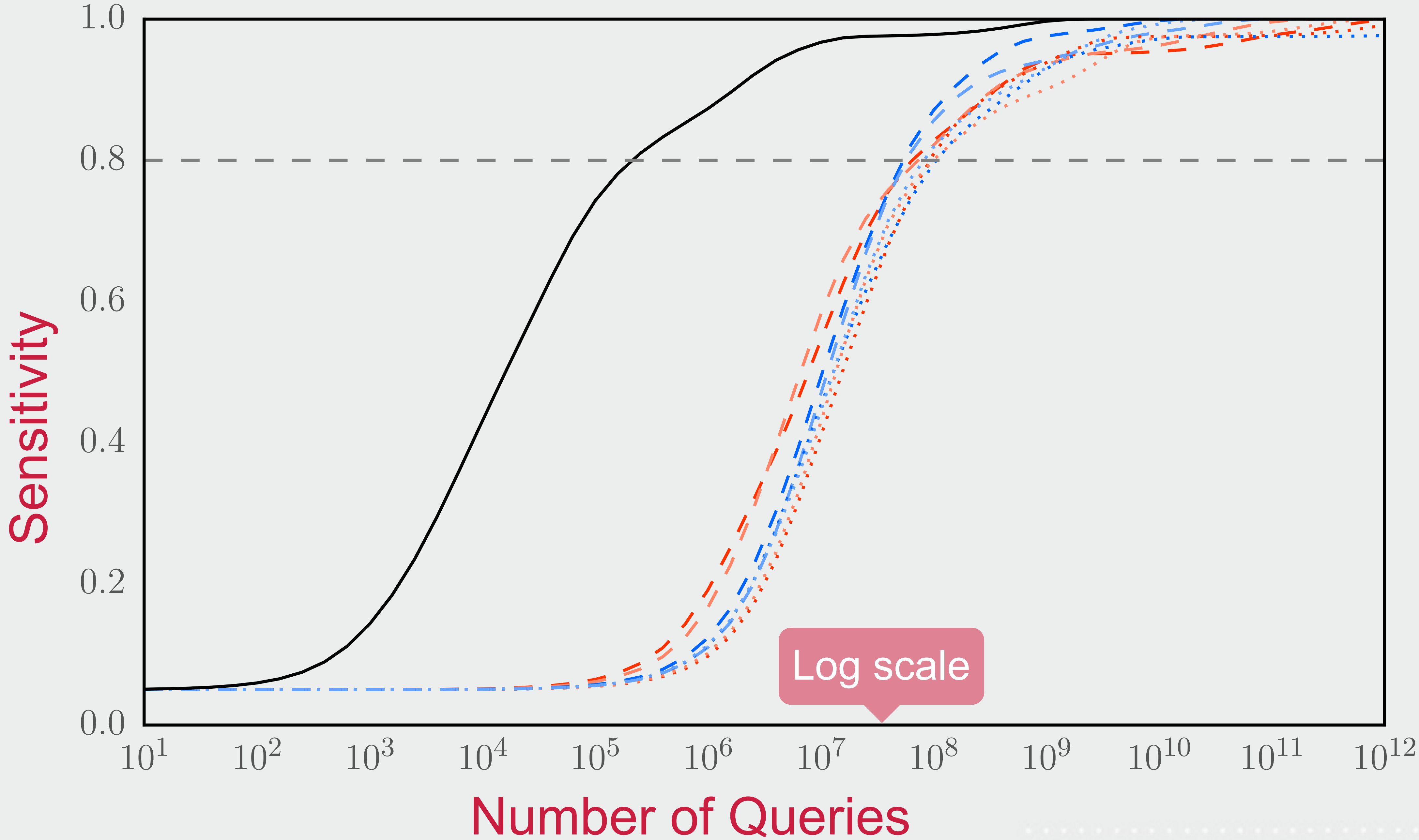
- **How many queries** are required for statistically significant conclusions?
- **Sensitivity** (power) **analysis**
  - alpha=0.05, two sided
  - AB Testing: **independent** t-test
  - Interleaving (TDI): **paired** t-test

