

# Data Anonymization Techniques



Memorial

University of Newfoundland  
Department of Computer Science

Edward K. Brown, Ph.D., LL.B.

[brown@cs.mun.ca](mailto:brown@cs.mun.ca)

# Outline

- The Privacy Problem in large data collections
- Approaches to preserving privacy
- K-Anonymity algorithms
- Distributed applications
- Problems with model assumptions
- Future directions

# Privacy preservation problem

- Even with the removal of personal identifiers, inferences can be made about individuals
- In the worst case, the personal identity can be reconstructed from existing data

Subject re-identify →

<i>Name</i>	<i>Age</i>	<i>sex</i>	<i>Zipcode</i>
<i>Sami</i>	<i>19</i>	<i>M</i>	<i>02141</i>
<i>Amy</i>	<i>21</i>	<i>F</i>	<i>02144</i>
<i>Shelly</i>	<i>33</i>	<i>M</i>	<i>02142</i>
<i>Publicly available data</i>			

<i>Age</i>	<i>Sex</i>	<i>Zipcode</i>	<i>Disease</i>	<i>Cure</i>
<i>19</i>	<i>M</i>	<i>02141</i>	<i>Hepatitis</i>	<i>Yes</i>
<i>20</i>	<i>F</i>	<i>02144</i>	<i>Hepatitis</i>	<i>Yes</i>
<i>23</i>	<i>F</i>	<i>02145</i>	<i>AIDS</i>	<i>No</i>
<i>Hospital's data</i>				

# Approaches to Preserving Privacy

- Data Mining Approaches
  - Modify or perturb the data so certain rules cannot be inferred
  - Cryptographic/Probabilistic Approaches
    - response to specific queries give probabilistic results
    - multiple public keys for each users provide access to aggregate or individual data
  - Reconstruction Techniques (estimation)
- Perturbation (Statistical) techniques
- Anonymization Approach

# Anonymization

- External uses of person-specific data
  - intentional or incidental release of data
- Re-identification problem
- K-anonymization framework

Subject re-identify →

<i>Name</i>	<i>Age</i>	<i>sex</i>	<i>Zipcode</i>
<i>Sami</i>	<i>19</i>	<i>M</i>	<i>02141</i>
<i>Amy</i>	<i>21</i>	<i>F</i>	<i>02144</i>
<i>Shelly</i>	<i>33</i>	<i>M</i>	<i>02142</i>
<i>Publicly available data</i>			

<i>Age</i>	<i>Sex</i>	<i>Zipcode</i>	<i>Disease</i>	<i>Cure</i>
<i>19</i>	<i>M</i>	<i>02141</i>	<i>Hepatitis</i>	<i>Yes</i>
<i>20</i>	<i>F</i>	<i>02144</i>	<i>Hepatitis</i>	<i>Yes</i>
<i>23</i>	<i>F</i>	<i>02145</i>	<i>AIDS</i>	<i>No</i>
<i>Hospital's data</i>				

# Anonymization

- External uses of person-specific data
  - intentional or incidental release of data
- Re-identification problem
- K-anonymization framework
- 

Subject re-identify →

<i>Name</i>	<i>Age</i>	<i>sex</i>	<i>Zipcode</i>
<i>Sami</i>	<i>19</i>	<i>M</i>	<i>02141</i>
<i>Amy</i>	<i>21</i>	<i>F</i>	<i>02144</i>
<i>Shelly</i>	<i>33</i>	<i>M</i>	<i>02142</i>
<i>Publicly available data</i>			

<i>Age</i>	<i>Sex</i>	<i>Zipcode</i>	<i>Disease</i>	<i>Cure</i>
<i>19-20</i>	<i>*</i>	<i>0214*</i>	<i>Hepatitis</i>	<i>Yes</i>
<i>19-20</i>	<i>*</i>	<i>0214*</i>	<i>Hepatitis</i>	<i>Yes</i>
<i>23</i>	<i>F</i>	<i>02145</i>	<i>AIDS</i>	<i>No</i>
<i>Hospital's data</i>				

