Learning from End User Data with Shuffled Differential Privacy over Kernel Densities

Tal Wagner

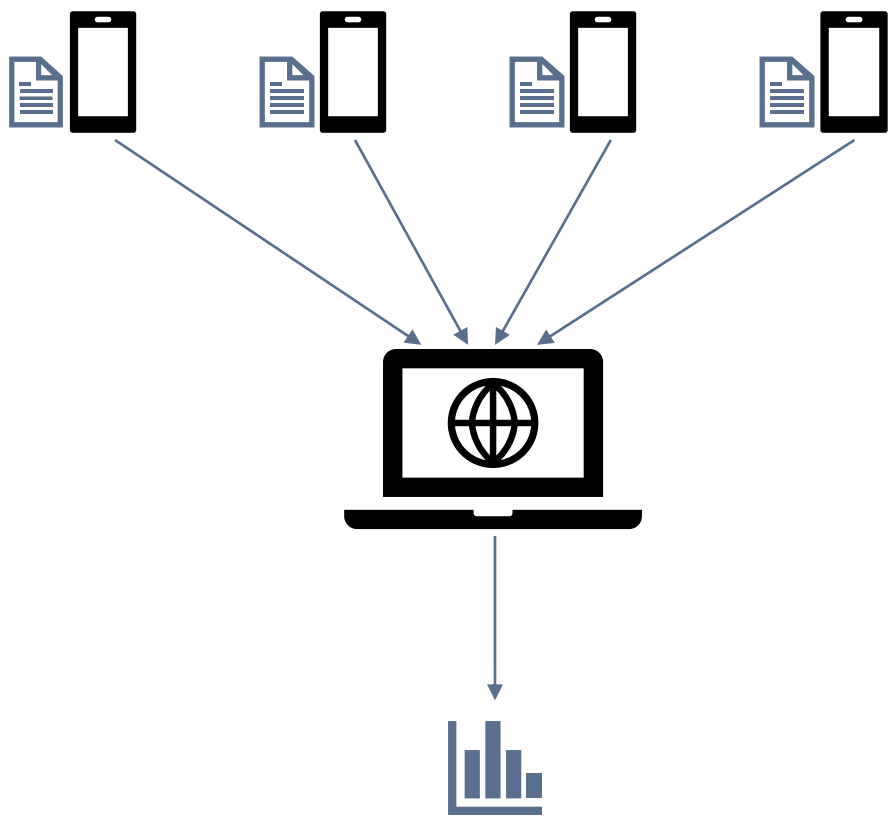
Tel Aviv University

ביה"ס למדעי המחשב
ובינה מלאכותית ע"ש בלווטניק
הפקולטה למדעים מדויקים
ע"ש רימונד ובברלי סאקלר
אוניברסיטת תל אביב

The Blavatnik School of
Computer Science and AI
The Raymond and Beverly Sackler
Faculty of Exact Sciences
Tel Aviv University