University of Amsterdam

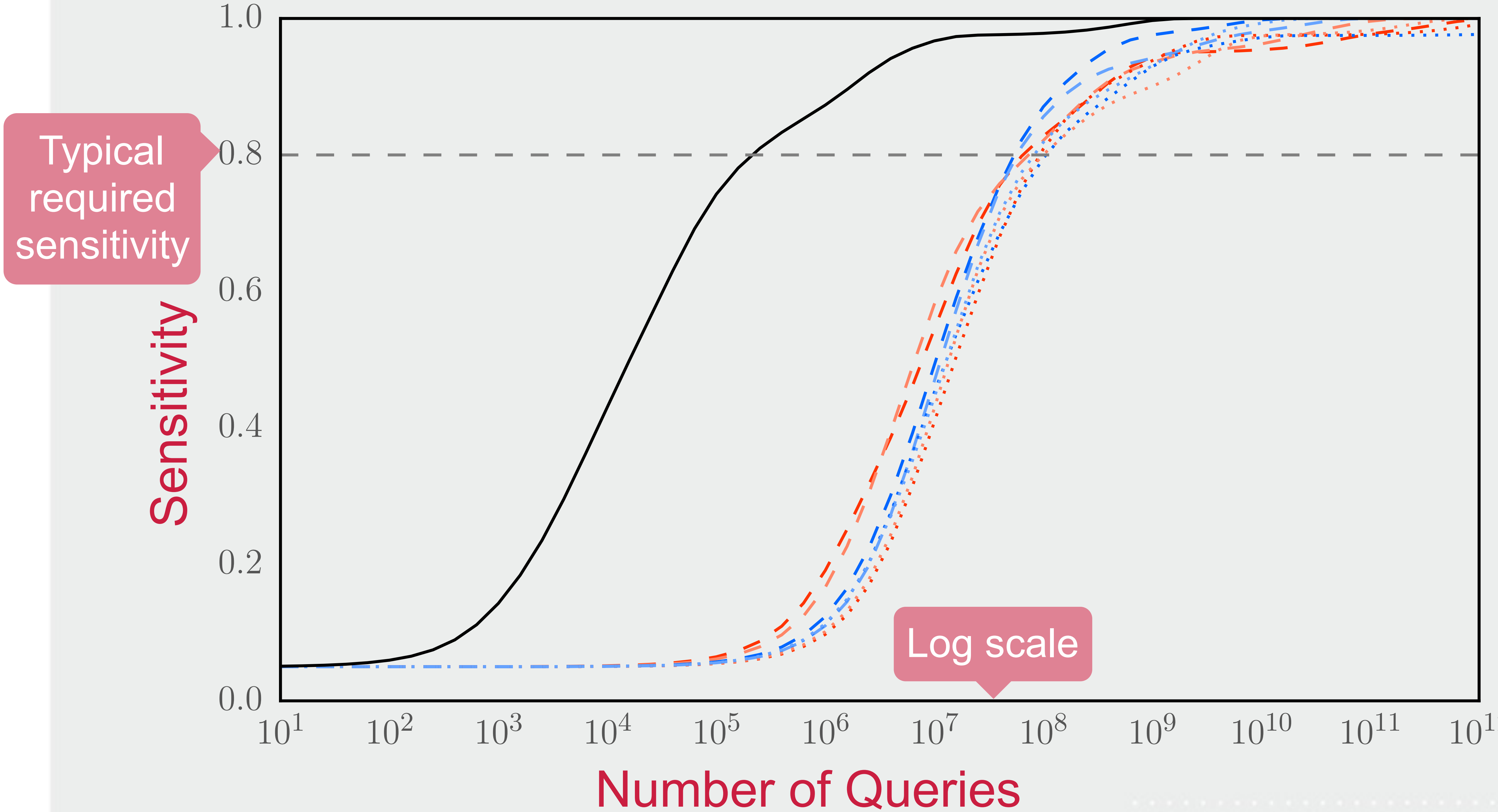
Data - Analysis - **Sensitivity**



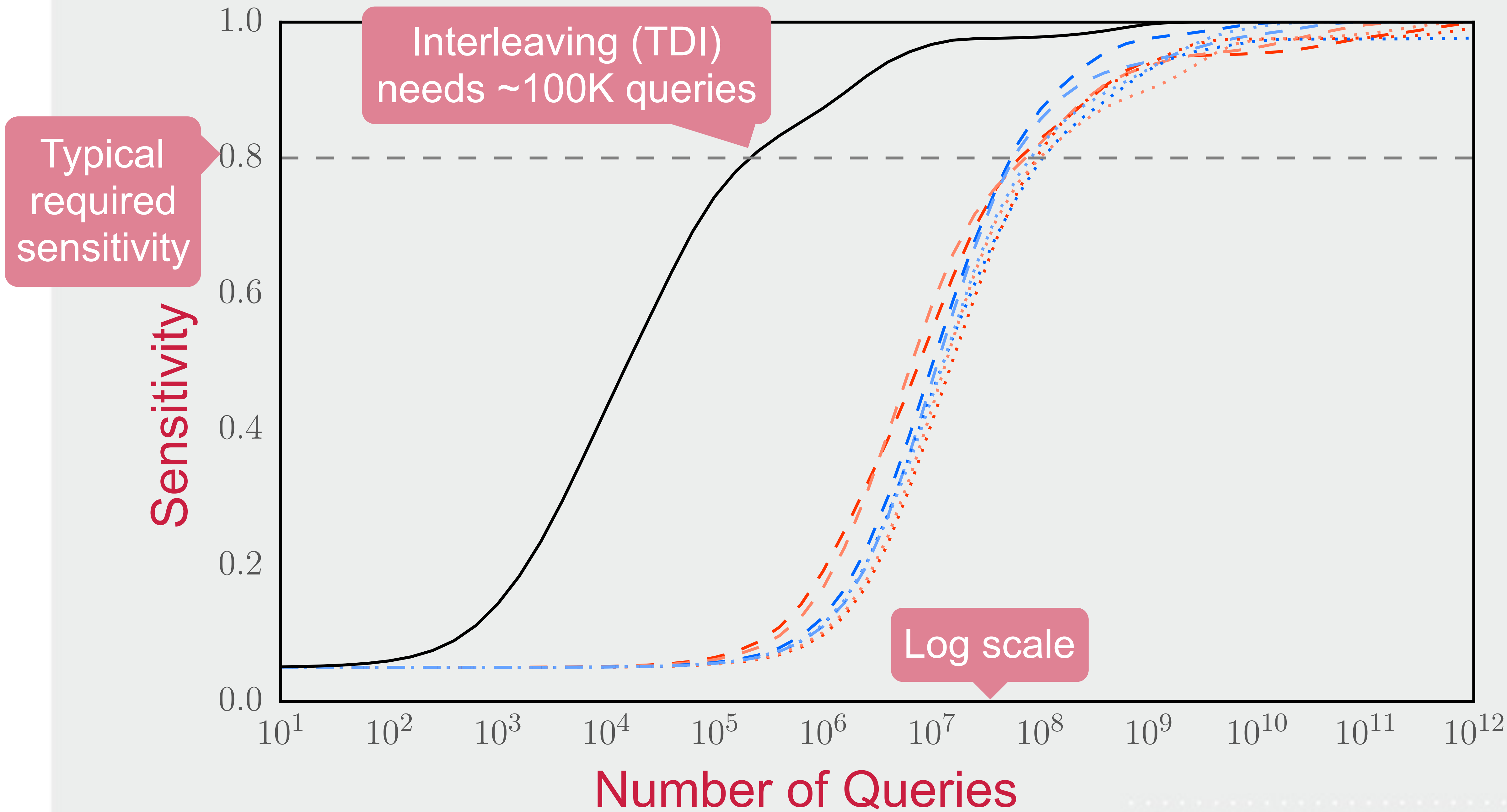
Data - Analysis - **Sensitivity**



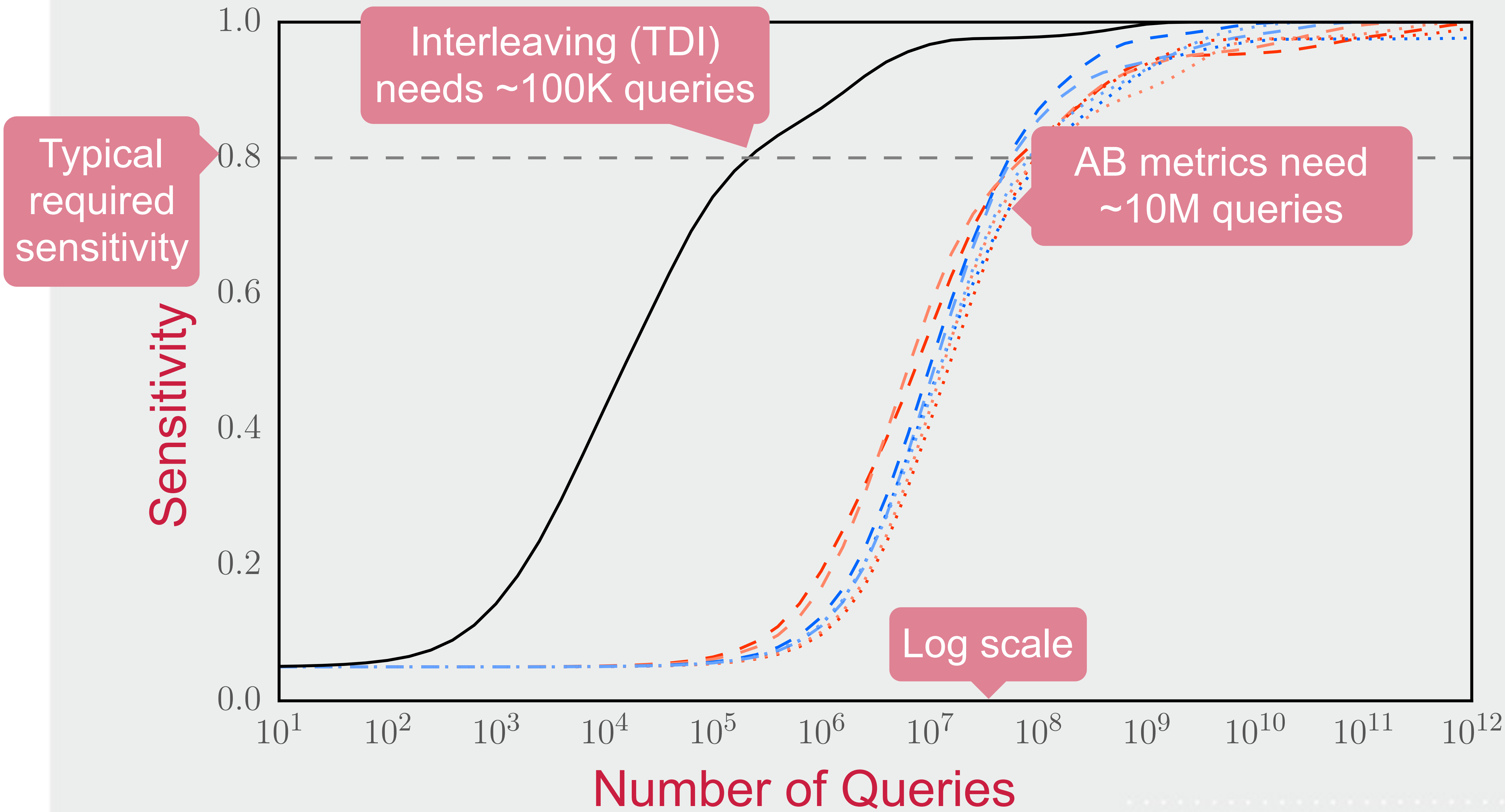
Data - Analysis - **Sensitivity**



