BLENDER: Enabling Local Search with a Hybrid Differential Privacy Model

Brendan Avent[1], Aleksandra Korolova[1], David Zeber[2], Torgeir Hovden[2], Benjamin Livshits[3]

University of Southern California[1] Mozilla[2] Imperial College London[3]

Full paper available here.

Local Search

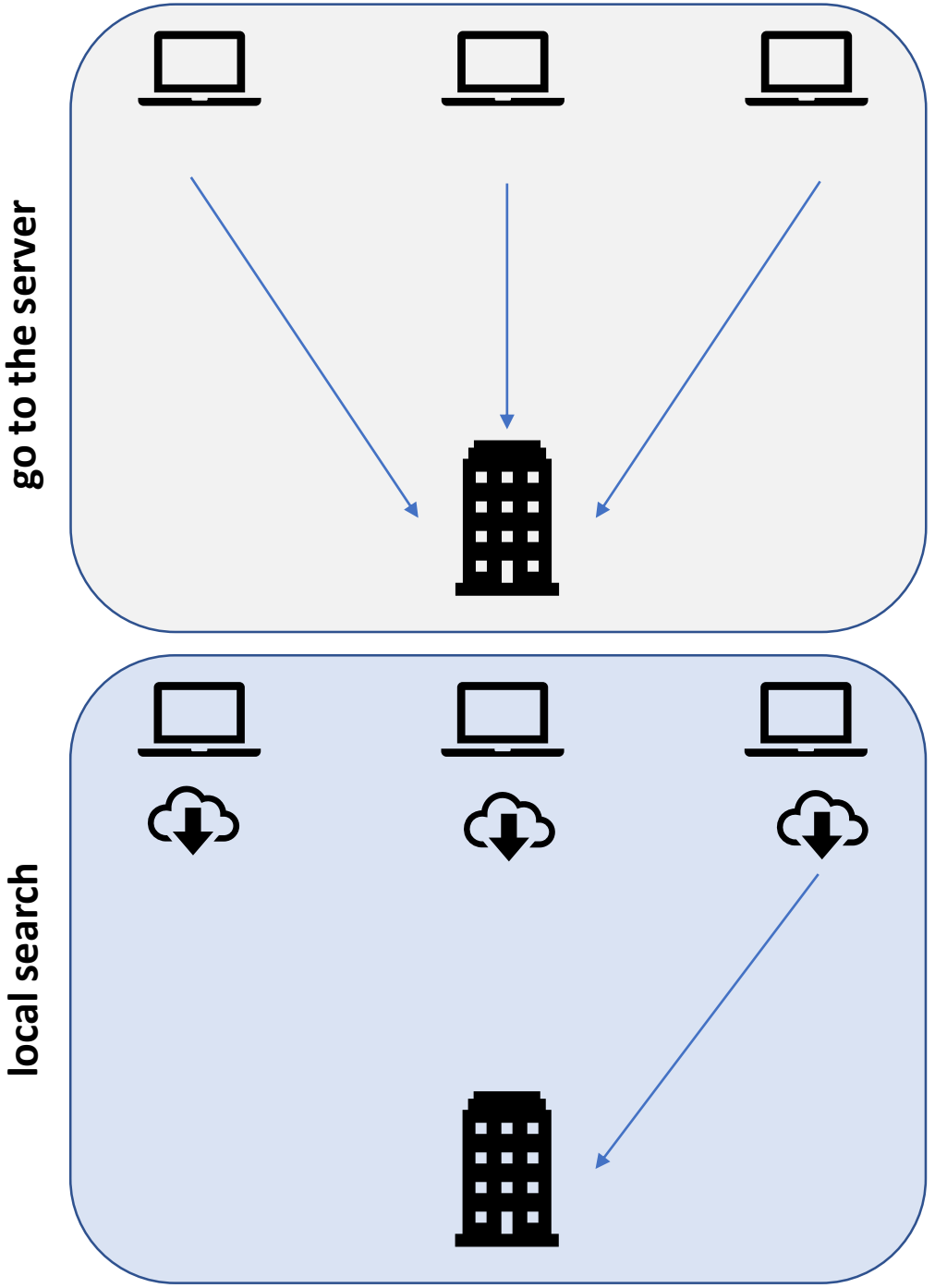
Goal

To make popular queries and their corresponding URLs available *locally* on users' devices

Why its needed?

Caching popular search data avoids many round-trips to a server

- Reduces latency in web-browsing
- Useful for temporary network disruptions
- Enables new browser features



Local Search with Privacy

Why is privacy needed?

- Local search is generated from user data
- Want differential privacy guarantees

Local Search with Privacy

Why

-
-

Algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private iff for all neighboring databases D and D' differing in the value of precisely one user's data, the following inequality is satisfied for all possible sets of outputs $Y \subseteq Range(\mathcal{A})$:

$$\Pr[\mathcal{A}(D) \in Y] \leq e^{\epsilon} \Pr[\mathcal{A}(D') \in Y] + \delta$$

Local Search with Privacy

Why is privacy needed?

- Local search is generated from user data
- Want differential privacy guarantees

Why is differentially private local search hard?

Differential Privacy Models

trusted curator model

- Central curator collects the data from all users, then performs privatization
- Most differentially private algorithms **are** in this model

Requires the users to trust the curator with their private data

local model

- Each user privatizes their own data, then sends it to a central curator
- Requires less trust from users

Harsh utility trade-offs compared to trusted curator model algorithms
[Chan et al 2012; Duchi et al 2013; Kairouz et al 2014, 2016]

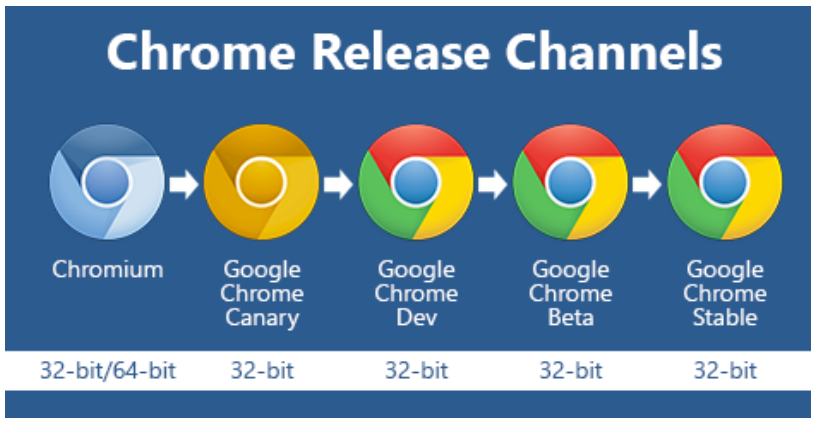
Hybrid Model

a more realistic privacy model

Users Have Heterogeneous Privacy Preferences

Firefox Browser Privacy Notice

Our pre-release versions (Beta/Developer Edition, Nightly, and TestFlight) may have different privacy characteristics. Pre-release versions automatically send Telemetry data to Mozilla.