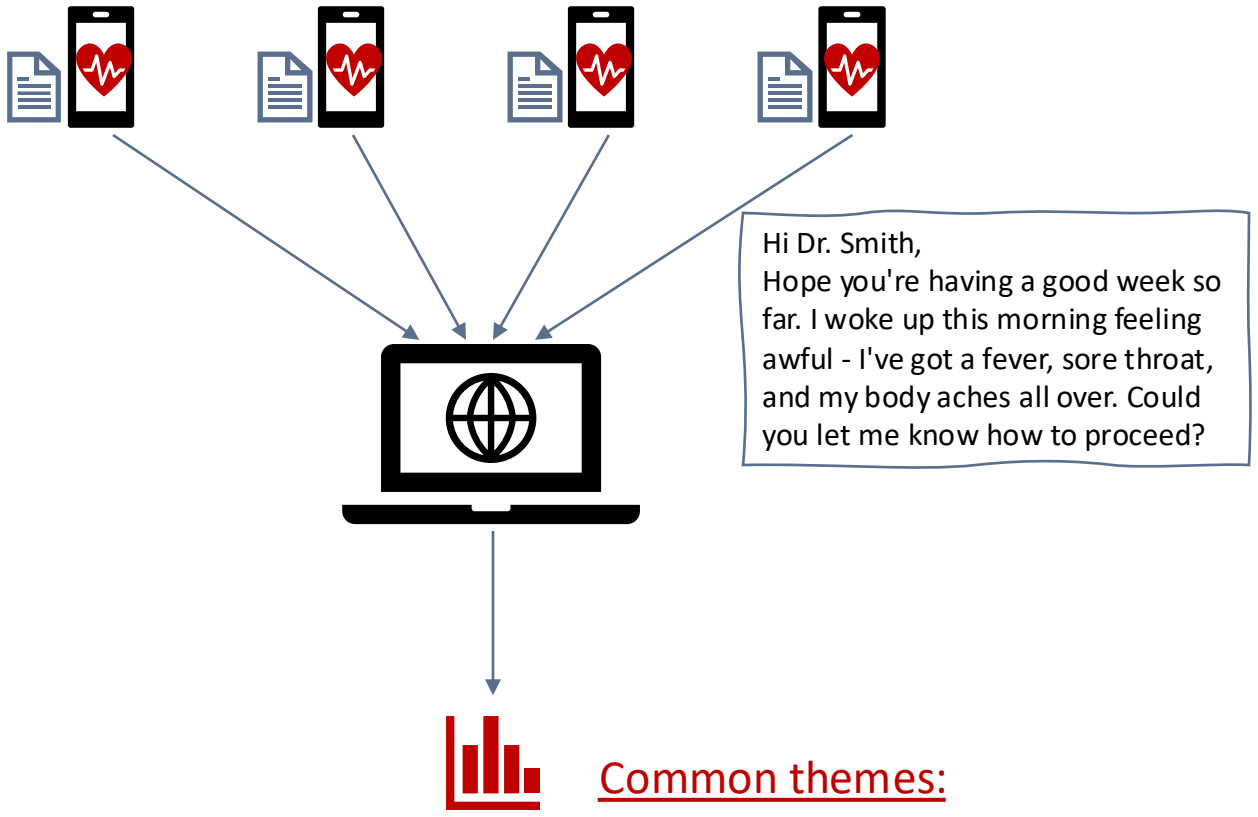
Problem: Private Distributed Topical Decoding

- Network of end users
- Each has a private text
- Users communicate data to a central untrusted server
- **Server goal**: Detect topical themes (what user texts are about) without reading any text

