ICML | 2020

Thirty-seventh International Conference on Machine Learning

Fast and Private Submodular and k-Submodular Functions Maximization with Matroid Constraints

Akbar Rafiey

Yuichi Yoshida

SFU SIMON FRASER UNIVERSITY

NII Inter-University Research Institute Corporation / Research Organization of Information and Systems National Institute of Informatics

Core massage

- What is the problem?
- What do we want to achieve?
- What do we achieve in this paper?

What is the problem?

Analyst: wants to do statistical analysis of data

Examples:
- medical data ,
- web search data,
- social networks,
- Salary data
- Etc,

What do we want to achieve?

We need an algorithm such that:

- It returns almost a correct answer to a query
- It is efficient and fast
- Preserves privacy when we have sensitive data.

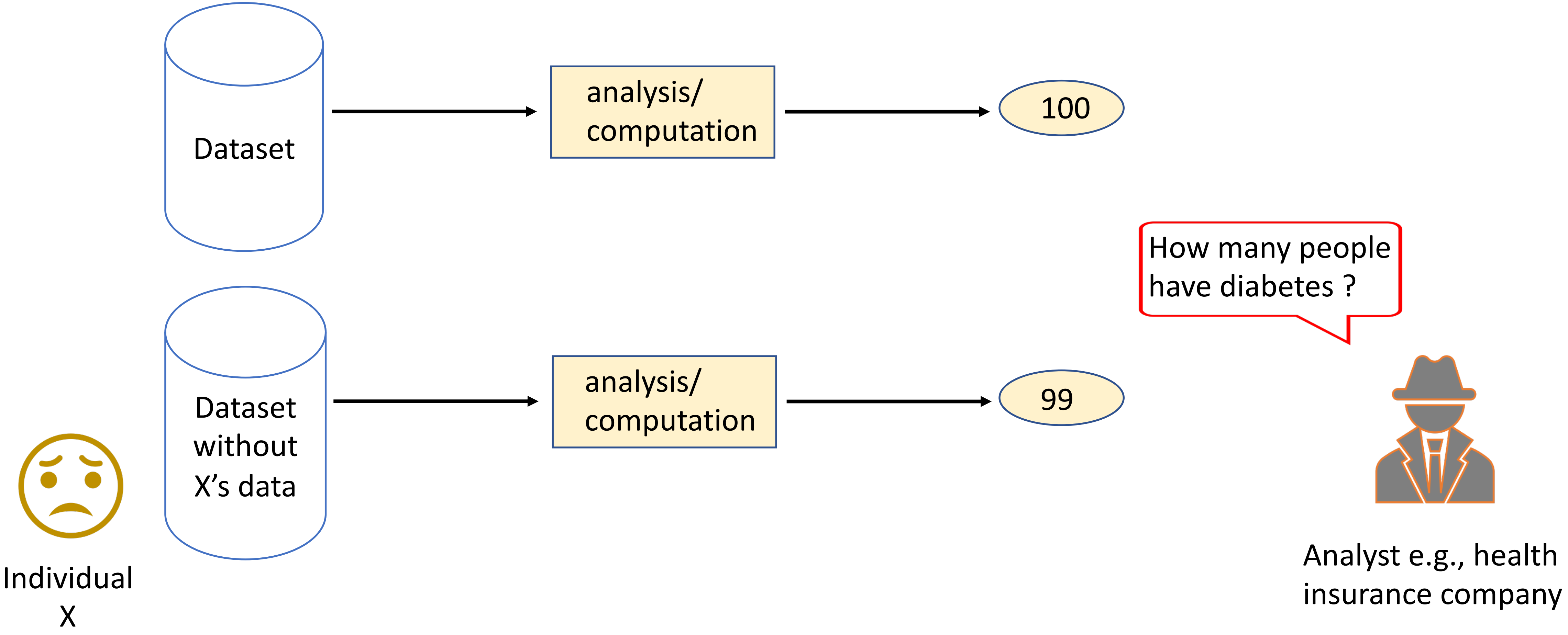
What we achieve in this paper?(part 1)