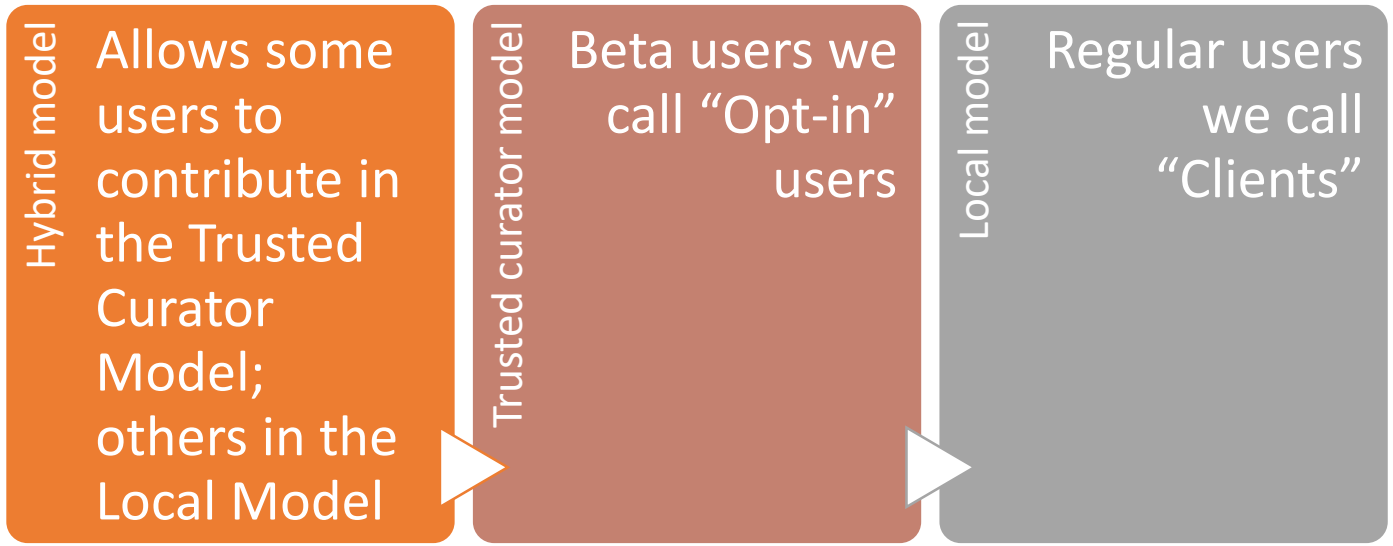


the INQUIRER

ificial Intelligence | Internet of Things | Open Source | Hardware | Software | Securi

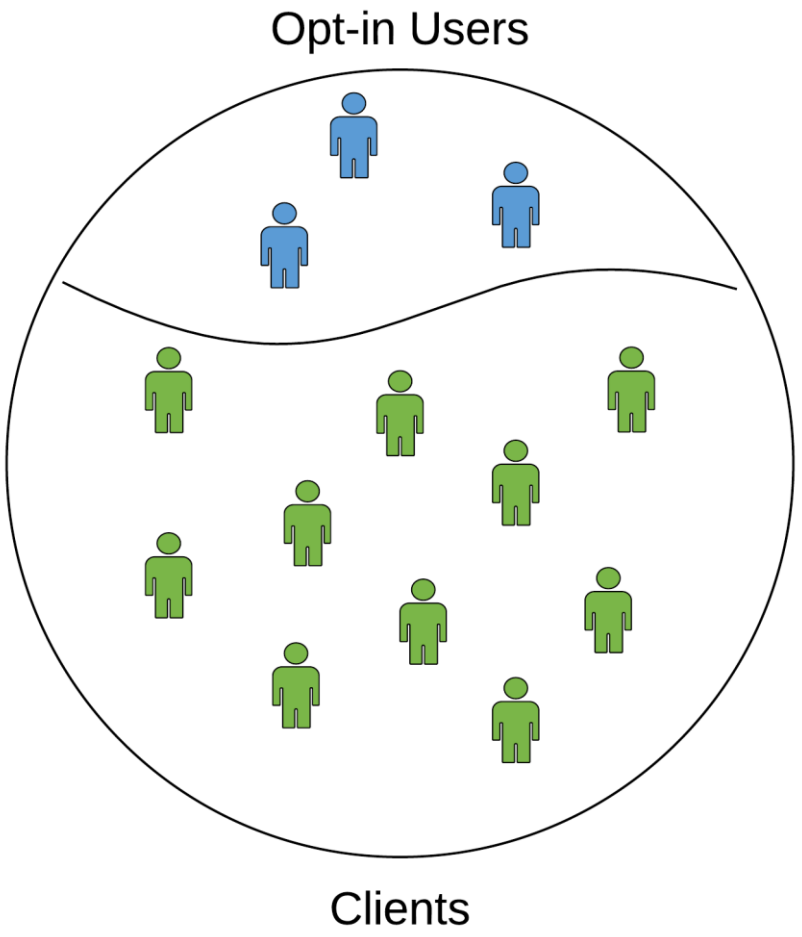
Microsoft reminds privacy-concerned Windows 10 beta testers that they're volunteers