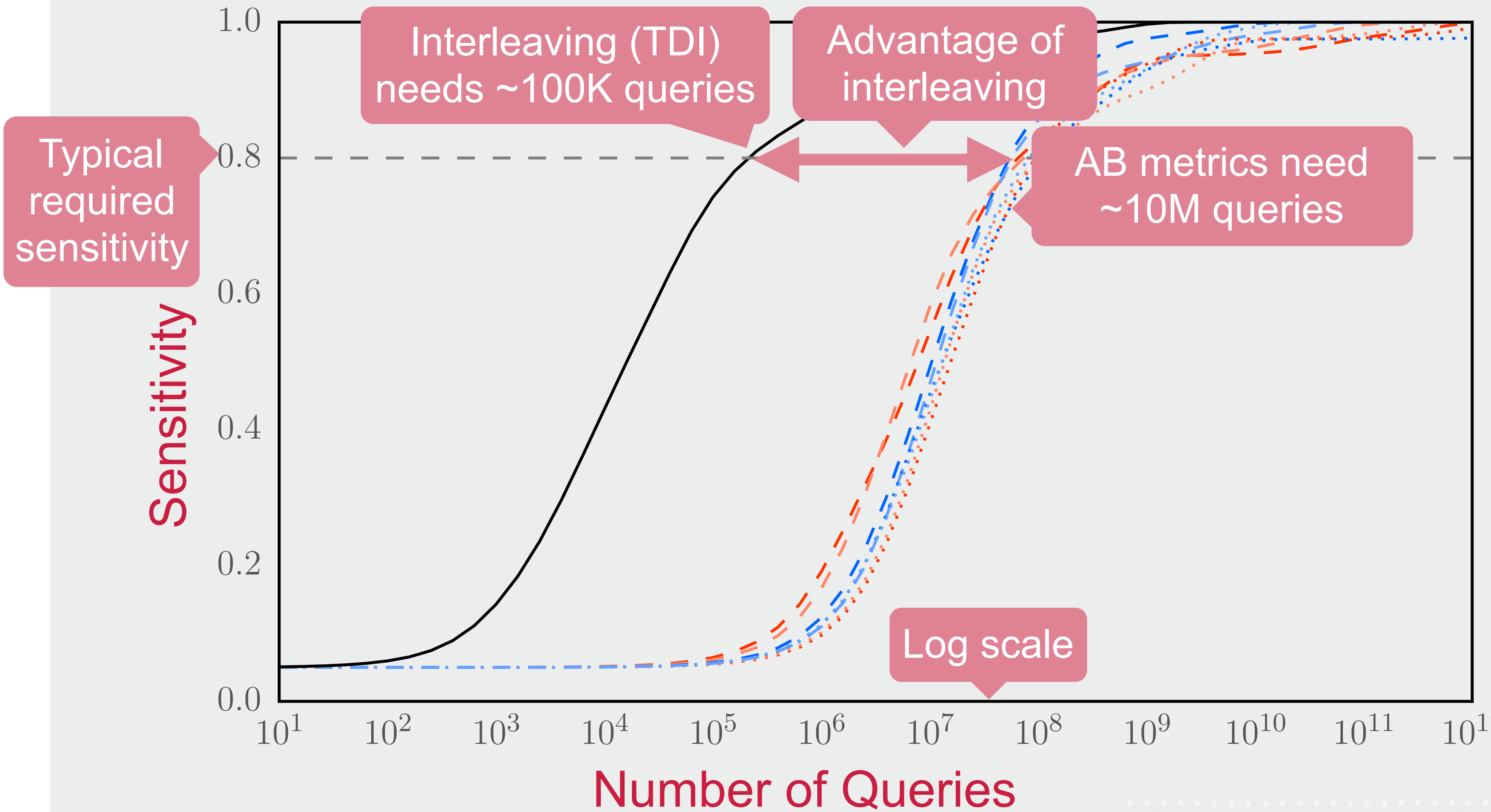
Data - Analysis - **Sensitivity**



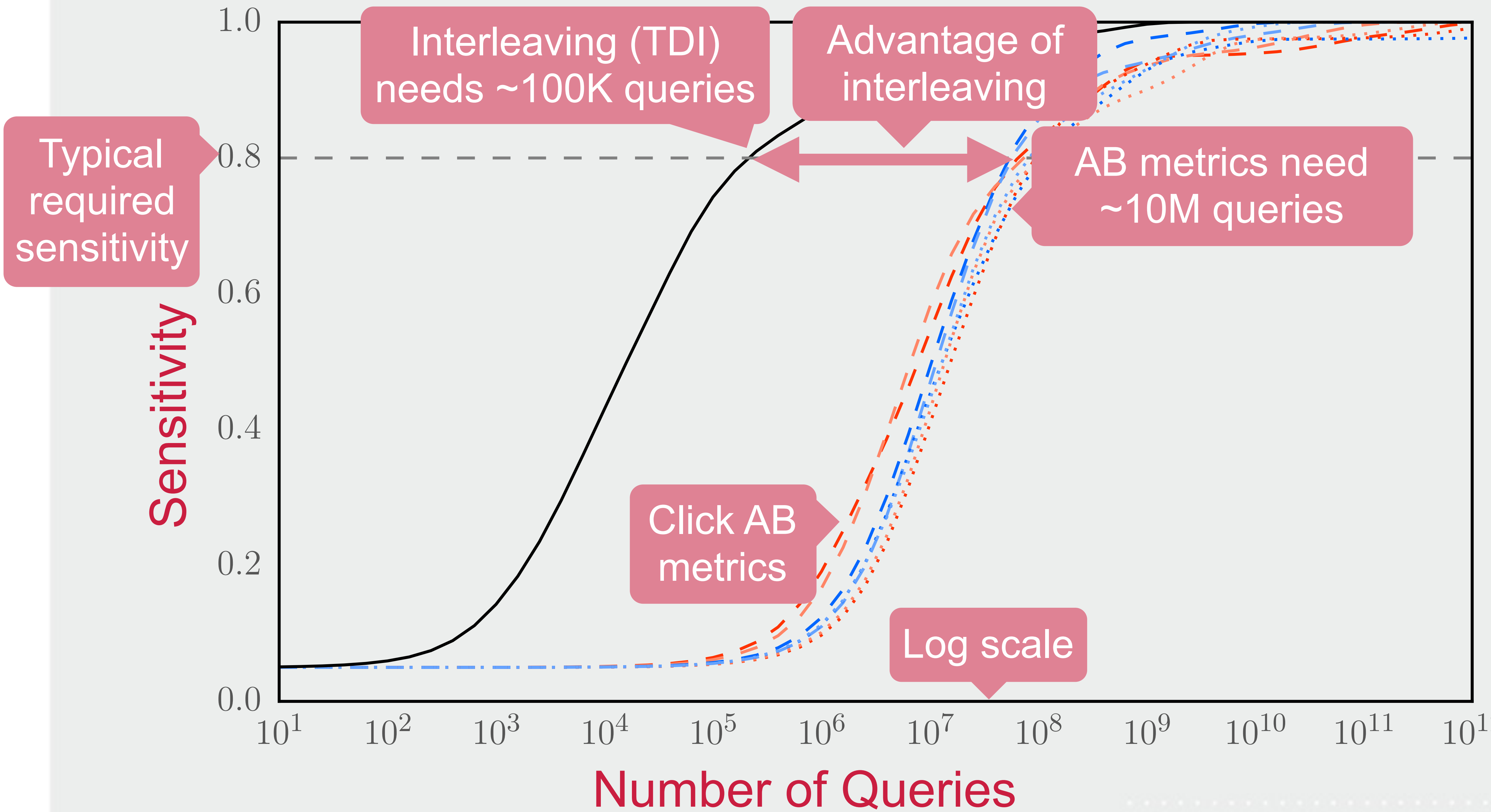
Data - Analysis - **Sensitivity**



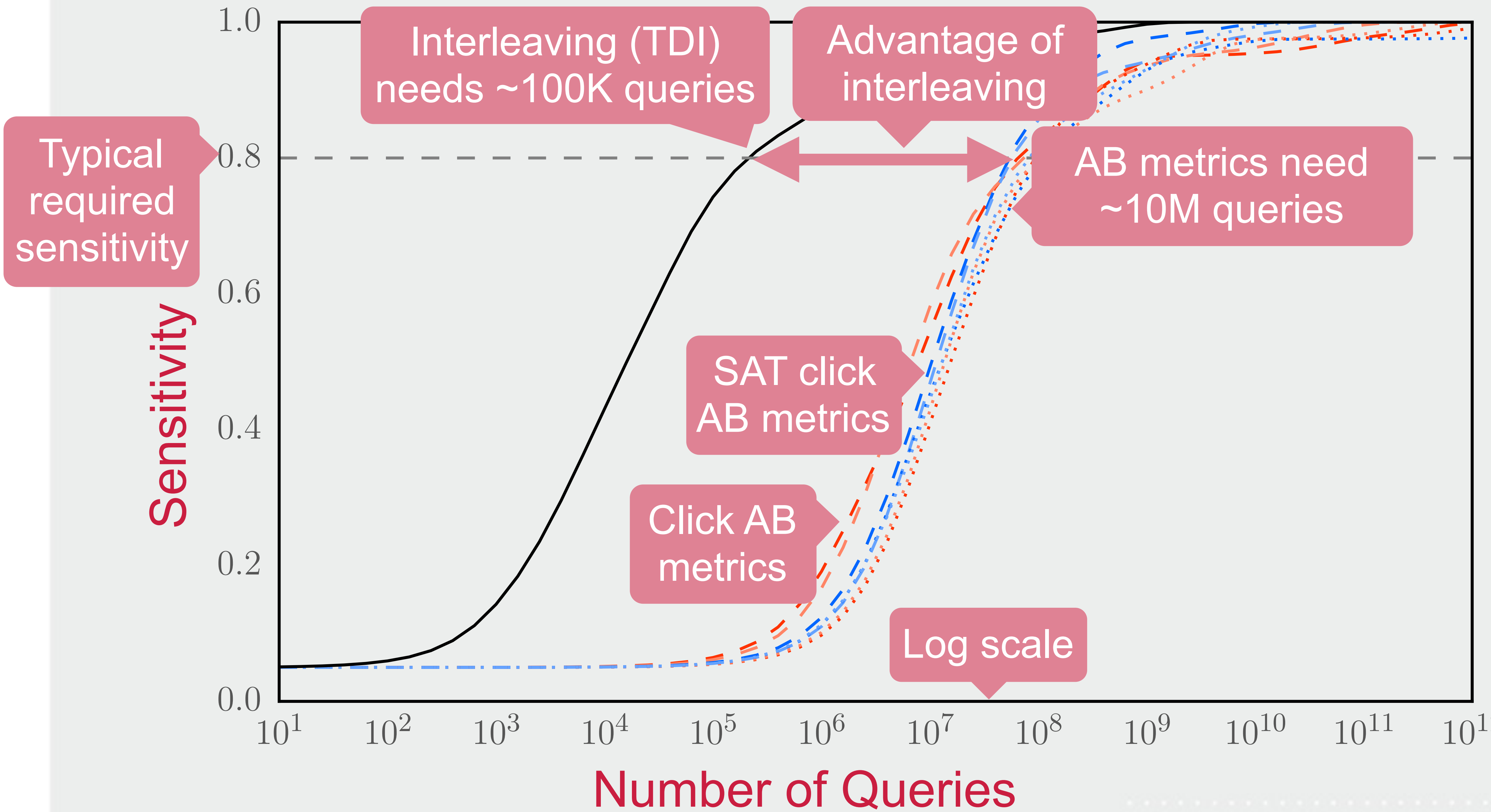
Data - Analysis - **Sensitivity**



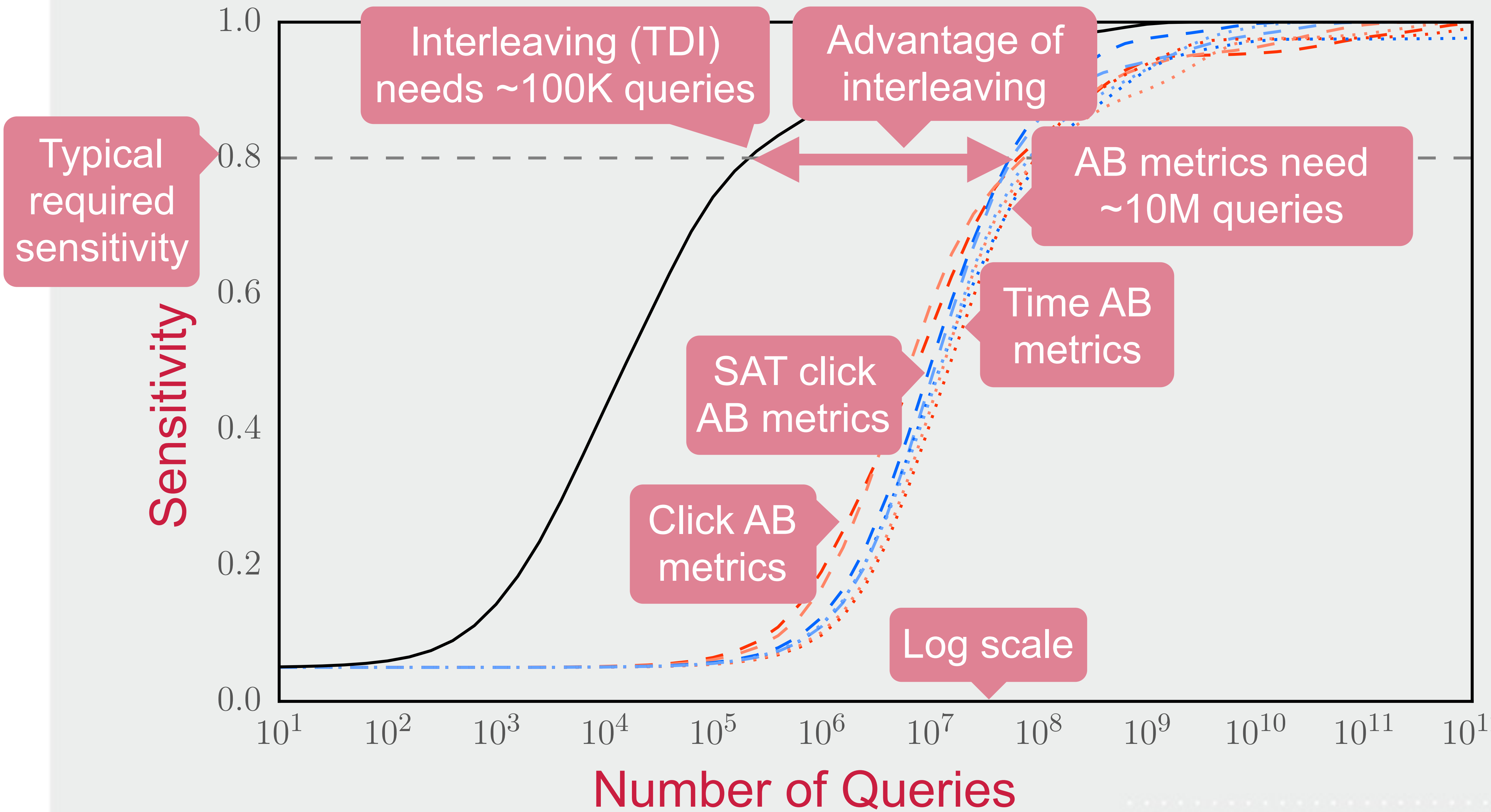