- We consider a class of set function queries, namely submodular set functions
- We present an algorithm for submodular maximization and prove:
  - It is computationally efficient,
  - Outputs solutions close to an optimal solution
  - Preserves privacy of dataset

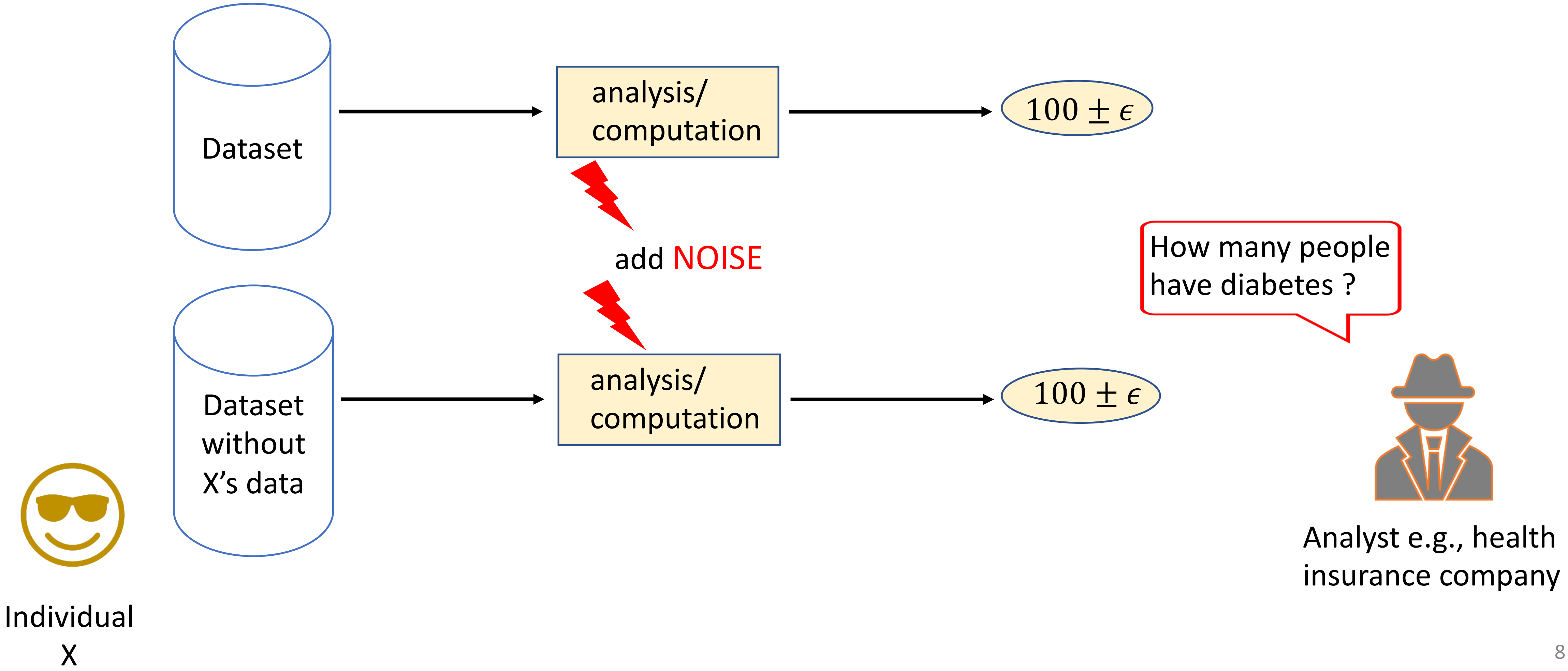
What we achieve in this paper?(part 2)

- Further, we consider a generalization of submodular functions, namely k-submodular functions.
- This allows to capture more problems.
- We present an algorithm for k-submodular maximization and prove:
  - It is computationally efficient,
  - Outputs solutions close to an optimal solution
  - Preserves privacy of dataset