## • ***K-anonymization problem***

- **Quasi-Identifier Attribute Set:** is the columns that may be used to re-identify the subject of the table, such as Age, Sex, and Zipcode.
- **Frequency List:** Let  $T$  be a table where each row has  $n$  columns. Frequency List  $L$  is a binary relation:
  - $L = \{ \langle t_0, \dots, t_n \rangle v \mid \text{where each } \langle t_0, \dots, t_n \rangle \text{ is a unique combination of column values of } T \text{ and } v \text{ is number of occurrence of } \langle t_0, \dots, t_n \rangle \text{ in } T \}$
- **K-Anonymity:** Table  $T$  is  $K$ -anonymous if each  $v$  in its frequency list is  $\geq K$ .
- **K-Anonymization:** A modified table  $MT$  is a  $K$ -Anonymization of table  $T$  if  $MT$  alters the attribute values of  $T$  and is  $K$ -anonymous.
- **Generalization:** Generalizing an attribute is just replacing an attribute value with a less specific attribute value.
  - *For example, if  $\{42135, 42136, 42137\} \subset \text{ZIPcodes}$  then this set of attribute values can be generalized in a semantically consistent manner to  $\{4213^*\}$  by omitting the rightmost digit.*

# Efficiency issues

- optimal solution (anonymity with minimal information loss) known to be NP-hard
- heuristic approaches used

# ***Datafly Algorithm***

**Input:** Private Table **PT**, quasi-identifier set **QI** = { $A_1, \dots, A_n$ }, anonymity parameter **k**, and hierarchies **DGHA<sub>i</sub>**, where  $i=1, \dots, n$ .

**Output:** **MT** - a  $K$ -anonymization of **PT[QI]**.

**Method:**

1. **MT** **PT[QI]**; **freq** *Frequency List* of **MT**
2. **while there exists** frequencies in **freq** less than **k** which together represent at least **k** rows **do**
  - 2.1. **let**  $A_j$  be attribute in **MT** having the most number of distinct values
  - 2.2. **MT**[ $A_j$ ] *generalized* values of  $A_j$  according to **DGHA<sub>j</sub>**
  - 2.3. recalculate **freq** *Frequency List* of **MT**
3. *suppress* rows in **MT** occurring less than  $k$  times.
4. enforce  $k$  requirement on suppressed rows in **MT**
5. return **MT**

# GreedyRelease Algorithm

**Input:** Private Table **PT**, quasi-identifier set **QI** = { $A_1, \dots, A_n$ }, anonymity parameter **k**, and hierarchies **DGHA<sub>i</sub>**, where  $i=1, \dots, n$ .

**Output:** **MT** - a  $K$ -anonymization of **PT[QI]**.

**Method:**

1. **MT** = { }; **freq** = Frequency List of **PT[QI]**.

2. **while there exists** frequencies in **freq** less than **k** which together represent at least **k** rows **do**

2.1. **MT<sub>i</sub>** = {  $r$  : row of **PT[QI]** | **freq**( $r$ )  $\geq k$  }

2.2. **MT** = **MT**  $\cup$  **MT<sub>i</sub>**

2.3 **PT[QI]** = **PT[QI]** - **MT<sub>i</sub>** (Set difference)

2.4. **let**  $A_j$  be attribute in **PT[QI]** having the most number of distinct values

2.5. **PT[QI]** = generalized values of  $A_j$  according to **DGHA<sub>j</sub>**

2.6 recalculate **freq** = Frequency List of **PT[QI]**

3. return **MT**

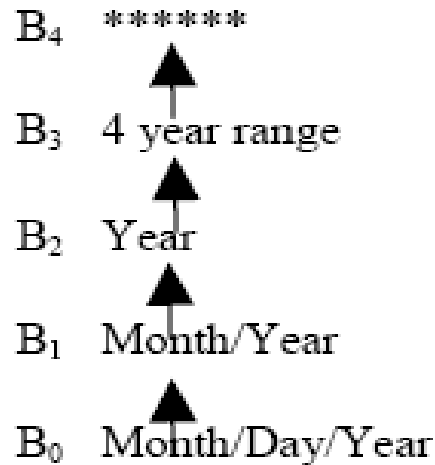
## ***Datafly vs Greedyrelease on a small data set***