Data - Analysis - **Sensitivity**



Data - Analysis - **Sensitivity**



Data - Analysis - **Sensitivity**



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

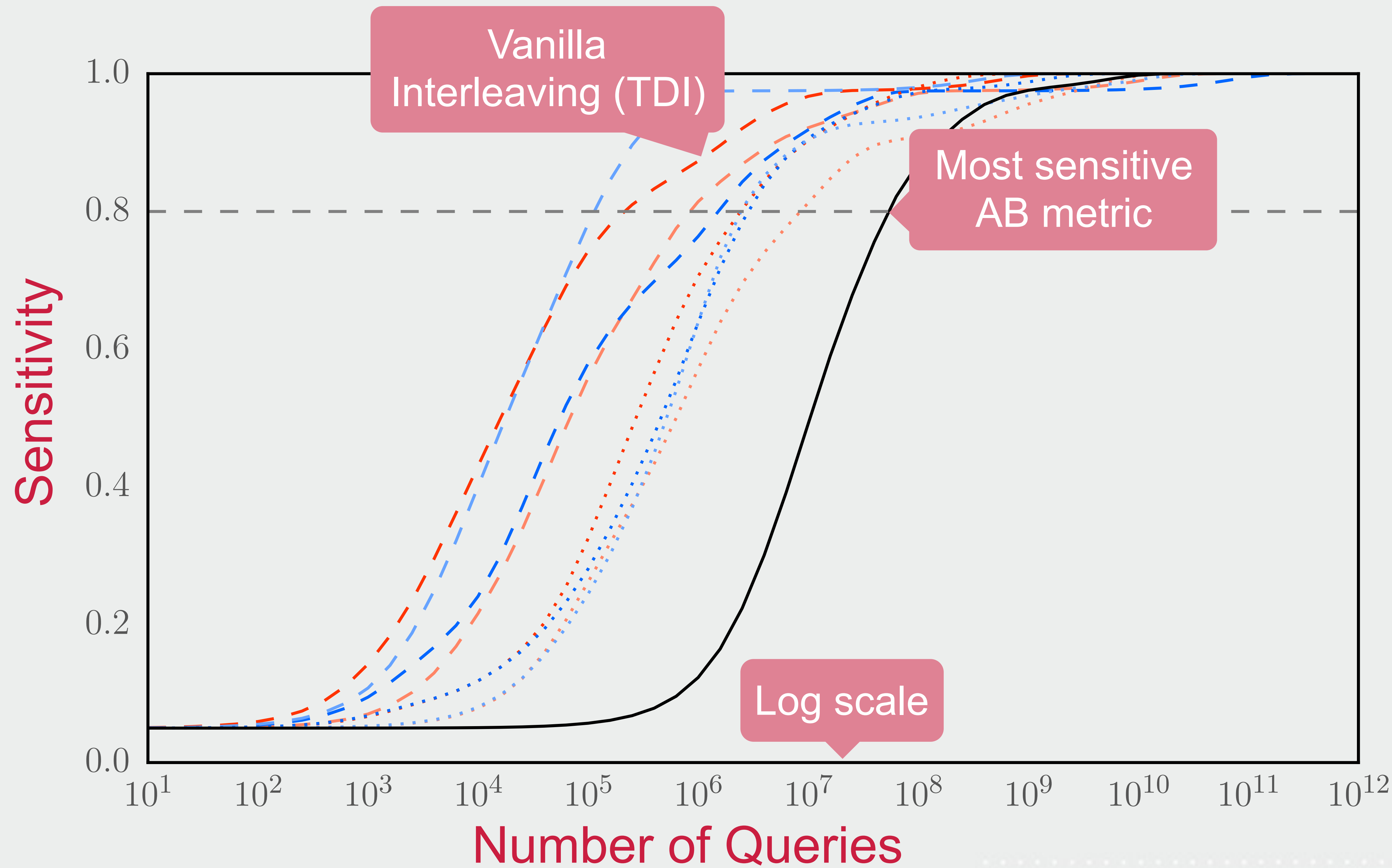
Filter out clicks,
can reduce sensitivity