Differential privacy:
A rigorous notion of privacy

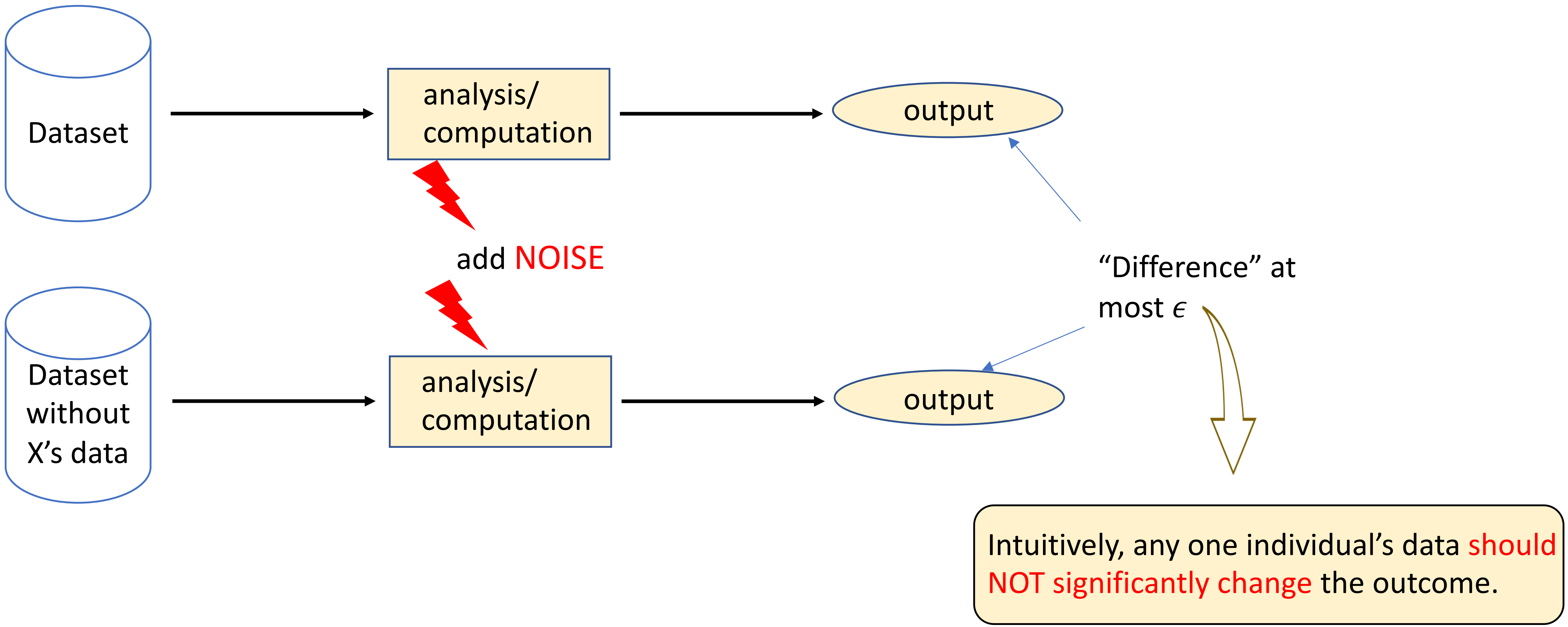


Differential privacy:
A rigorous notion of privacy



Differential privacy:

A rigorous notion of privacy



Differential Privacy (definition)

- For $\epsilon, \delta \in R_+$, we say that a *randomized* computation M is (ϵ, δ)-*differentially private* if
  1. for any neighboring datasets $D \sim D'$, and
  2. for any set of outcomes $S \subseteq$ range(M),

$$\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$$

Neighboring datasets: two datasets that differ in at most one record.