If you don't like it, don't participate

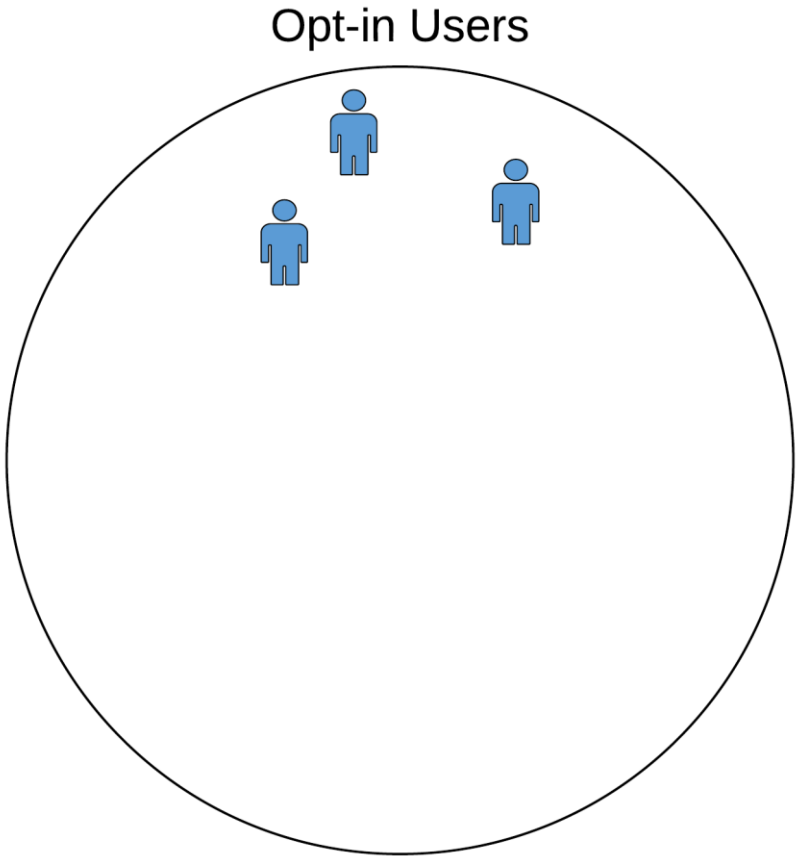


Hybrid Model for Differential Privacy

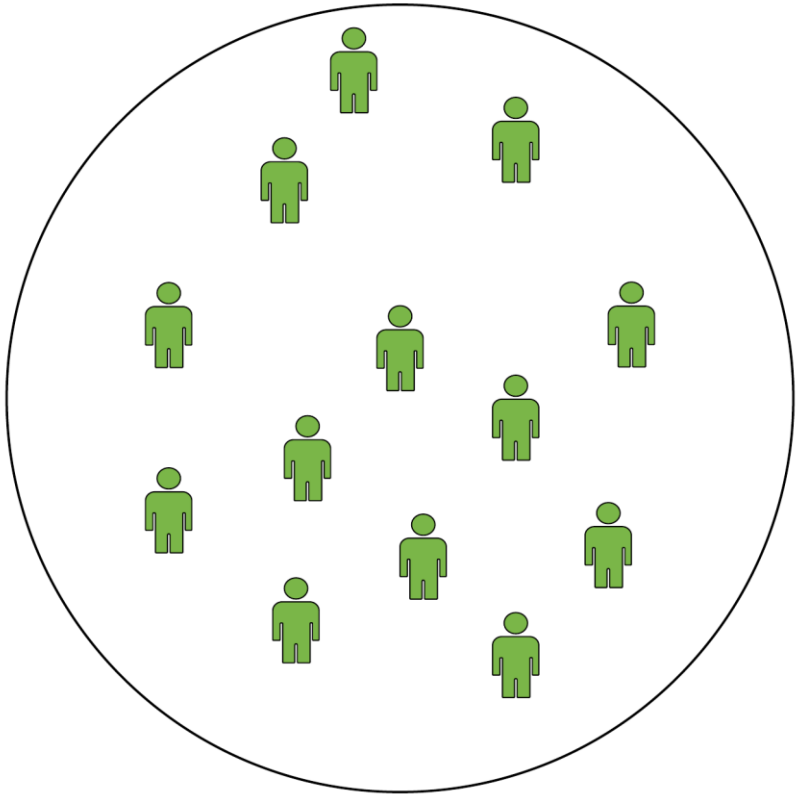
Why a Hybrid Model?



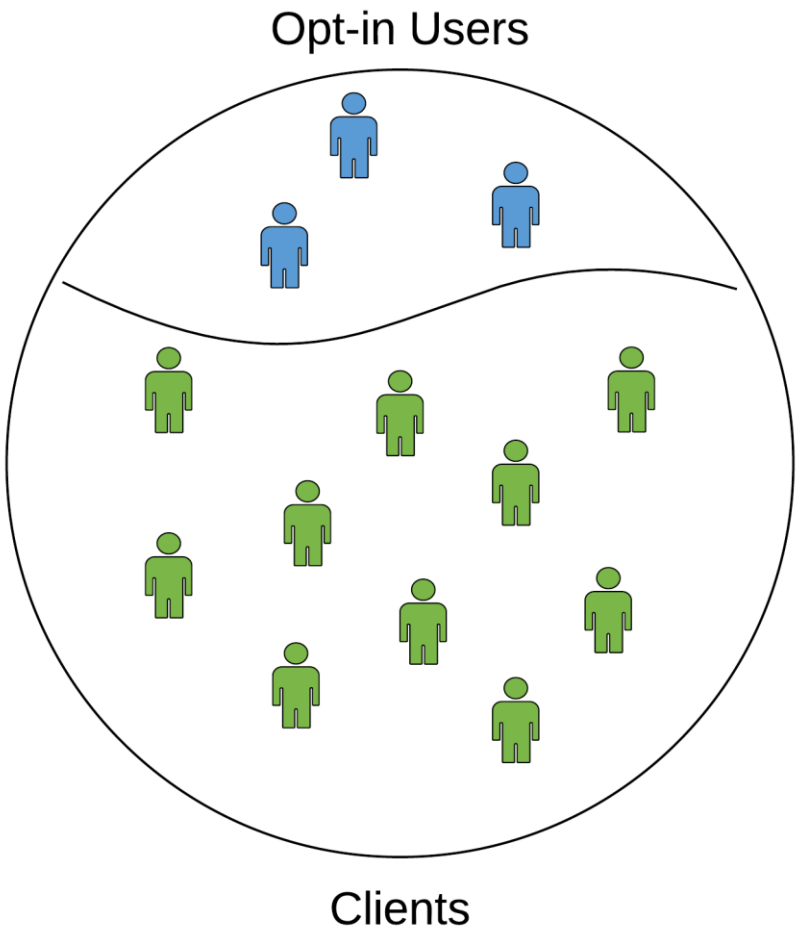
Why a Hybrid Model?



Why a Hybrid Model?



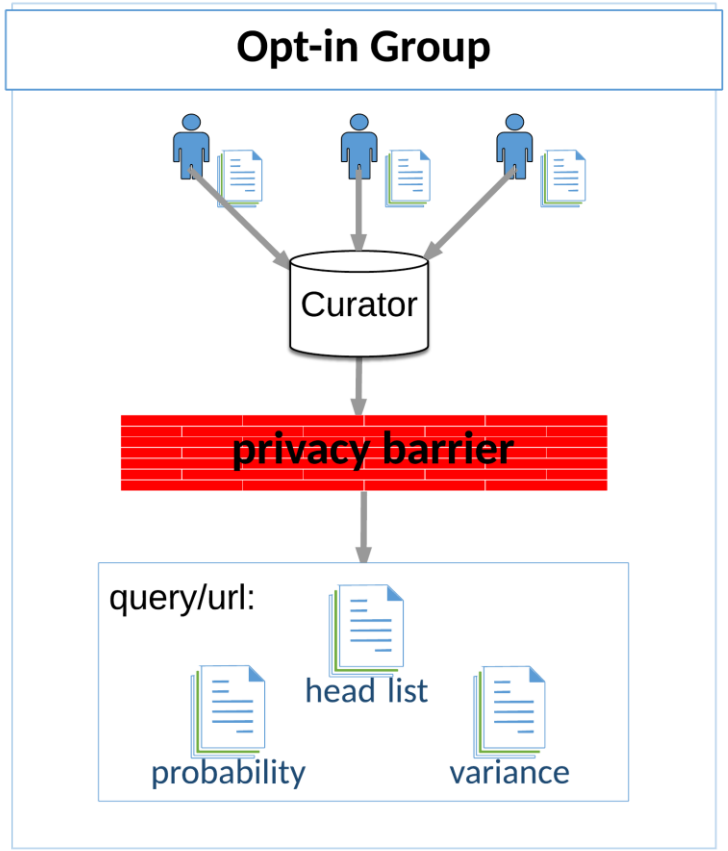
Why a Hybrid Model?