Methods - **Matching AB Metric**

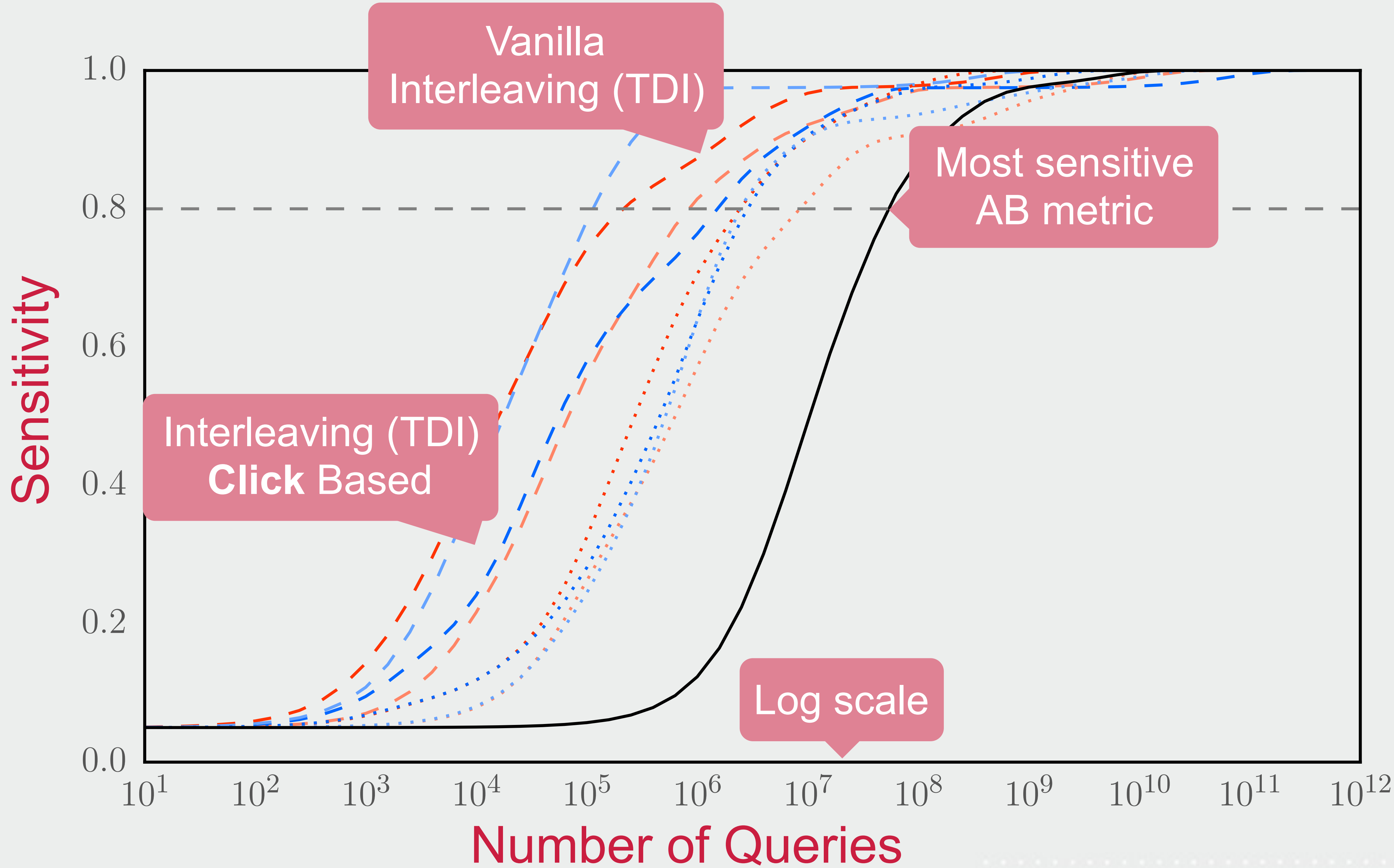
- **Interleaving** traditionally counts **all clicks**
- **Instead** of counting all clicks …
- … we propose to **match AB metrics**
  - Count only **certain** clicks
    - @1
    - SAT
  - Measure **time** to click

Filter out clicks,
can reduce sensitivity

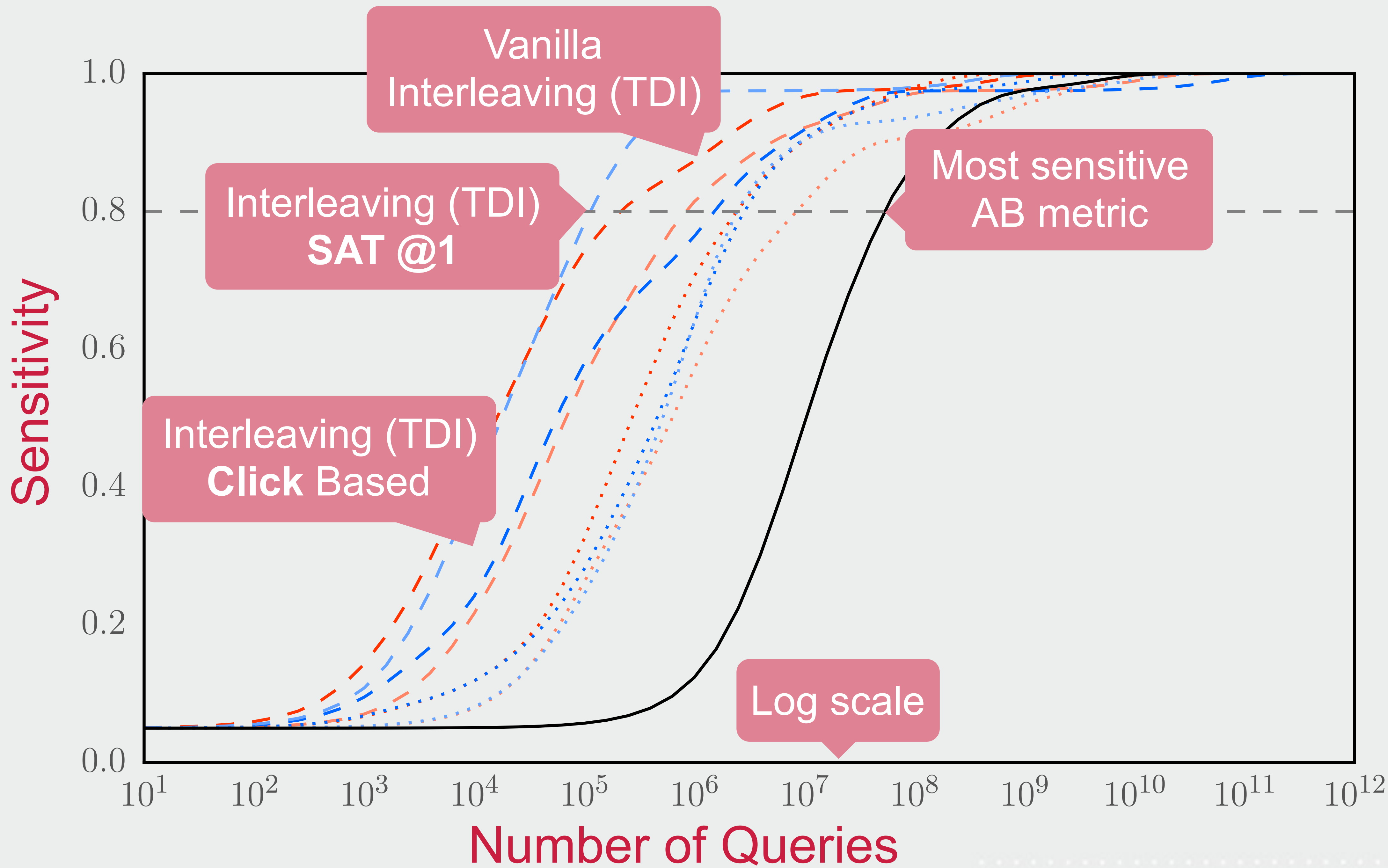
Methods - Matching AB Metric - **Sensitivity**



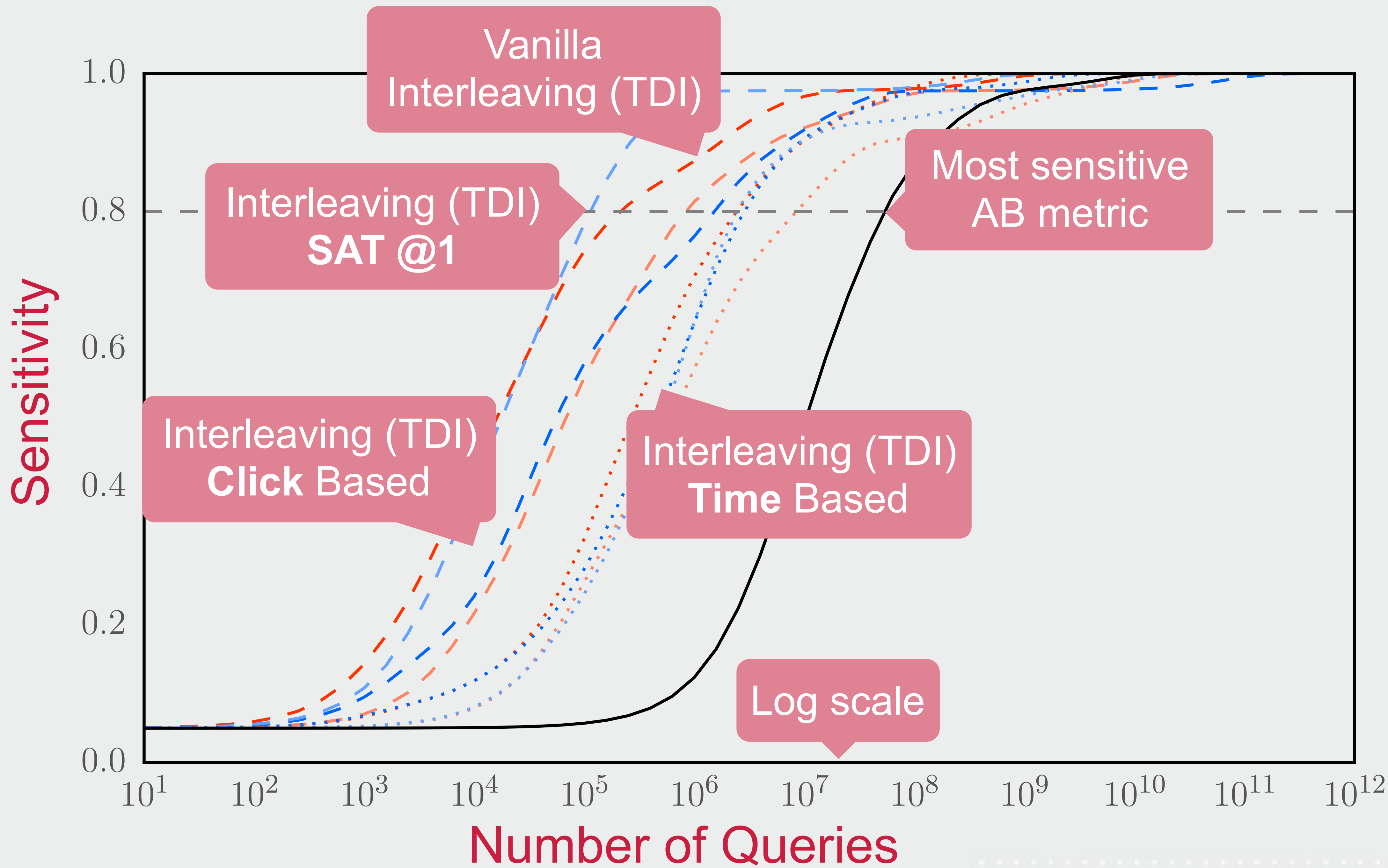
Methods - Matching AB Metric - **Sensitivity**