Set function queries

m features

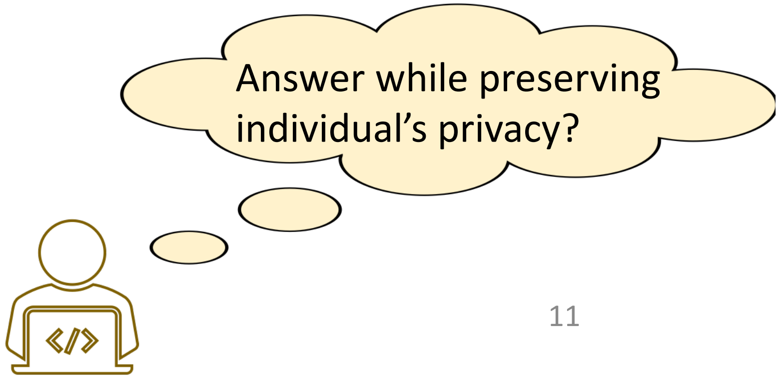
Id	gender	diabetes	….	asthma	Class
1	F	0	….	1	C1
2	M	1	….	1	C1
3	F	0	….	1	C1
4	M	1	….	0	C1
5	F	0	….	0	C1
6	NA	1	….	0	C1
7	F	0	….	1	C2
8	M	1	….	1	C2
9	NA	0	…..	1	C2
10	M	1	….	1	C2

Dataset D

Set function $f_D: 2^E \rightarrow R$

- Given dataset D, function $f_D(S)$ measures "values" of set S in dataset D
- $f_D(\{gender, diabetes\}) = 5$
- $f_D(\{asthma\}) = 7$

Query: what are k most informative features ?



Submodular Function

- In words: the marginal contribution of any element e to the value of the function $f(S)$ diminishes as the input set S increases.

- Mathematically, a function $f: 2^E \rightarrow R$ is submodular if
  - for all $A \subseteq B \subseteq E$,
  - and all elements $e \in E \setminus B$ we have

$$f(\mathrm{A} \cup \{e\}) - f(A) \geq f(B \cup \{e\}) - f(B)$$

diminishing gain property

Problem

- Design a framework for differentially private submodular maximization under matroid constraint.

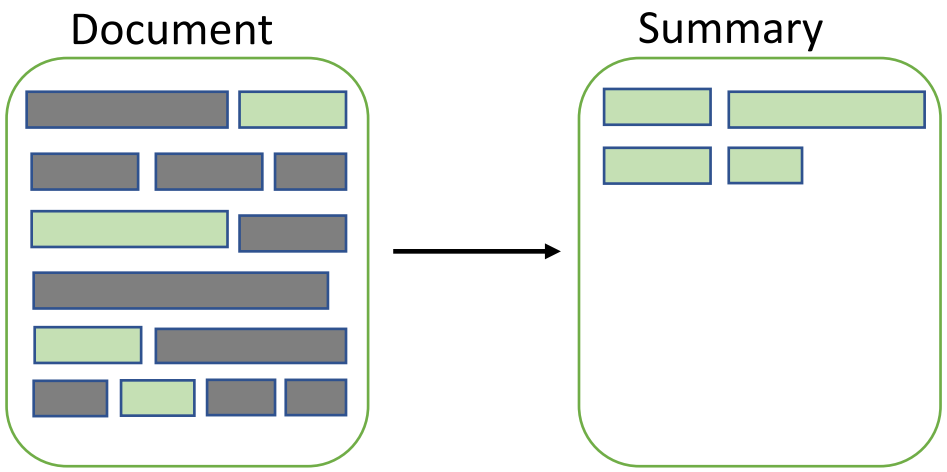
- A pair $M = (E, I)$ of a set E and $I \subseteq 2^E$ is called a *matroid* if
  - $\emptyset \in I$,
  - $A \in I$ for any $A \subseteq B \in I$,
  - for any $A, B \in I$ with $|A| < |B|$, there exists $e \in B \setminus A$ such that $A \cup \{e\} \in I$.