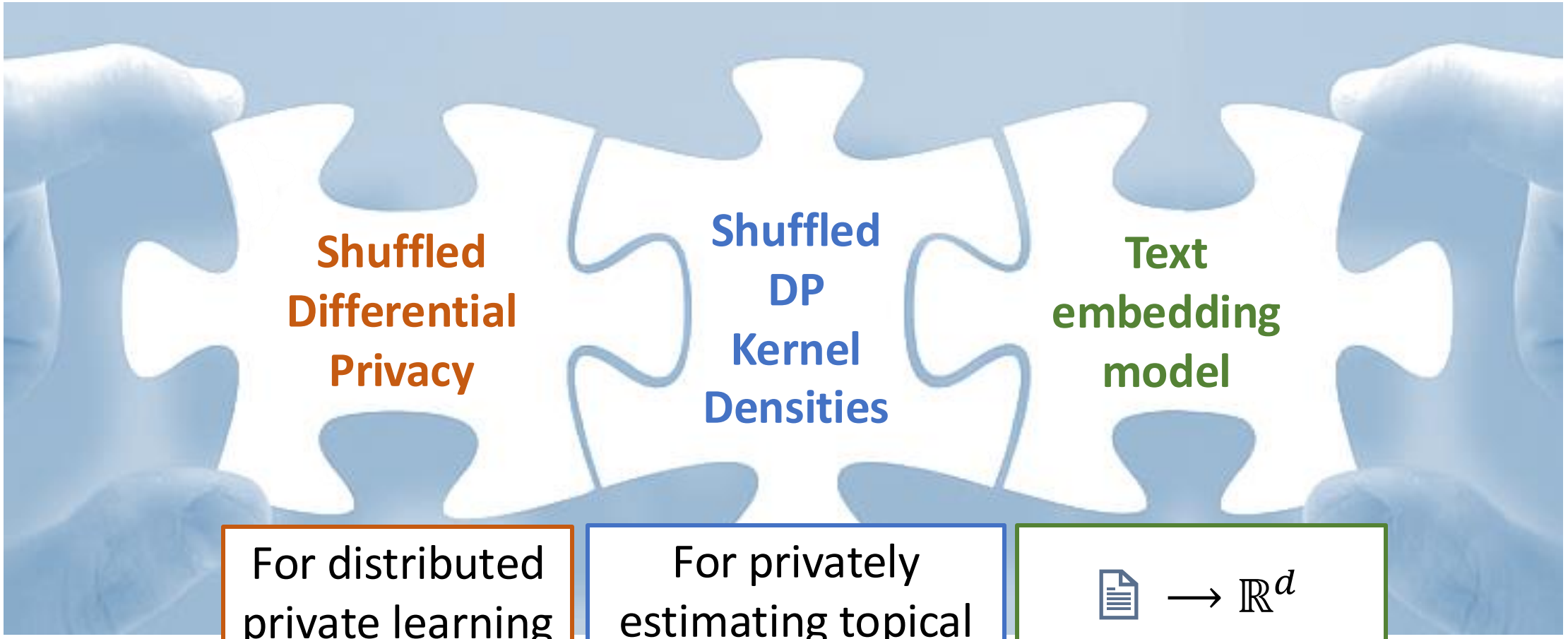


Example: Healthcare Apps

- Network of end users
- Each has a private text
- Users communicate data to a central untrusted server
- **Server goal**: Detect topical themes (what user texts are about) without reading any text