Methods - Matching AB Metric - **Sensitivity**



Methods - Matching AB Metric - **Sensitivity**



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**

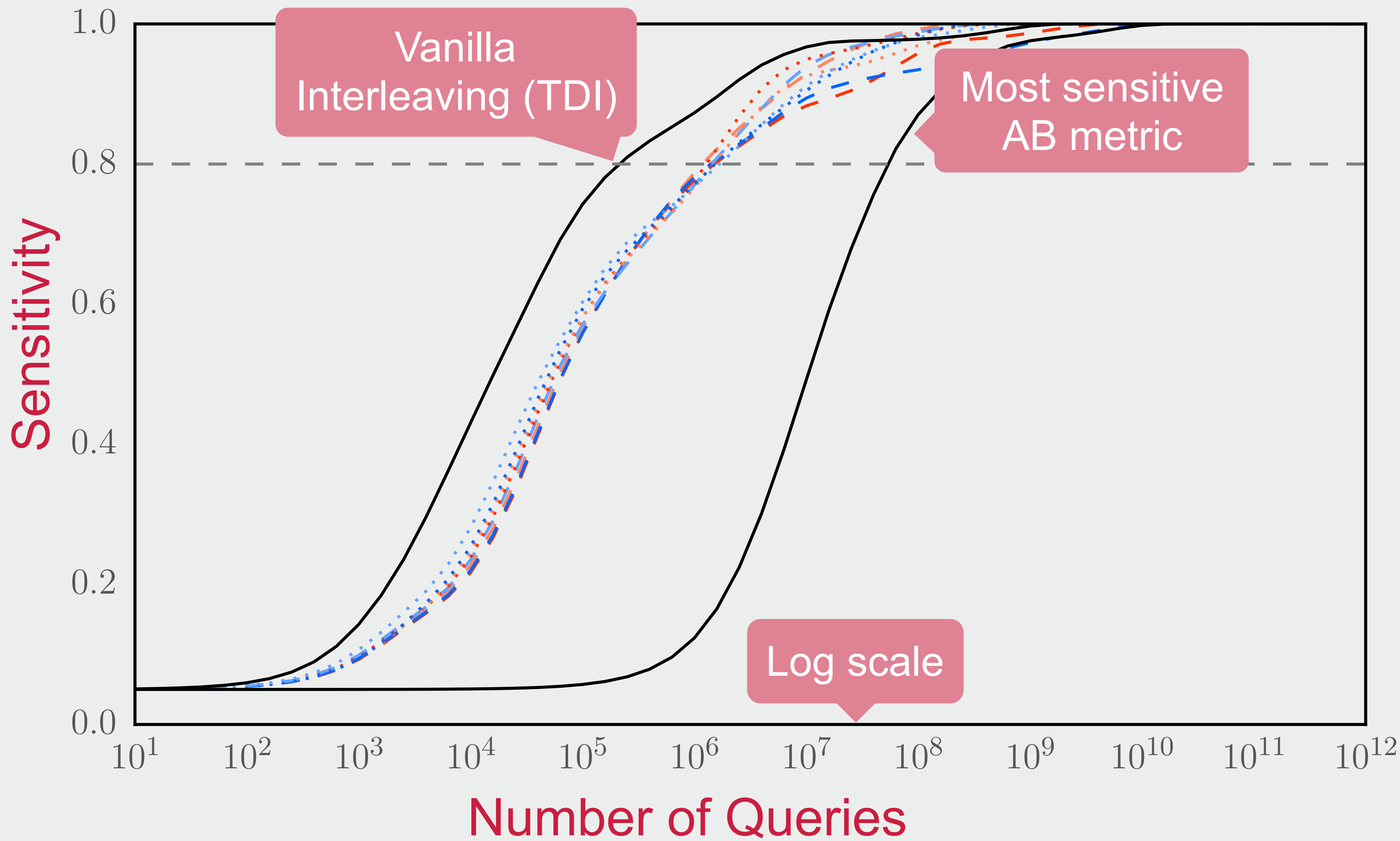
Remember, we have a model that predicts **SAT probability**

- 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

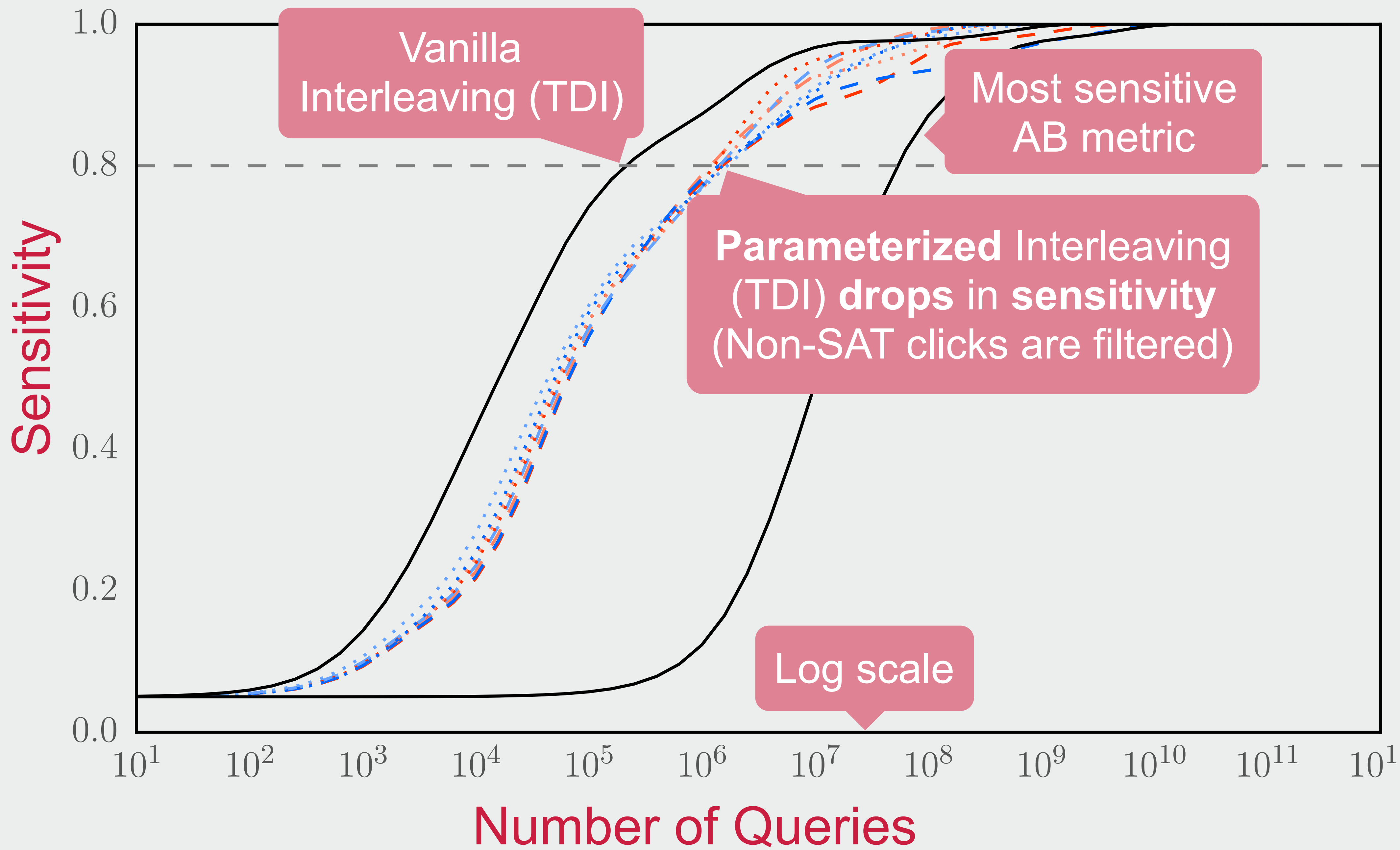
    Filter out non SAT clicks, **can reduce** sensitivity
- Find **optimal threshold t_S**
  - Maximize agreement for **each** AB metric
- Repeat n=100 times:
  - Take bootstrap sample
  - Grid search to find t_S that maximizes agreement
  - Report performance on “out of bag” sample

Methods - Parametrized Credit - **Sensitivity**



UNIVERSITY OF AMSTERDAM

Methods - Parametrized Credit - **Sensitivity**



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	
$AB_{T,S}$	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	**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**
$AB_{T,S}$	0.47	0.46
$AB_{T,S}$@1	0.42	0.52