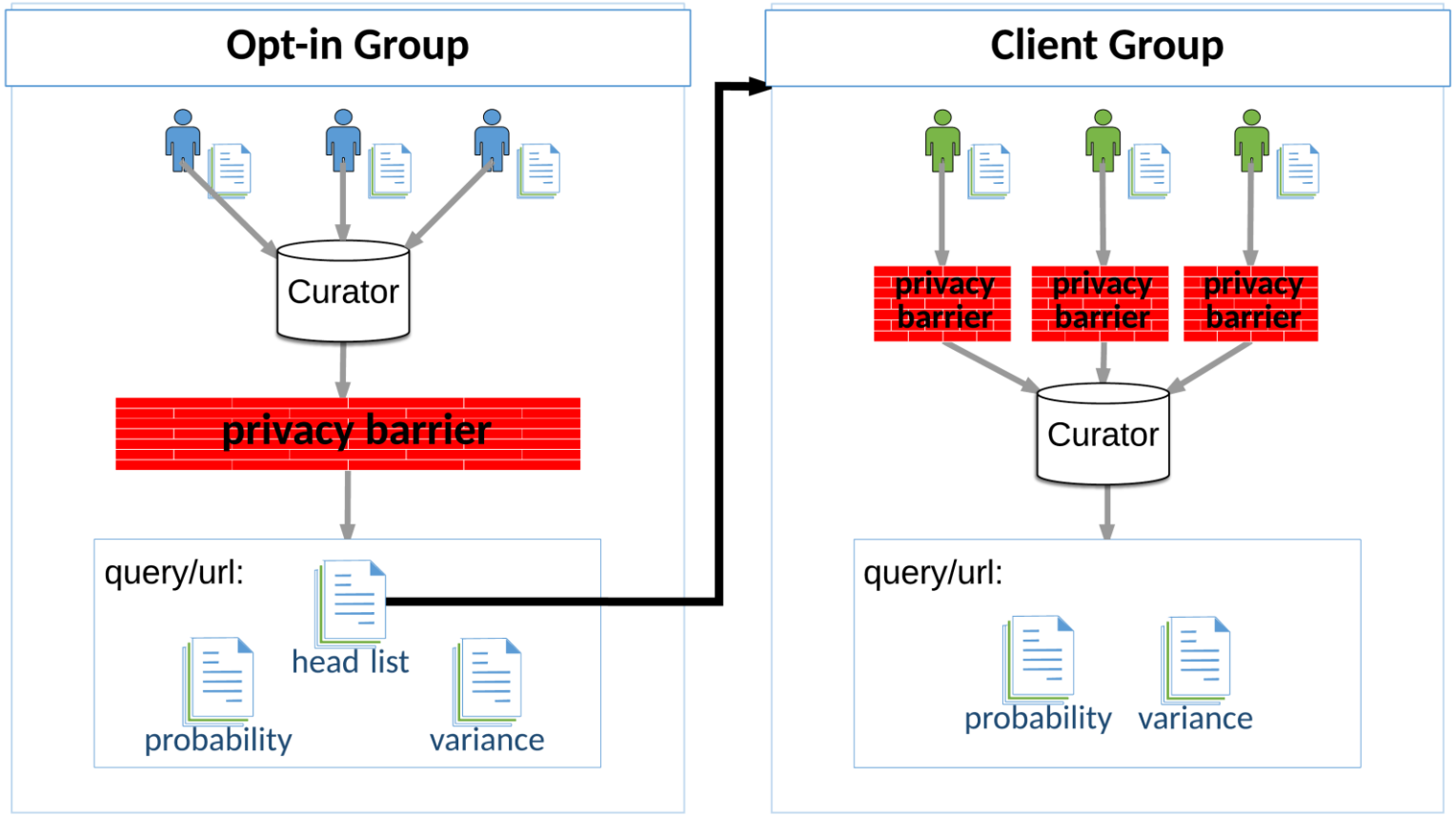
BLENDER

local search in the hybrid model

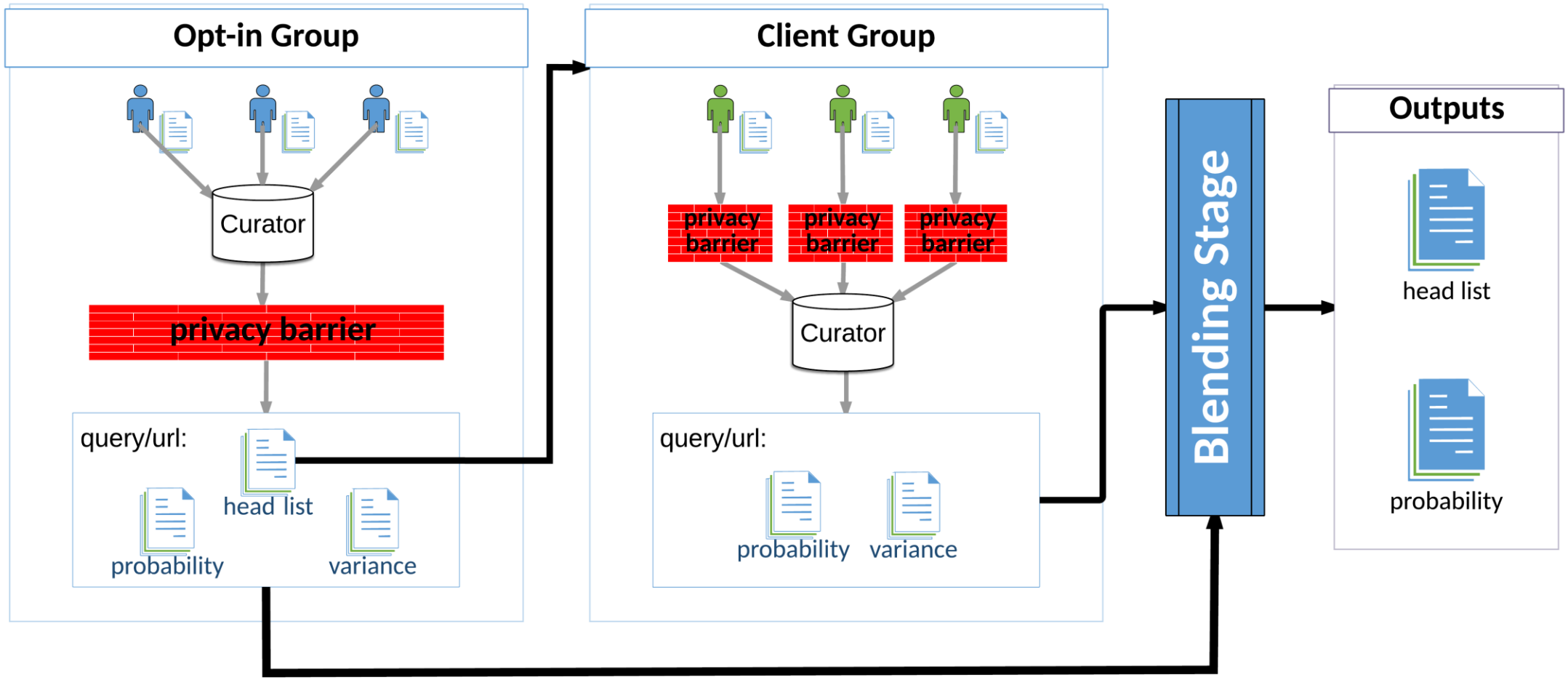
BLENDER Architecture



BLENDER Architecture



BLENDER Architecture



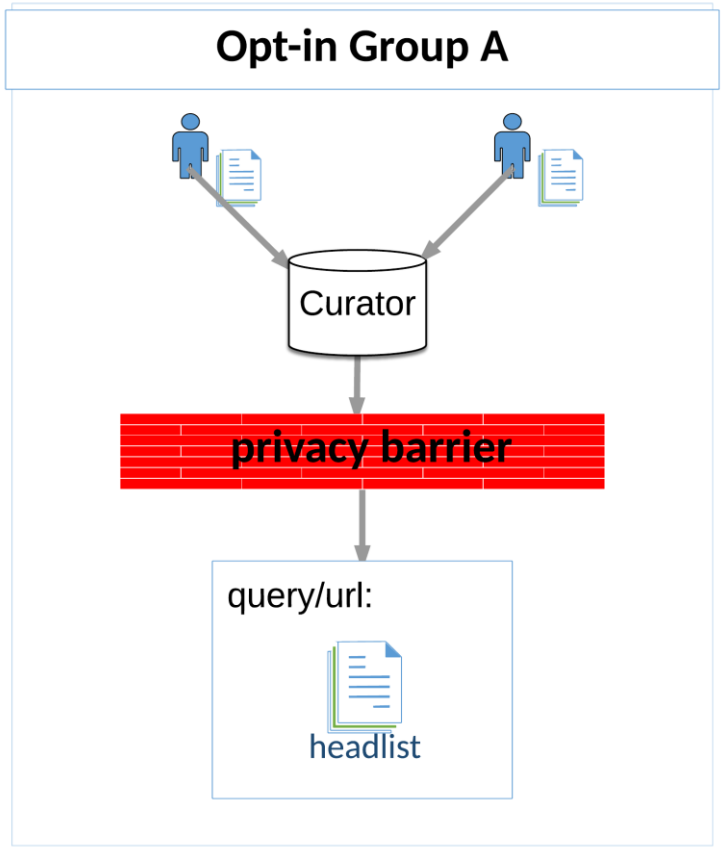
Opt-in Group Algorithm

Two-phase approach: Discovery and Estimation

Partition users into two disjoint groups

- Group A – Discovery phase
- Group B – Estimation phase

Opt-in Group Data: Discovery of Head List