Our Method Components

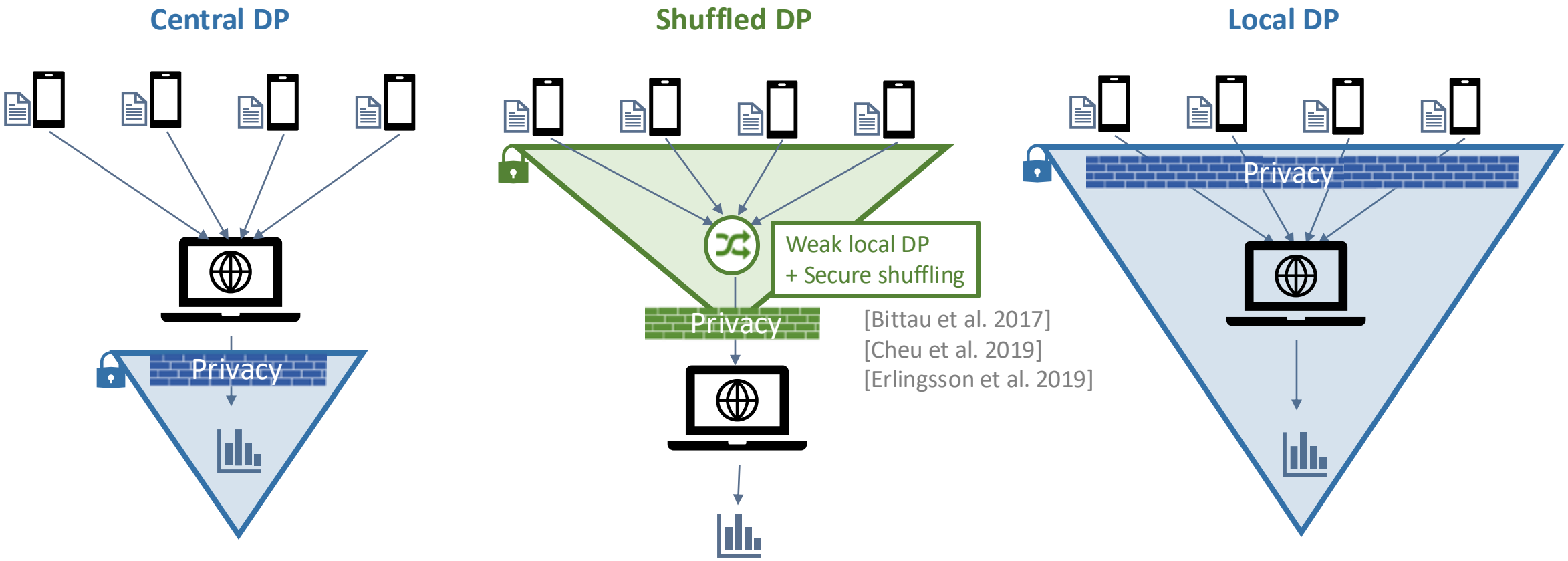


For distributed private learning

For privately estimating topical densities in $\mathbb{R}^d$

$\longrightarrow \mathbb{R}^d$

Distributed Privacy via Shuffled DP



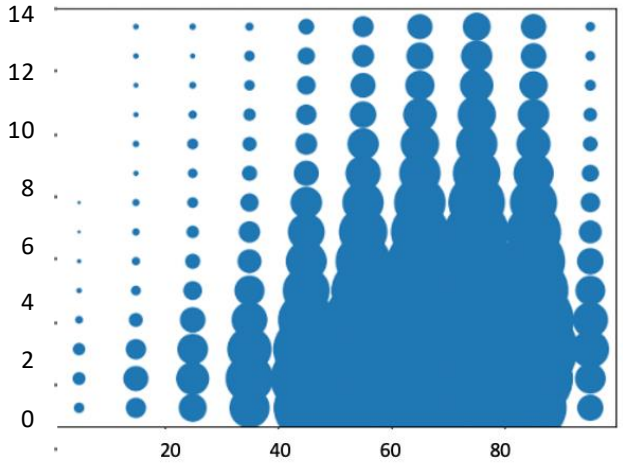
✗ No distributed privacy
✓ Good accuracy

✓ Distributed privacy
✓ Good accuracy

✓ Distributed privacy
✗ Bad accuracy

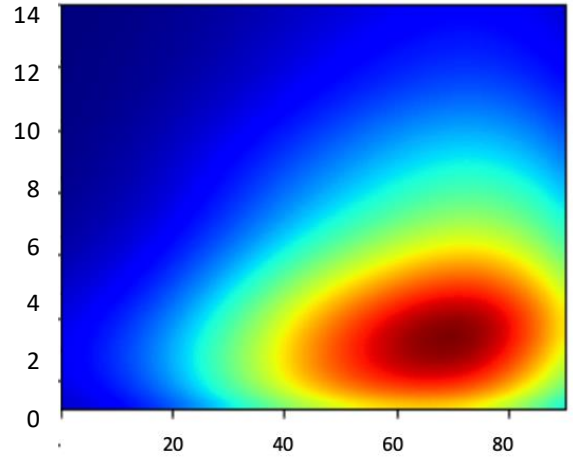
Kernel Density Estimation (KDE)

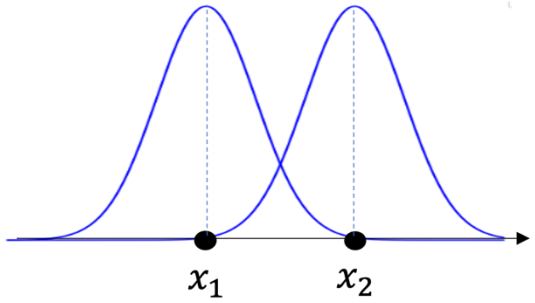
KDE turns a dataset $x_1, \ldots, x_n \in \mathbb{R}^d$ into a continuous density function