Methods - Parametrized Credit - **Agreement**

AB Metric	TDI (Vanilla)	TDI_S^{ts} (Click based)	$TDI_{T,S}^{ts}$ (Time based)
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**

AB Metric	TDI (Vanilla)	TDI_S^{ts} (Click based)	$TDI_{T,S}^{ts}$ (Time based)
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**
  - $w_S \cdot TDI_S^{ts} + w_T \cdot TDI_{T,S}^{ts}$

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

University of Amsterdam

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**

AB Metric	TDI	$TDI_{T,S}^{W}$ agreement	Click weight w_S	Time weight 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**

AB Metric	TDI	$TDI_{T,S}^{W}$ agreement	Click weight w_S	Time weight 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**

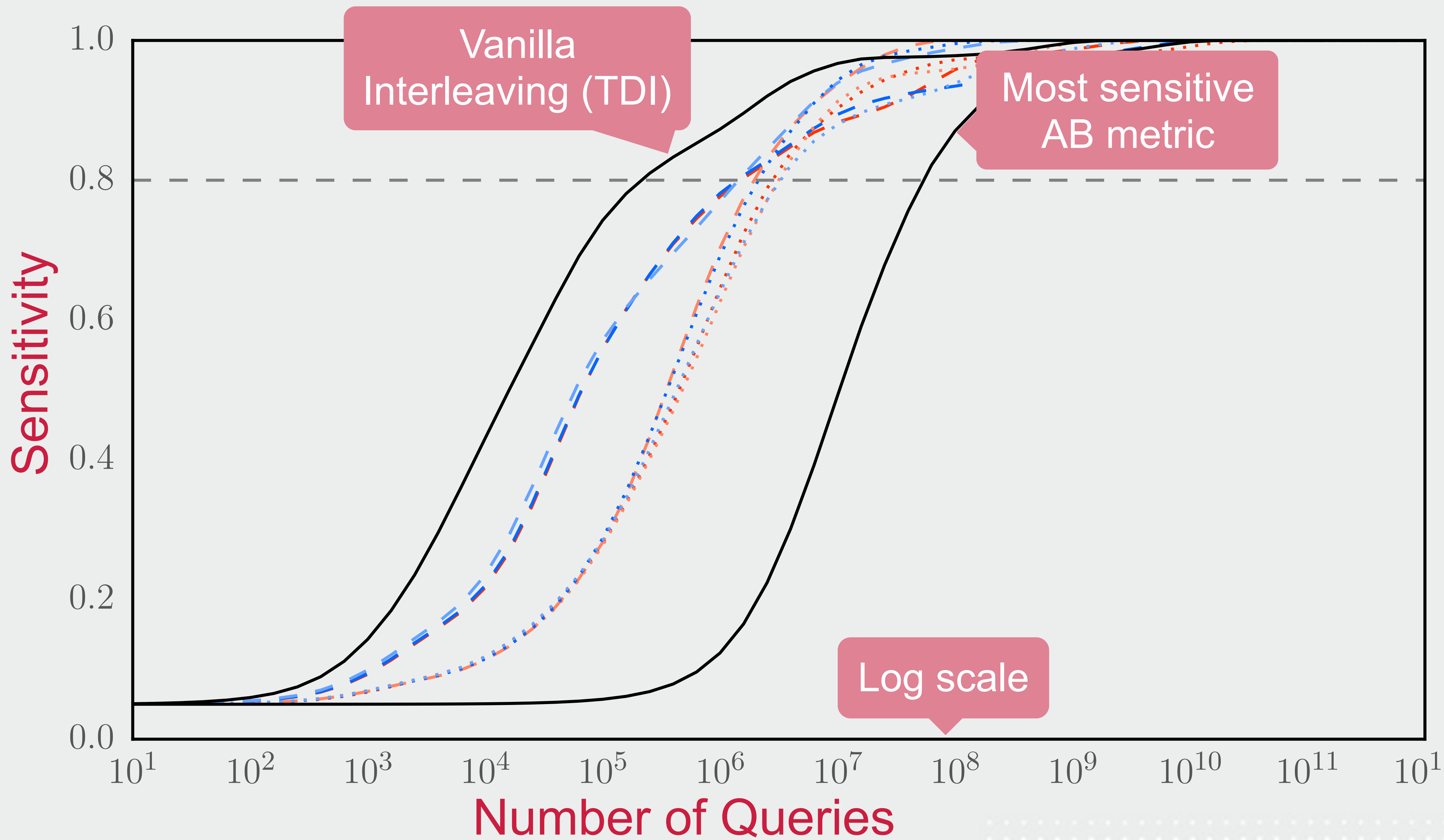
		$TDI_{T,S}^{W}$	Click weight	Time weight
AB Metric	**TDI**	**agreement**	w_S	w_T
AB	0.63	**0.84**	*1.00*	*0.00*
AB@1	0.71	**0.75**	*1.00*	*0.05*
AB_S	0.71	**0.85**	*1.00*	*0.00*
AB_S@1	0.76	**0.83**	*1.00*	*0.02*
AB_T	0.53	**0.68**	*0.99*	*0.90*
AB_T@1	0.45	0.56	*0.96*	*0.79*
$AB_{T,S}$	0.47	0.63	*0.91*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

Methods - Combined Credit - **Agreement**

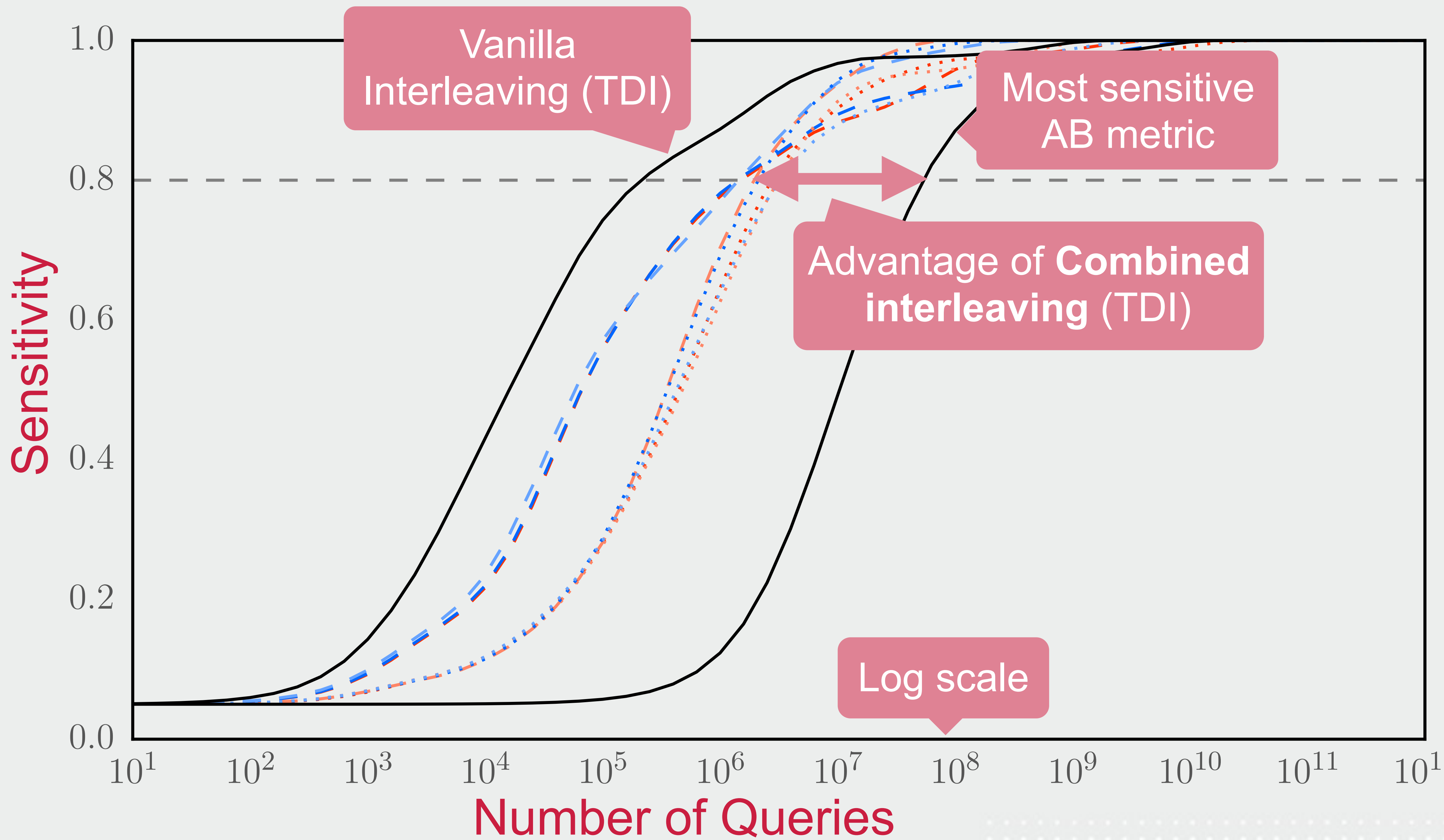
		$TDI_{T,S}^{W}$	Click weight	Time weight
AB Metric	TDI	agreement	w_S	w_T
AB	0.63	**0.84**	*1.00*	*0.00*
AB@1	0.71	**0.75**	*1.00*	*0.05*
AB_S	0.71	**0.85**	*1.00*	*0.00*
AB_S@1	0.76	**0.83**	*1.00*	*0.02*
AB_T	0.53	**0.68**	*0.99*	*0.90*
AB_T@1	0.45	0.56	*0.96*	*0.79*
$AB_{T,S}$	0.47	0.63	*0.91*	*0.88*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