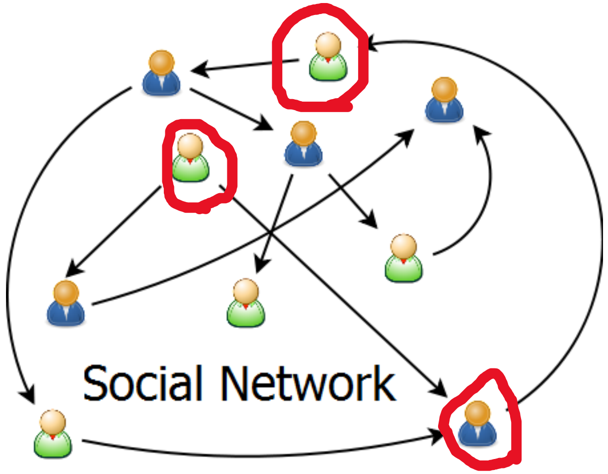
- Our objective: $\operatorname*{argmax}_{S \in I} f(S)$

Examples of submodularity

- Feature selection
- Influence maximization
- Facility location
- Maximum coverage
- Data summarization
  - Image summarization
  - Document summarization

….

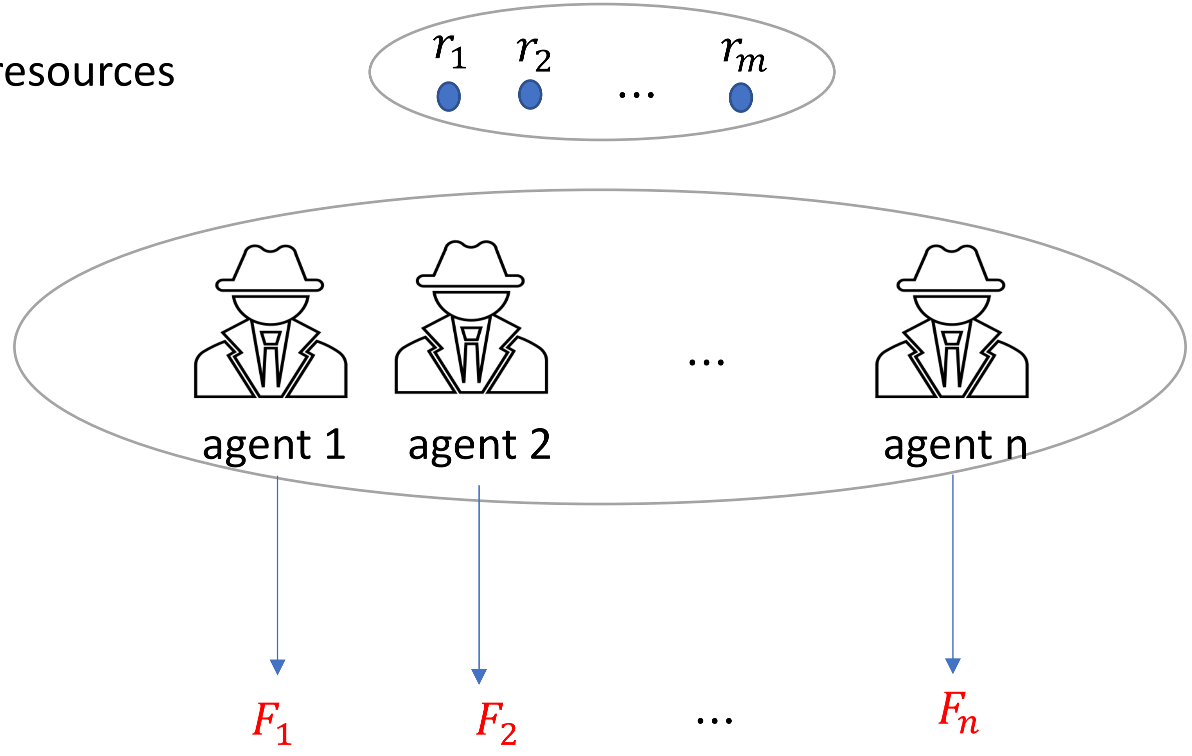




A toy example

Set E : m resources

n agents



Each agent has a private submodular function $F_i : 2^E \to R$

Objective: find $S \subseteq E$ in the matroid that maximizes

$$\sum_{i=1}^{n} F_i(S)$$

Our contributions

	non-private	previous result (Mitrovic et al.,)	our result
utility	$\left(1-\frac{1}{e}\right) OPT$	$\frac{1}{2} OPT - O(\frac{\Delta \cdot r(M) \cdot \ln(\lvert E \rvert)}{\epsilon})$	$\left(1-\frac{1}{e}\right) OPT - O(\sqrt{\epsilon} + \frac{\Delta \cdot r(M) \cdot \ln(\lvert E \rvert)}{\epsilon^3})$
privacy	--	$\epsilon . r(M)$	$\epsilon . r(M)^2$

- $\left(1-\frac{1}{e}\right) OPT$ is the best possible approximation ratio unless P=NP.