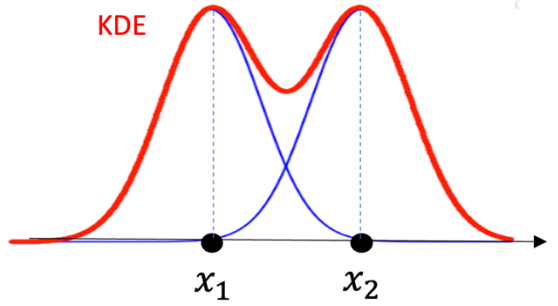


$$\forall\, y \in \mathbb{R}^d\text{:}$$

$$KDE(y) = \frac{1}{n}\sum_{i=1}^{n} e^{-\|y - x_i\|_2^2}$$

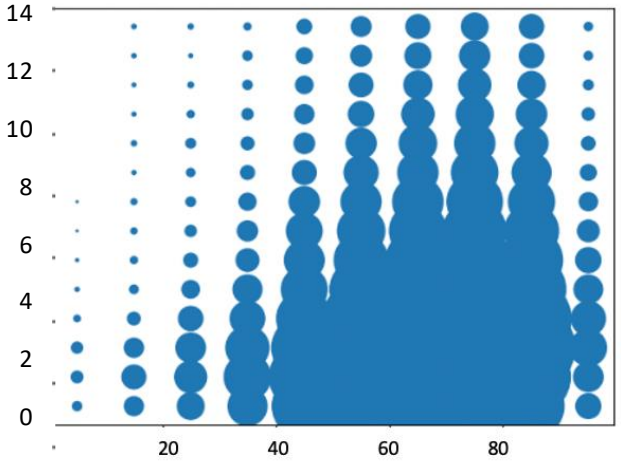




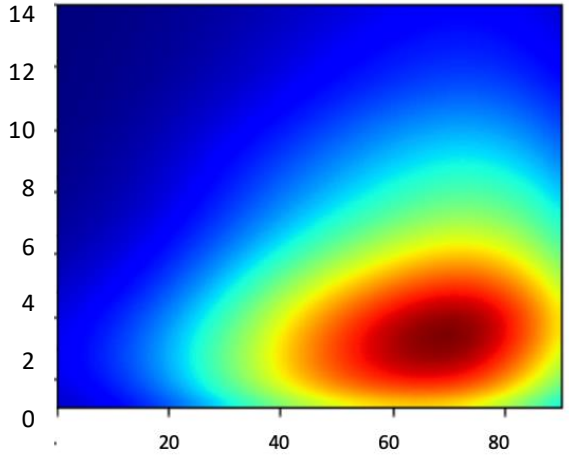


Private Kernel Density Estimation (DP KDE)