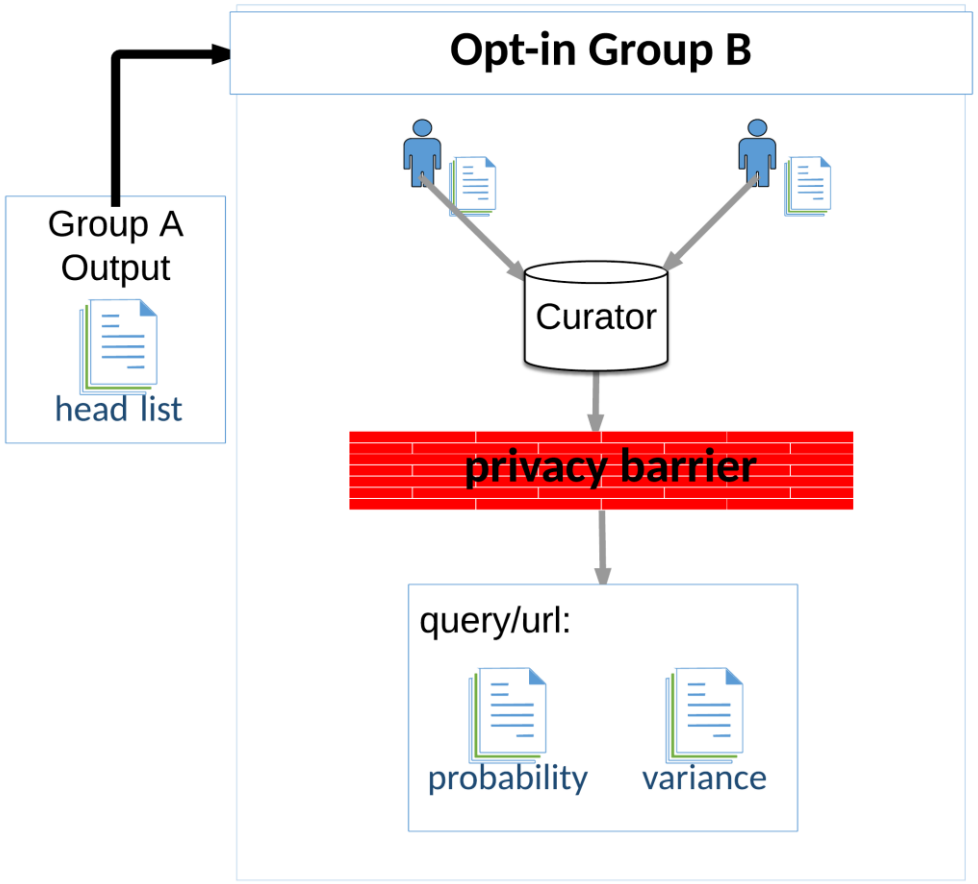


For each distinct <*query, URL*> record from Group A's data:

- Compute empirical probability
- Add Laplace noise to form noisy empirical probability
- If noisy empirical probability exceeds threshold, add record to the ***head list***

[Korolova et al, 2009]

Opt-in Group Data Usage: Estimation

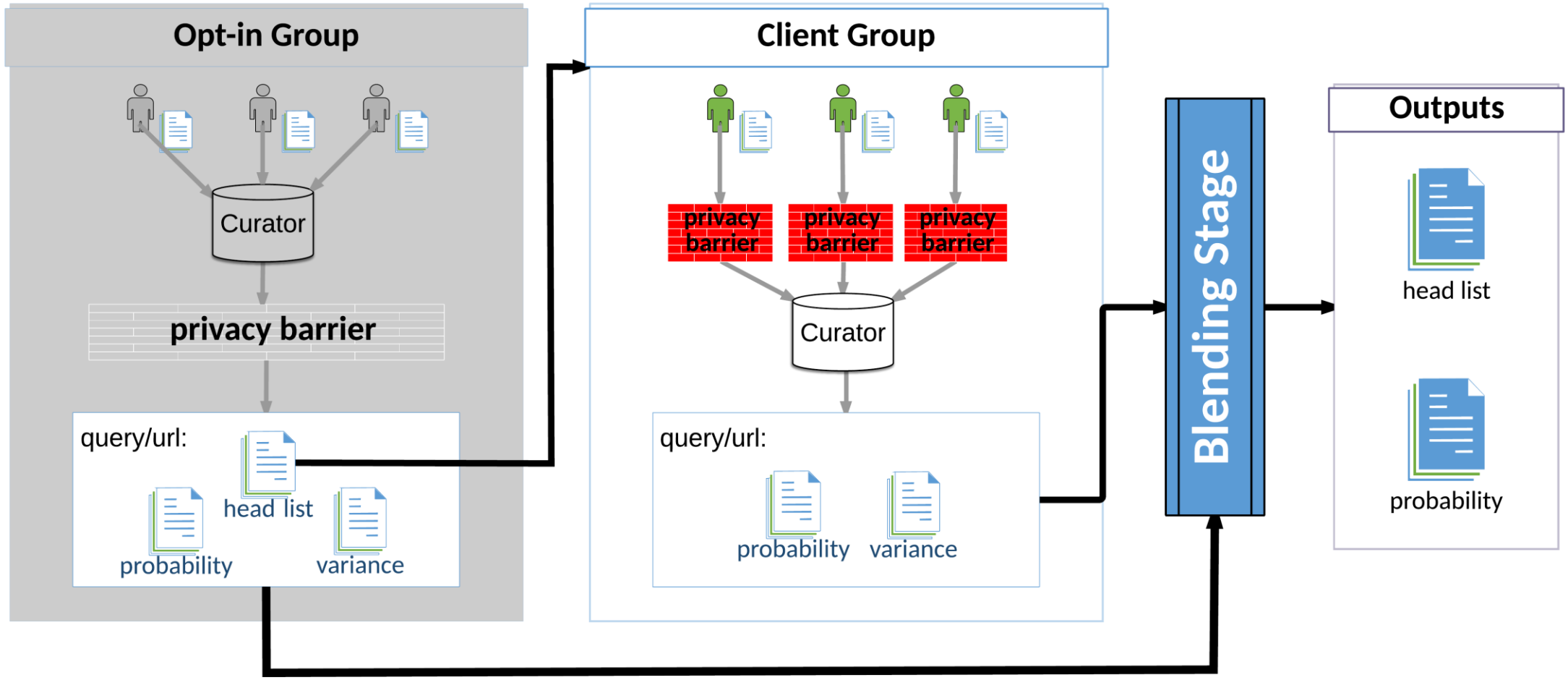


For each distinct <*query, URL*> record from Group B's data and using the privatized head list:

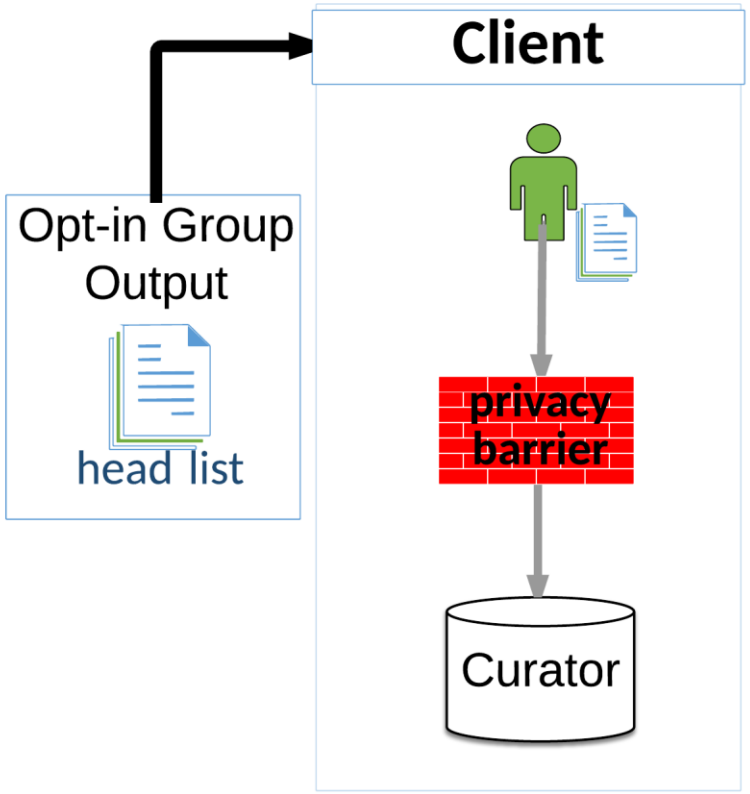
- Compute empirical probability
- Add Laplace noise to form noisy probability estimate
- Compute the sample variance of the probability estimate

[Dwork et al, 2006]

BLENDER: Client Group



Client Data Reporting

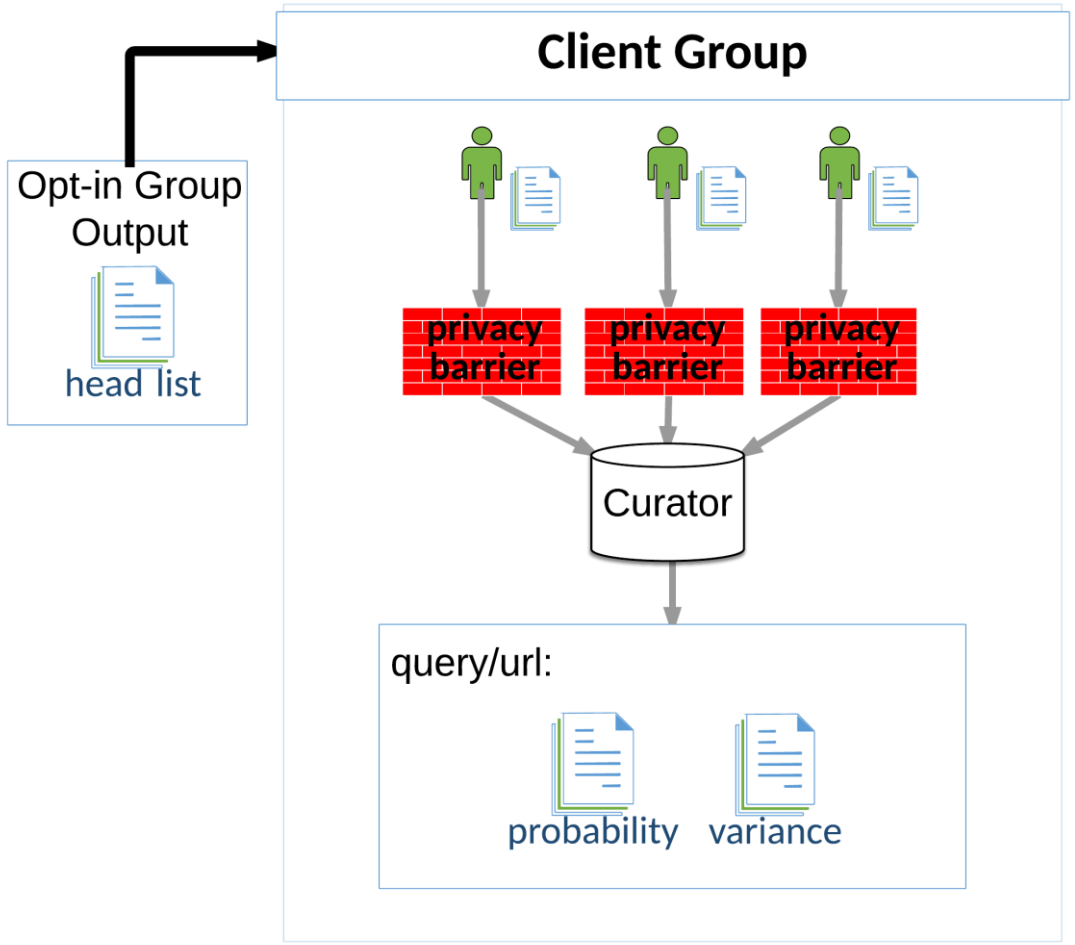


2-stage k-randomized response [Warner 1965]

1. Report the query truthfully with probability t,
   otherwise, report a query at random

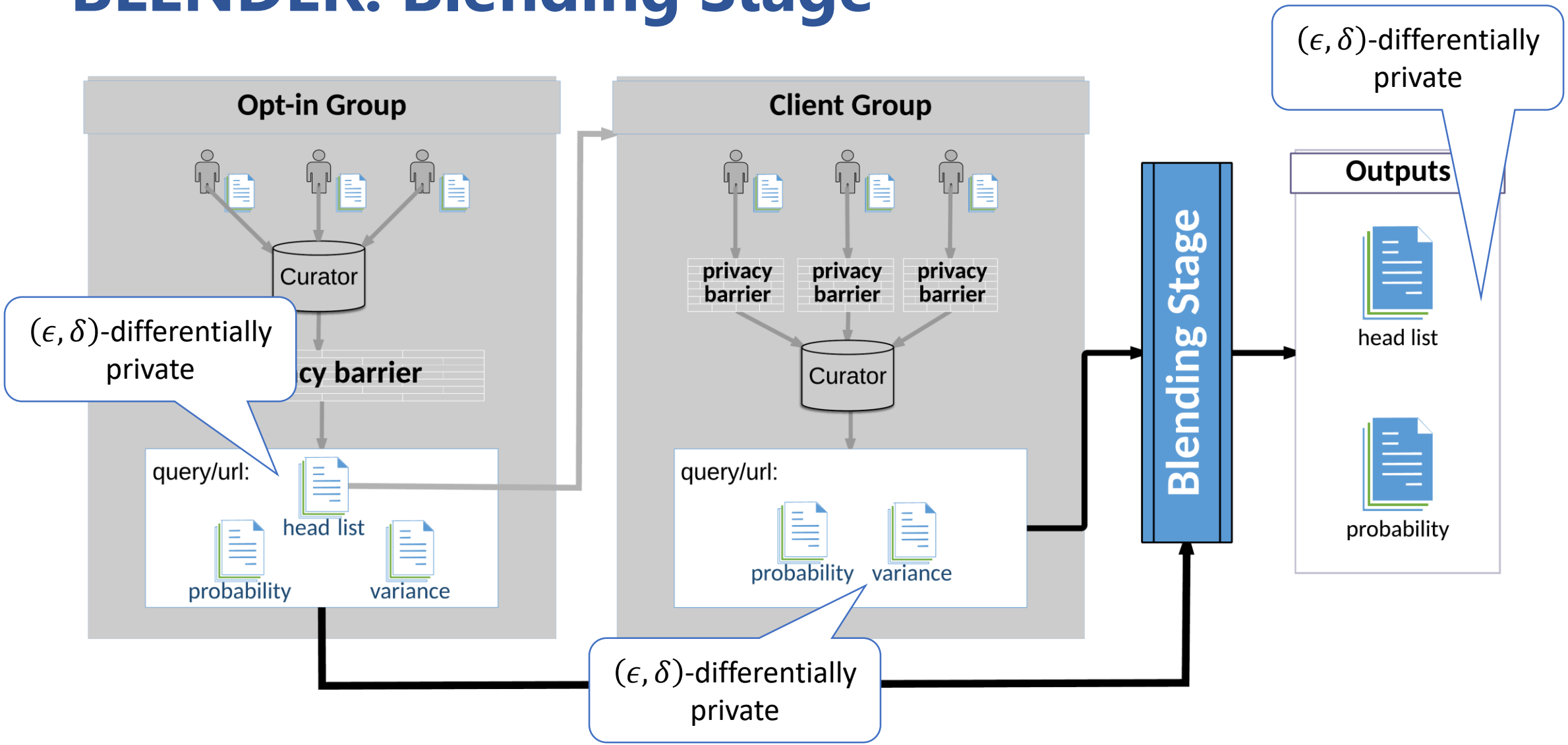
2. Report the URL truthfully with probability t_q,
   otherwise, report a URL at random