- Our algorithm uses almost cubic number of function evaluations $O(r(M) \cdot \lvert E \rvert^2 \cdot \ln(\frac{r(M)}{\epsilon}))$.
- Our privacy factor is worse than the previous work since we deal with multilinear extension.
- Please see our paper for details and proofs

Generalization of submodularity:

K-submodular functions

A function $f: (k+1)^E \to R_+$ defined on k-tuples of pairwise disjoint subsets of E is called *k-submodular* if for all k-tuples $S = (S_1, \dots, S_k)$ and $T = (T_1, \dots, T_k)$ of pairwise disjoint subsets of E,

$$f(S) + f(T) \geq f(S \sqcap T) + f(S \sqcup T)$$

where we define

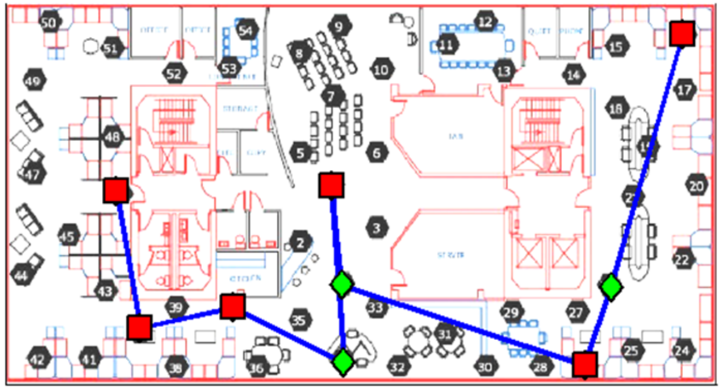
$$S \sqcap T = ((S_1 \cap T_1), \dots, (S_k \cap T_k))$$

$$S \sqcup T = ((S_1 \cup T_1) \setminus \left(\bigcup_{i \neq 1} S_i \cup T_i\right), \dots, (S_k \cup T_k) \setminus \left(\bigcup_{i \neq k} S_i \cup T_i\right))$$

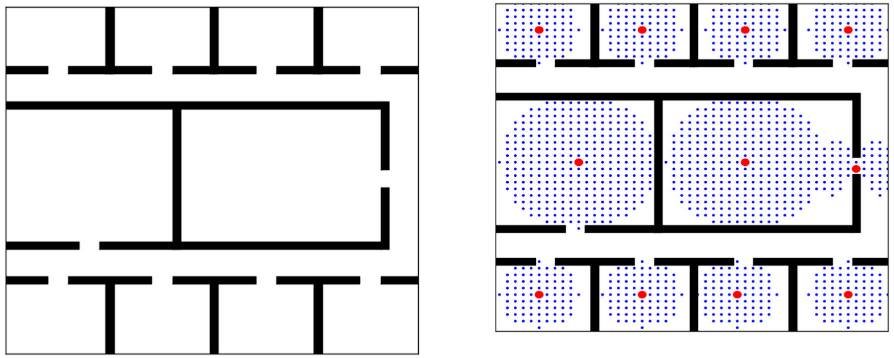
<u>A simpler definition</u>: A monotone function is k-submodular if each orthant (fix the domain of each element to be $\{0, i\}$ for some $i \in \{1,2, \dots, k\}$) is submodular.