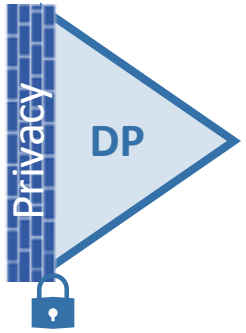
Sensitive user data



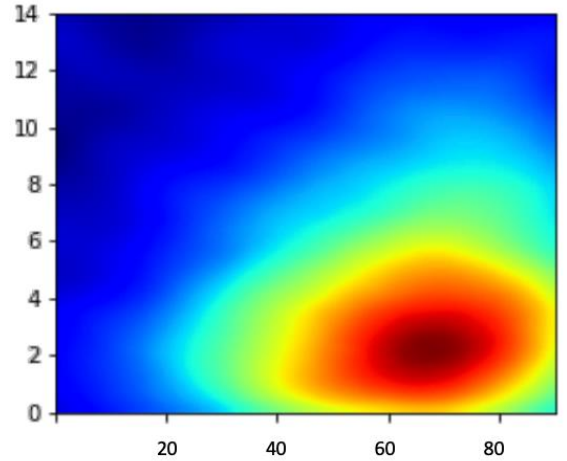
Exact KDE



Leaks private data



Differentially private KDE



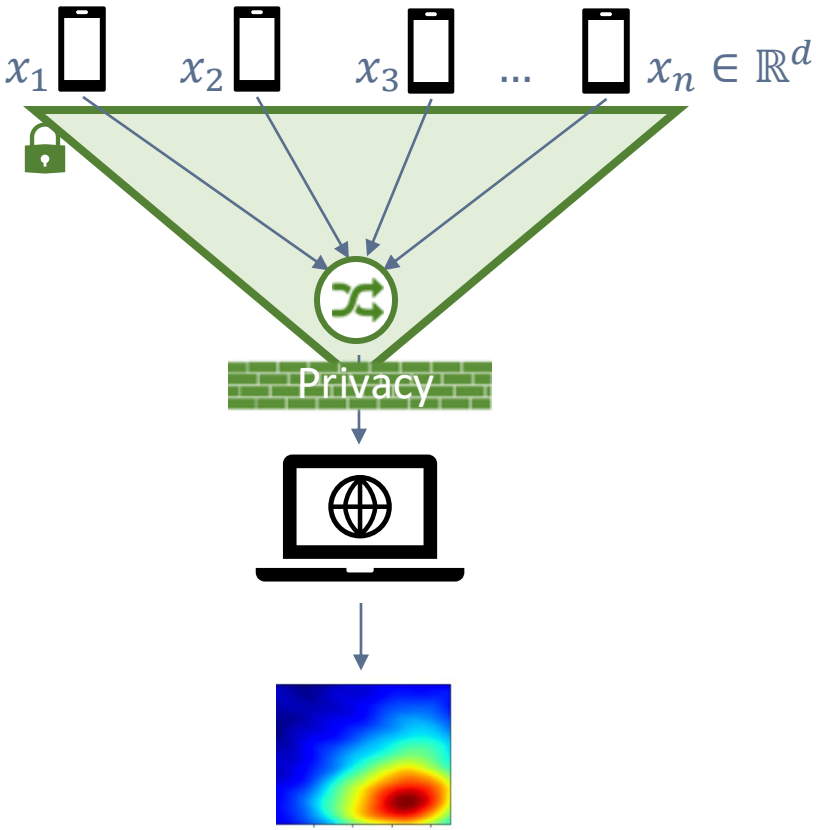
Protects private data

Main Theorem: Shuffled DP KDE

Theorem: Let $\varepsilon, \delta > 0$ such that $\varepsilon \lesssim \log(\delta^{-1})$. For every $\alpha \gtrsim \frac{\sqrt{\log(\delta^{-1})}}{\varepsilon n}$, there is an n-user **shuffled DP** protocol that outputs an approximate Gaussian KDE function $\widetilde{KDE}(\cdot)$ over $\mathbb{R}^d$, such that:

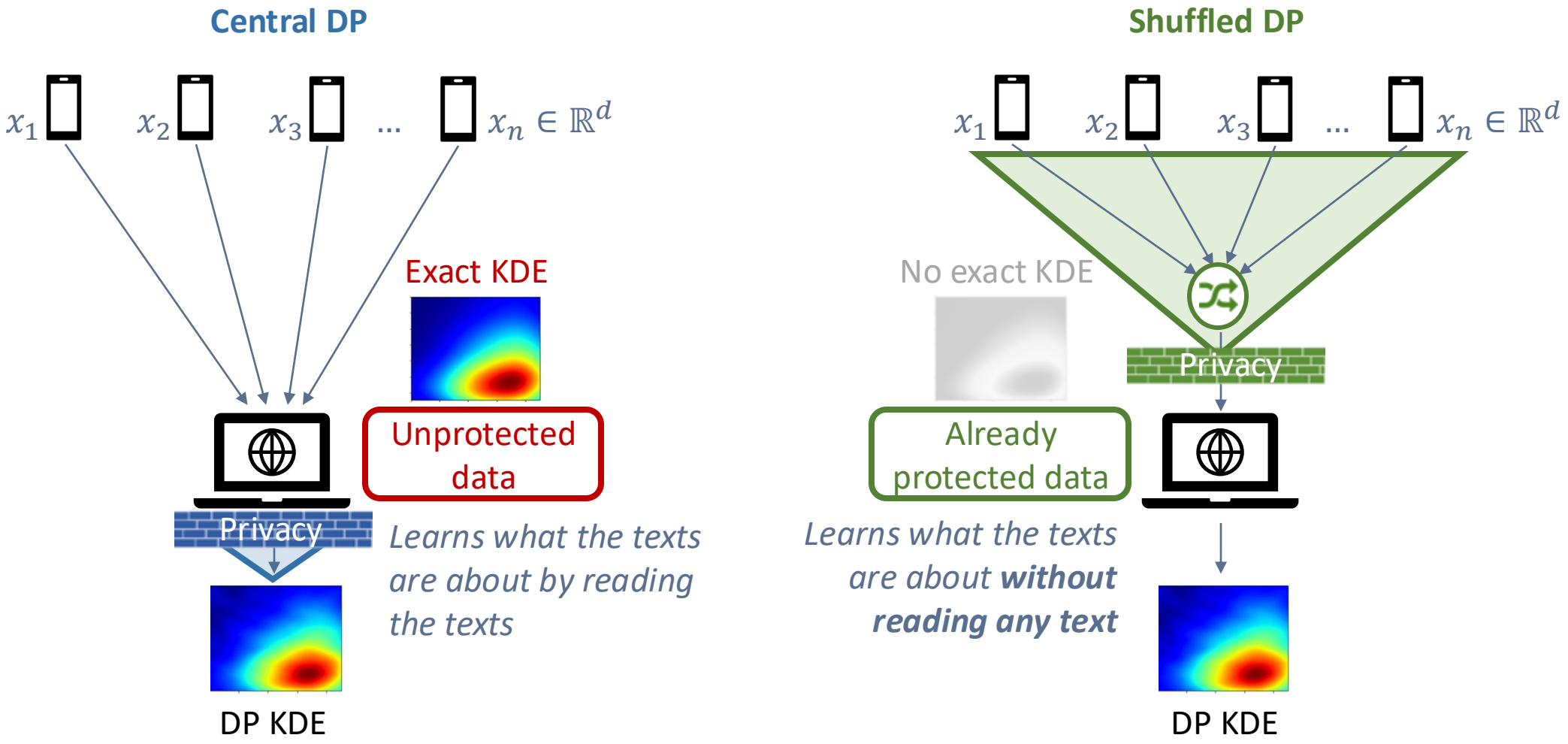
- Privacy: $\widetilde{KDE}(\cdot)$ is (ε, δ)-DP
- Accuracy:

$$\text{SupRMSE} = \sup_{y \in \mathbb{R}^d} \sqrt{\mathbb{E}\left[KDE(y) - \widetilde{KDE}(y)\right]^2} \leq \alpha$$