Server Aggregating Client Data



- Collects privatized reports from all users
- Aggregates the privatized reports into empirical probability estimates for each record
- Performs denoising procedure to generate unbiased probability estimates and variance estimates

BLENDER: Blending Stage



Evaluation

Measuring the utility of BLENDER

Experimental Datasets

	# Users	# Unique Queries	# Unique URLs	δ
AOL (2006)	0.5M	4.8M	1.6M	10^{-5}
Yandex (2013)	4.9M	13.2M	12.7M	10^{-7}

Measuring Utility

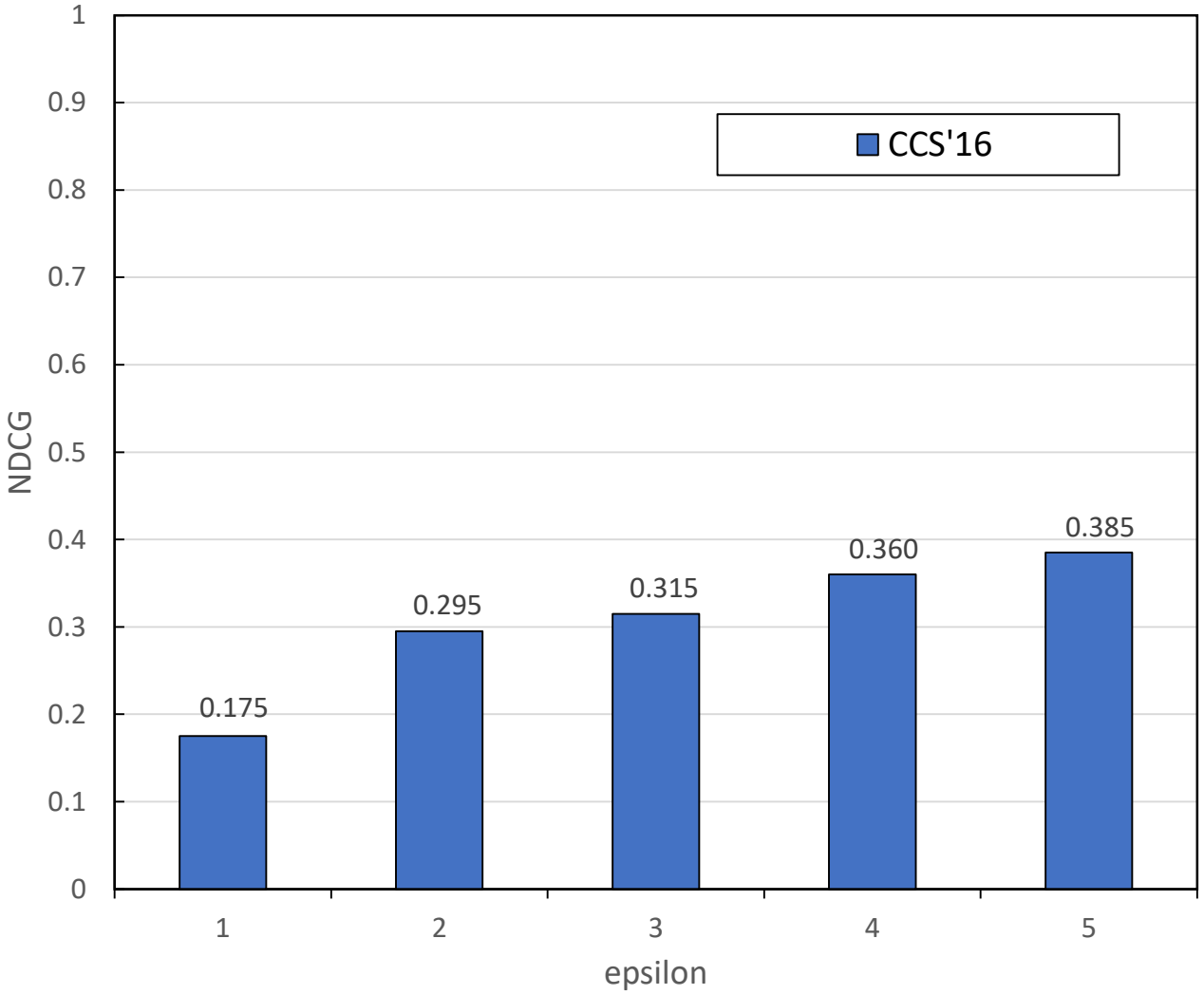
Normalized Discounted Cumulative Gain (NDCG)

- Standard measure of ranking quality
- $DCG = \sum_i \frac{2^{rel_i - 1}}{\log(i+1)}$
- $NDCG = \frac{DCG}{\text{Ideal } DCG}$

NDCG of NDCGs

1. Compute the NDCG for each query's URL list, $NDCG_{q_i}$
2. Generalized DCG for the query list:
$$\sum_i \frac{2^{rel_i - 1}}{\log(i+1)} \cdot NDCG_{q_i}$$
3. Normalize by analogous Ideal DCG