Examples of k-submodularity

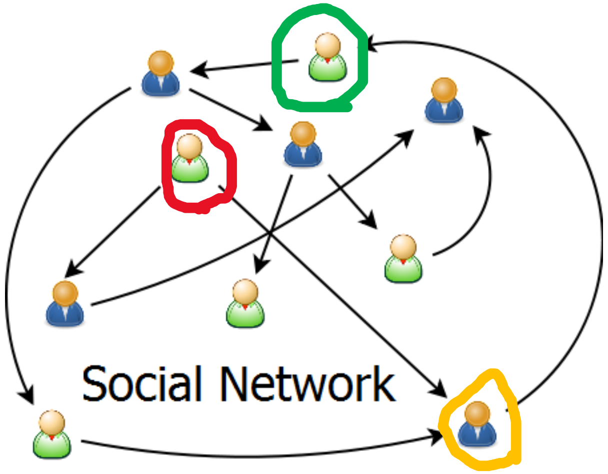
- Coupled feature selection
- Sensor placement with k kinds of measures
- Influence maximization with k topics
- Variant of facility location
- ….



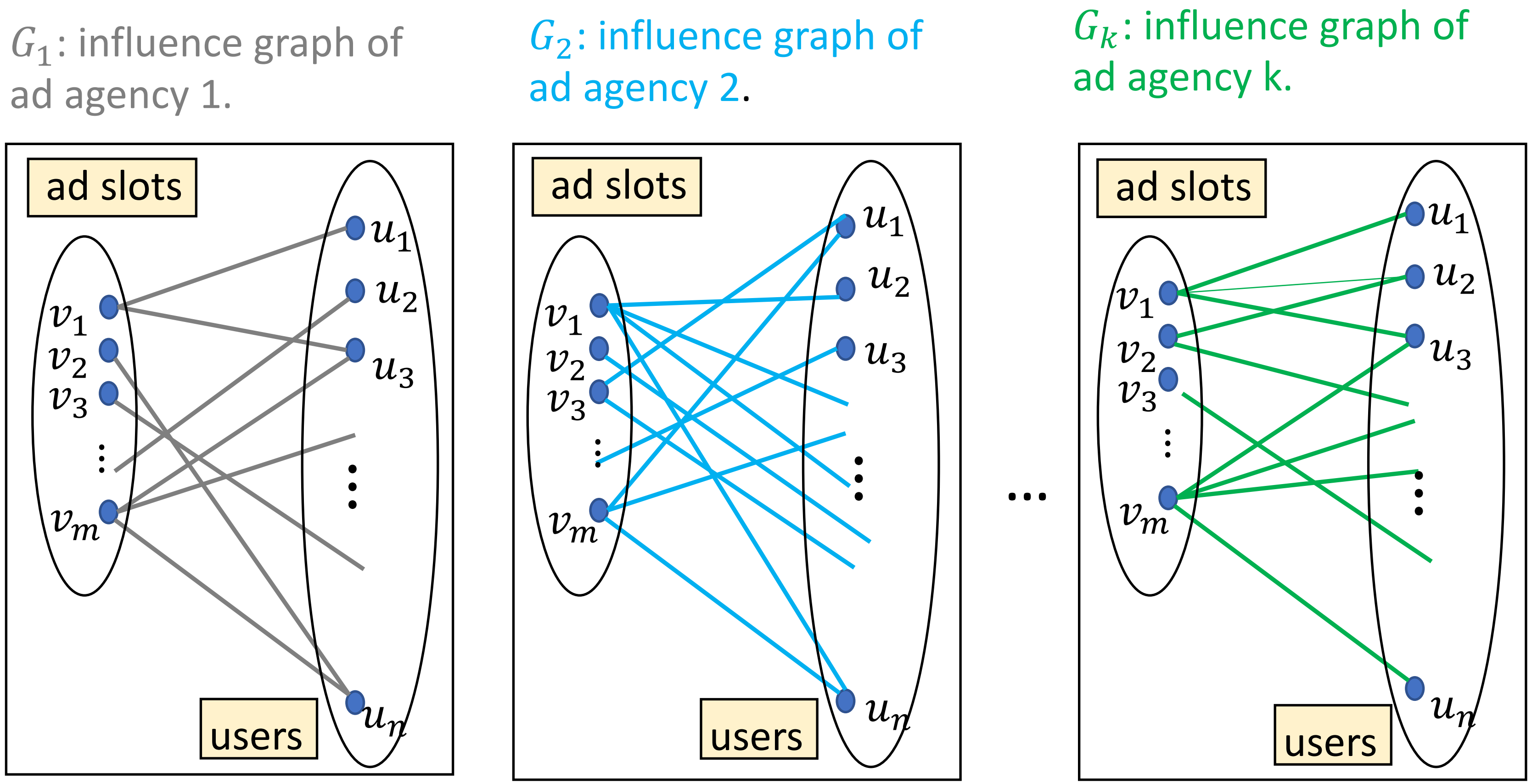
Picture from: **Near-optimal Sensor Placements** :
Maximizing Information while Minimizing Communication Cost.
A. Krause, A. Gupta, C. Guestrin, J. Kleinberg