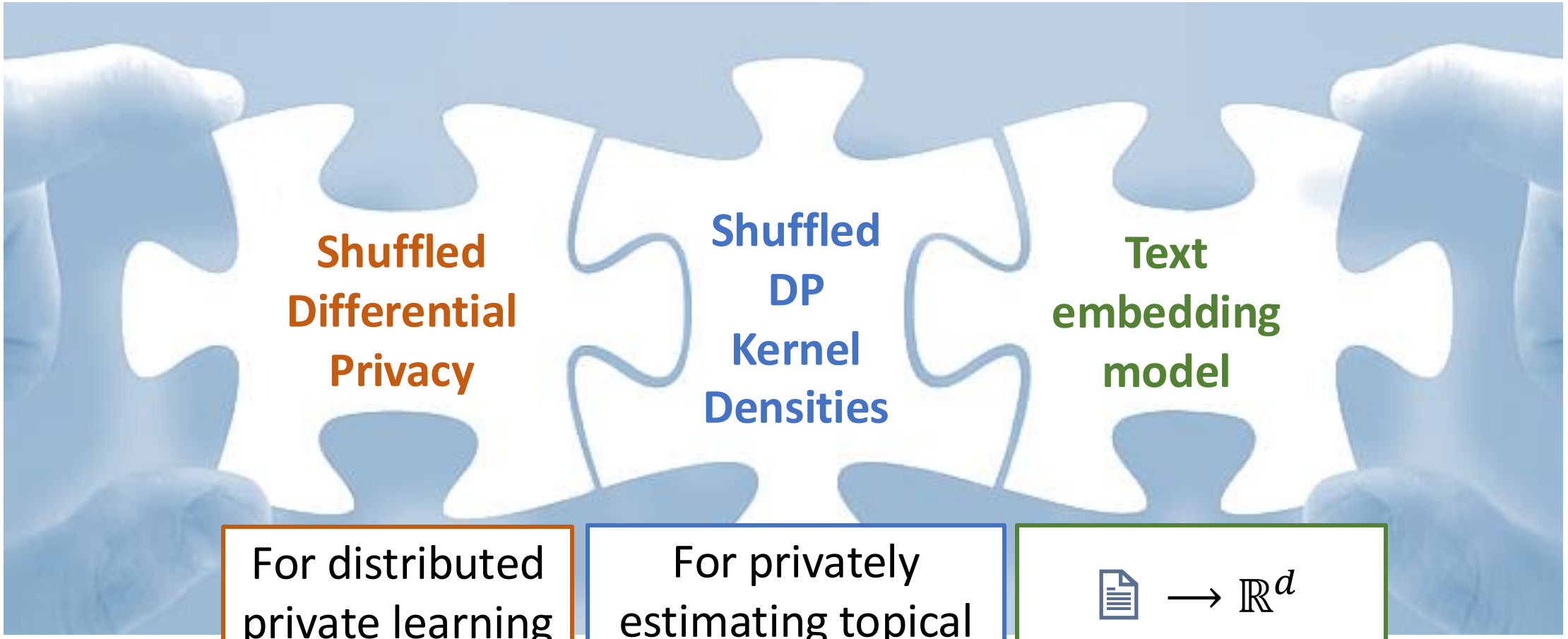


- Efficiency: User time $O\left(\frac{d}{\alpha^2}\right)$, communication $O\left(\frac{\log(\alpha^{-1})}{\alpha^2}\right)$ bits/user, server time $O\left(\frac{n}{\alpha^2}\right)$