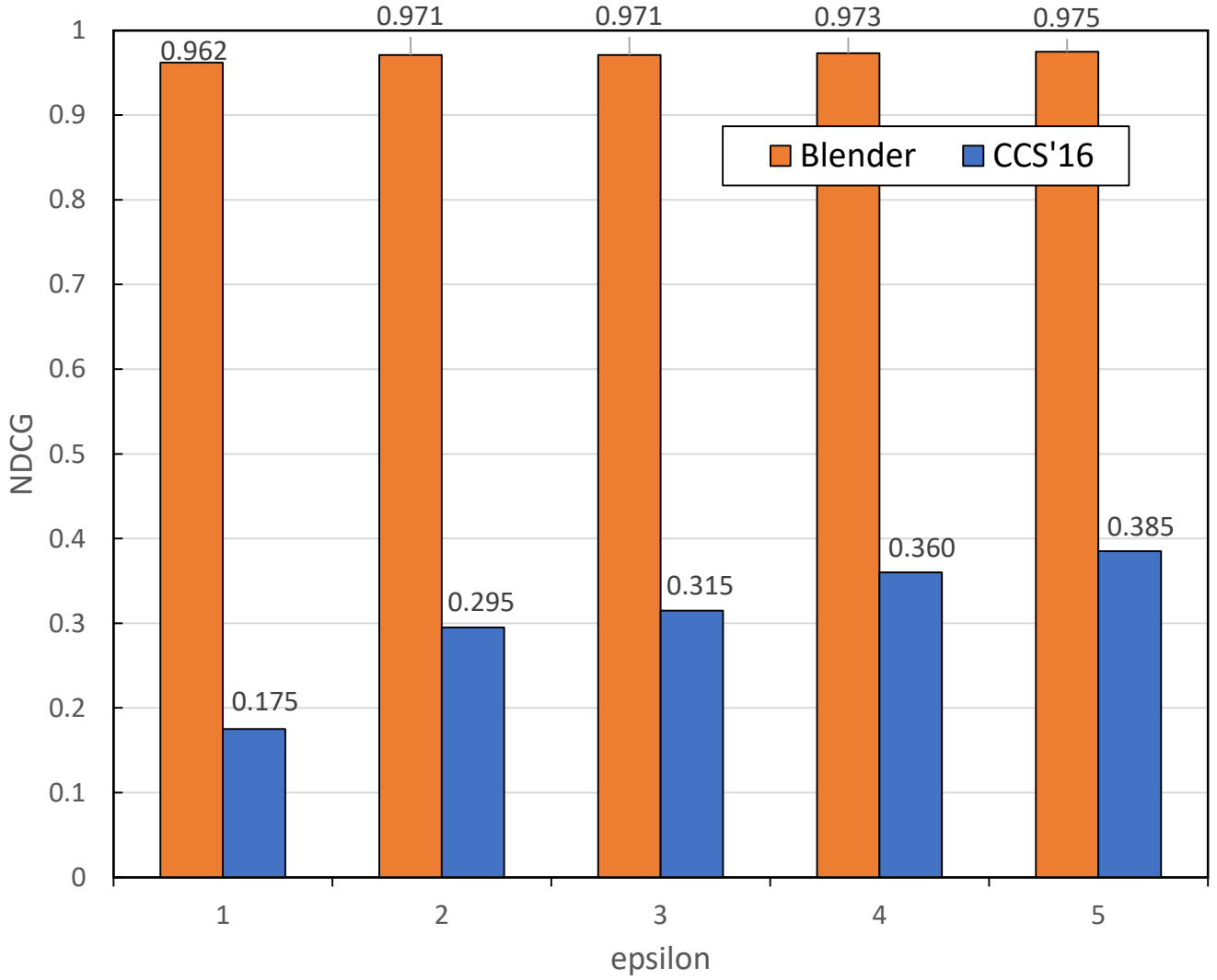
Comparison with Local Model [Qin et al, CCS 2016]



How does BLENDER compare to having all users use the Local Model?

AOL dataset
Head list size: 10

Comparison with Local Model [Qin et al, CCS 2016]



How does BLENDER compare to having all users use the Local Model?

AOL dataset
Head list size: 10

BLENDER
- 5% “opt-in” users
- 95% “client” users

Caveat: Slightly different versions of NDCG. See paper.

Effect of Opt-in User Percentage on NDCG

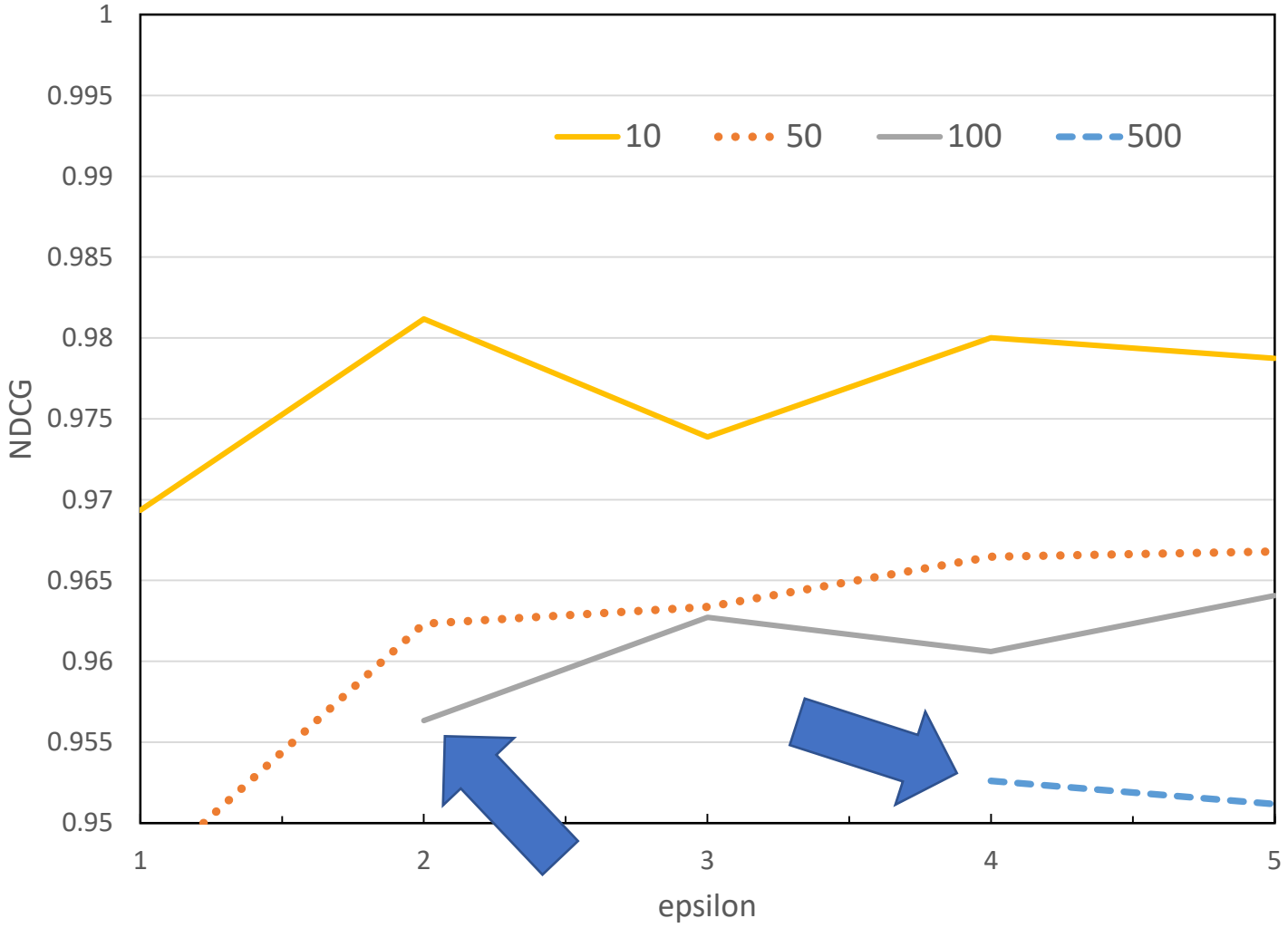
How does BLENDER's utility depend on the size of the opt-in user group?

Yandex dataset

$\epsilon = 4$

Head list sizes: 50, 100, 500

Effect of Privacy Budget on NDCG



How does BLENDER's utility depend on the privacy budget ϵ?

Yandex dataset
2.5% opt-in, 97.5% client
Head list sizes: 10, 50, 100, 500

Conclusions

Conclusions

Proposed a hybrid model for differential privacy

Constructed a blended approach within the hybrid model for local search

Achieved significant improvement on real world datasets with the blended approach

Future Work

- Improve on the sub-components of BLENDER to utilize state-of-the-art privatization methods
- Derive theoretical guarantees for the utility of BLENDER
- Reduce BLENDER's reliance on distributional assumptions
- Develop algorithms in the hybrid model for other applications

BLENDER: Enabling Local Search with a Hybrid Differential Privacy Model

Brendan Avent[1], Aleksandra Korolova[1], David Zeber[2], Torgeir Hovden[2], Benjamin Livshits[3]

[1]*University of Southern California*

[2]*Mozilla*

[3]*Imperial College London*

Thank you

Full paper available here.