All significantly better than TDI

Methods - Combined Credit - **Sensitivity**

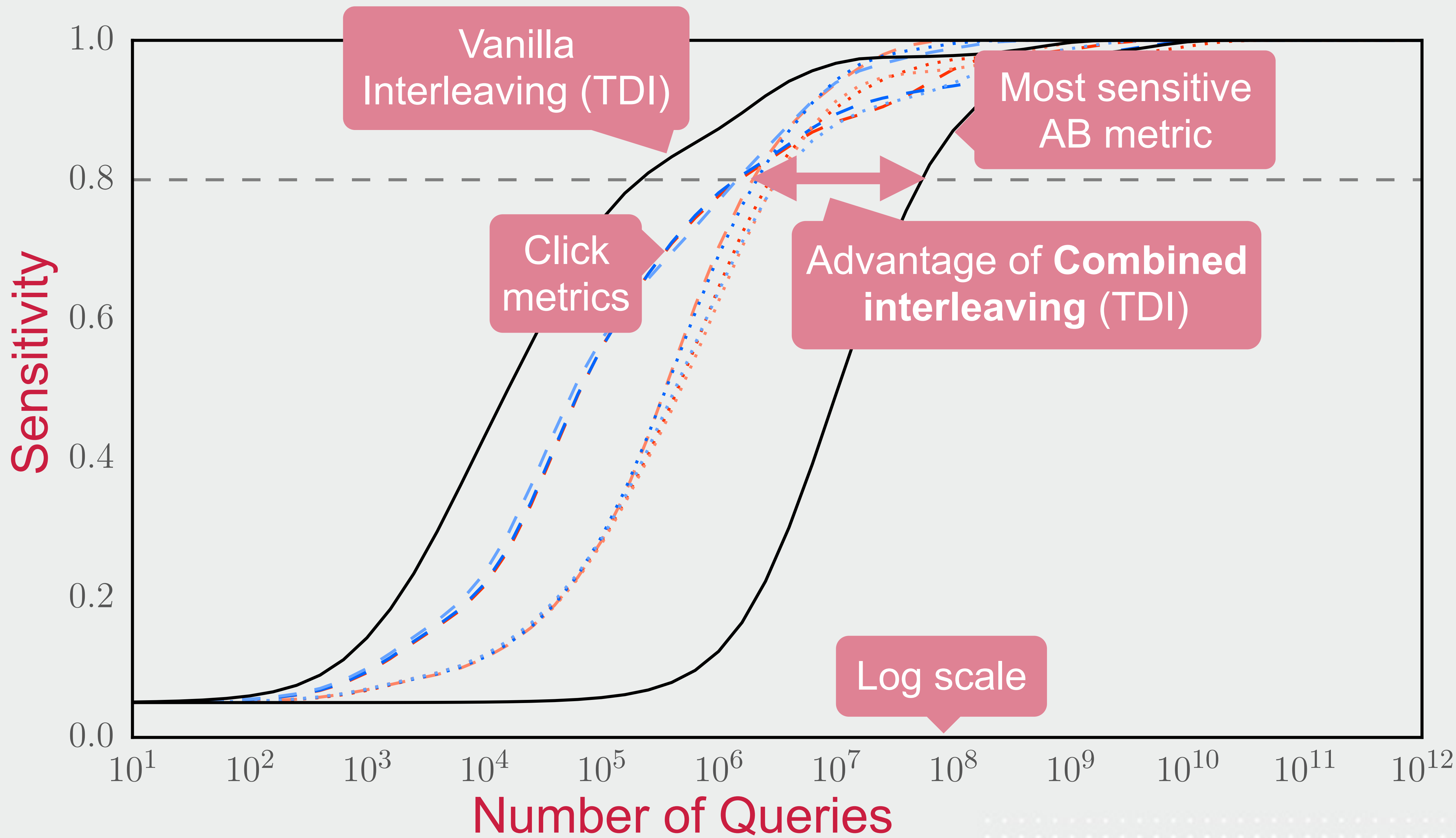


Methods - Combined Credit - **Sensitivity**

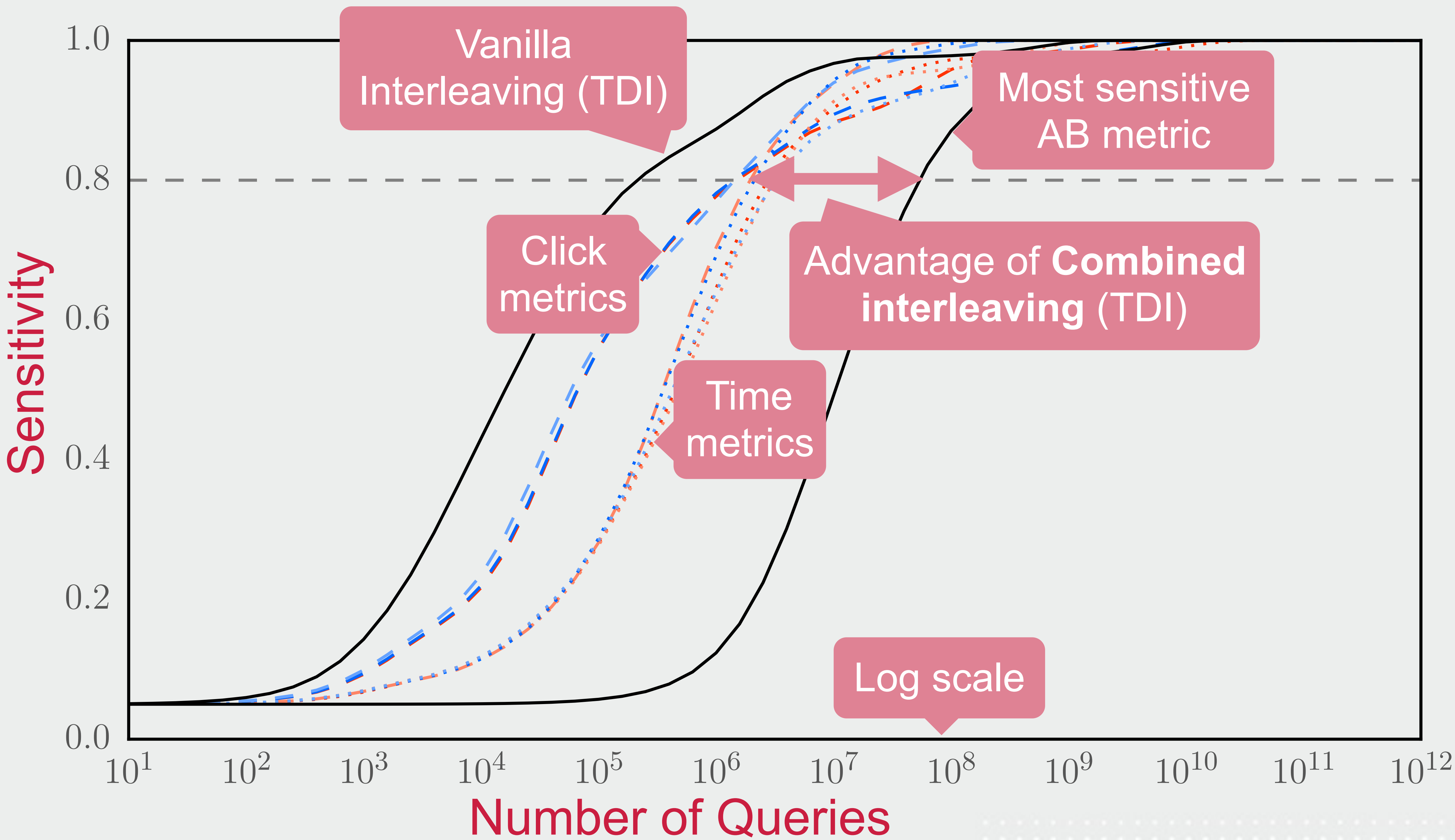


University of Amsterdam

Methods - Combined Credit - **Sensitivity**



Methods - Combined Credit - **Sensitivity**



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 just credit **matching** AB metrics
  - **Unpredictable** performance

Conclusions - **Methods**

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

Conclusions - **Methods**

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

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**
- Apply to **other types of ranking systems**

Take Away

Take Away

- **Richer** user **signals** in **interleaving**

Take Away

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

Take Away

- **Richer** 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

- **Richer** user **signals** in **interleaving**
- **Agreement** of interleaving with an AB metric can be made as high as **87%**
- While maintaining **high sensitivity** of interleaving
- **Weak signals can be measured with a strong (but biased) proxy**

UNIVERSITY OF AMSTERDAM

Take Away

- **Richer** user **signals** in **interleaving**
- **Agreement** of interleaving with an AB metric can be made as high as **87%**
- While maintaining **high sensitivity** of interleaving
- **Weak signals can be measured with a strong (but biased) proxy**

Supported by ACM SIGIR Travel Grant

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