Central vs. Shuffled DP KDE



Putting Things Together



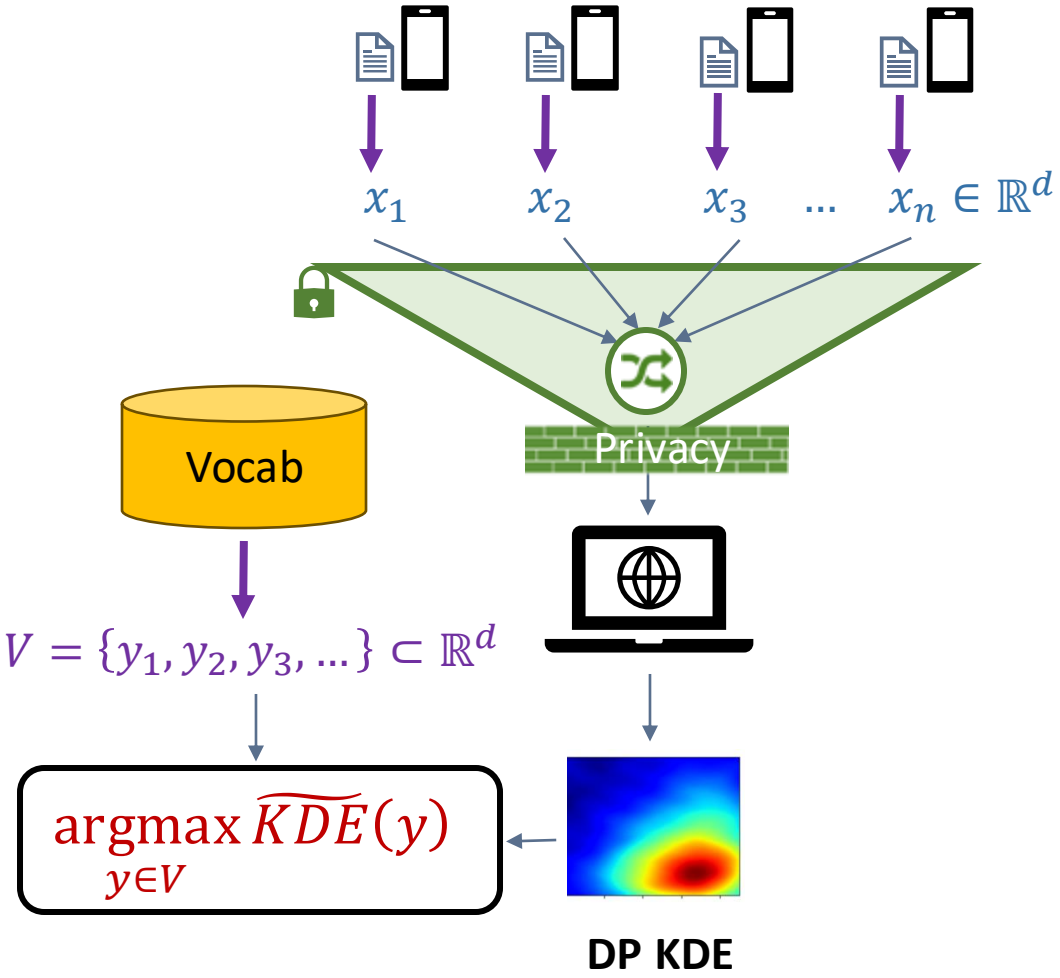
For distributed private learning

For privately estimating topical densities in $\mathbb{R}^d$

$\longrightarrow \mathbb{R}^d$

Private Distributed Topical Decoding

- Locally embed user texts in $\mathbb{R}^d$ with a public embedding model (e.g., SentenceBert)
- Compute shuffled DP KDE function $\widetilde{KDE}(\cdot)$
- Take a public vocabulary (e.g., GloVe 400k)
- Embed vocabulary in $\mathbb{R}^d$ with the same public embedding model
- Output the vocabulary terms with the highest private density estimate $\widetilde{KDE}(\cdot)$



Some Experiments

Paper:

- Privacy parameters: $\varepsilon \approx 3.2, \delta = 10^{-6}$

- **Experiment 1:**
  - Most user texts are about **artists** (DBPedia-14 dataset)
  - 8% of users have off-topic texts
  - Top-3 DP KDE terms: ***artist, lyricists, musician***

- **Experiment 2:**
  - Most user texts are **sports news articles** (AG news dataset)
  - 20% of users have off-topic news articles
  - Top-3 DP KDE terms: ***injury, semifinalists, finalists***

Thank You