Picture from: On Bisubmodular Maximization
A. P. Singh, A. Guillory, J. Bilmes



A toy example



Edges incident to a user u_i in $G_1, \ldots, G_k$ are sensitive data about u_i.

Objective: allocate at most B$\leq m$ ad slots to ad agencies so that it maximizes number of influenced users.

Our contributions

	non-private	previous result	our result
utility	$\frac{1}{2}OPT$	×	$\frac{1}{2}OPT - O(\frac{\Delta \cdot \mathrm{r(M)} \cdot \ln(\|E\|)}{\epsilon})$
privacy	×	×	$\epsilon . r(M)$

- Our algorithm is the first differentially private k-submodular maximization algorithm.
- $\left(\frac{1}{2}\right) OPT$ is asymptotically tight assuming P≠NP.
- Our algorithm uses almost linear number of function evaluations *i.e.,* $O(k \cdot |E| \cdot \ln(r(M)))$.

Thanks!

Definition of submodular function

A function $f: 2^E \rightarrow R$ is submodular if
- for all $A \subseteq B \subseteq E$,
- and all elements $e \in E \setminus B$ we have

$$f(\mathrm{A} \cup \{e\}) - f(A) \geq f(B \cup \{e\}) - f(B)$$

Applications

- Viral marketing
- Information gathering
- Feature selection for classification
- Influence maximization in social network
- Document summarization…

What is our objective?

We need an optimization method such that
- It returns almost an optimal solution
- It is efficient and fast
- Preserves individuals' privacy when we have sensitive data: medical data ,web search data, social networks

Differential privacy

A rigorous notion of privacy that allows statistical analysis of sensitive data while providing strong privacy guarantees.

Result 1

We present a differentially private algorithm for submodular maximization and:
- Prove that our algorithm returns a solution with quality at least

$$\left(1 - \frac{1}{e}\right) OPT + small\ additive\ error$$

- Prove that our algorithm preserve privacy
- Improve the number of function evaluations via a sampling technique while still preserving privacy

Result 2 (generalization of submodularity)

We present the first differentially private algorithm for k-submodular maximization and:
- Prove that our algorithm returns a solution with quality at least

$$\left(\frac{1}{2}\right) OPT + small\ additive\ error$$

- Prove our algorithm preserve privacy
- Reduce number of function evaluations to almost linear by a sampling technique while preserving privacy