<b><i>Race</i></b>	<b><i>BirthDate</i></b>	<b><i>Gender</i></b>	<b><i>ZIP</i></b>
<b><i>Black</i></b>	<b><i>8/20/1965</i></b>	<b><i>M</i></b>	<b><i>02141</i></b>
<b><i>Black</i></b>	<b><i>8/14/1965</i></b>	<b><i>M</i></b>	<b><i>02141</i></b>
<b><i>Black</i></b>	<b><i>8/23/1965</i></b>	<b><i>F</i></b>	<b><i>02138</i></b>
<b><i>Black</i></b>	<b><i>8/24/1965</i></b>	<b><i>F</i></b>	<b><i>02138</i></b>
<b><i>Black</i></b>	<b><i>10/7/1964</i></b>	<b><i>F</i></b>	<b><i>02138</i></b>
<b><i>Black</i></b>	<b><i>10/1/1964</i></b>	<b><i>F</i></b>	<b><i>02138</i></b>
<b><i>White</i></b>	<b><i>10/23/1964</i></b>	<b><i>M</i></b>	<b><i>02142</i></b>
<b><i>White</i></b>	<b><i>8/15/1965</i></b>	<b><i>F</i></b>	<b><i>02143</i></b>
<b><i>White</i></b>	<b><i>10/13/1964</i></b>	<b><i>M</i></b>	<b><i>02139</i></b>
<b><i>White</i></b>	<b><i>10/5/1964</i></b>	<b><i>M</i></b>	<b><i>02139</i></b>
<b><i>White</i></b>	<b><i>3/13/1967</i></b>	<b><i>M</i></b>	<b><i>02138</i></b>
<b><i>White</i></b>	<b><i>3/21/1967</i></b>	<b><i>M</i></b>	<b><i>02138</i></b>
<b><i>PT table prior to processing BirthData colum have most distinct value (12) .</i></b>			

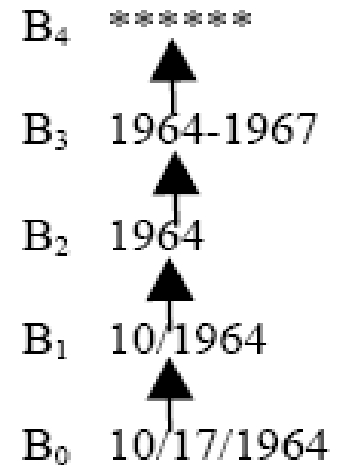
## Datafly vs Greedyrelease on a smaller data set

<i>Race</i>	<i>BirthDate</i>	<i>Gender</i>	<i>ZIP</i>
<i>Black</i>	8/1965	<i>M</i>	02141
<i>Black</i>	8/1965	<i>M</i>	02141
<i>Black</i>	8/1965	<i>F</i>	02138
<i>Black</i>	8/1965	<i>F</i>	02138
<i>Black</i>	10/1964	<i>F</i>	02138
<i>Black</i>	10/1964	<i>F</i>	02138
<i>White</i>	10/1964	<i>M</i>	02142
<i>White</i>	8/1965	<i>F</i>	02143
<i>White</i>	10/1964	<i>M</i>	02139
<i>White</i>	10/1964	<i>M</i>	02139
<i>White</i>	3/1967	<i>M</i>	02138
<i>White</i>	3/1967	<i>M</i>	02138

*PT table after Generalizing BirthDate by both algorithms and ZIP column have most distinct value (5).*



DGH<sub>Birthdate</sub>

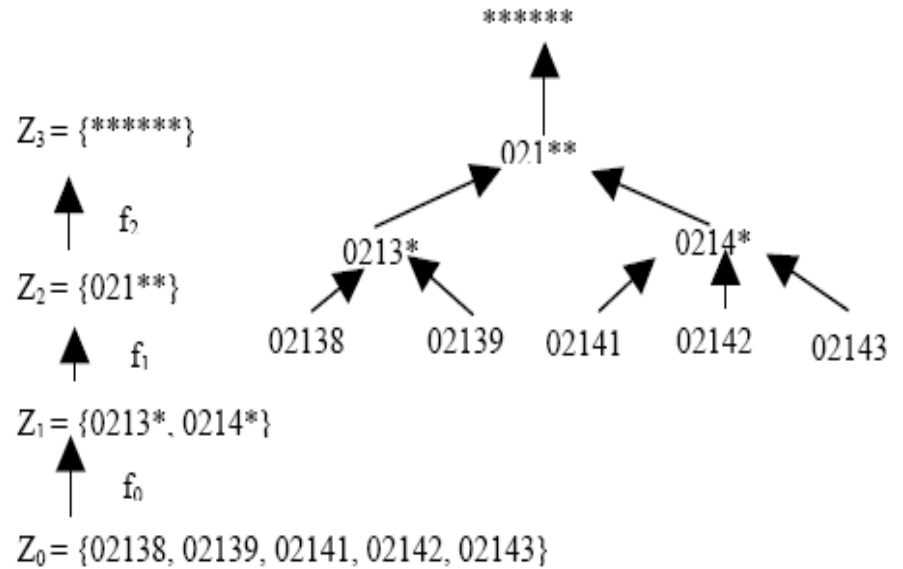


Value generalization example

## Datafly vs Greedyrelease on a smaller data set

Race	BirthDate	Gender	ZIP
Black	8/1965	M	0214
Black	8/1965	M	0214
Black	8/1965	F	0213
Black	8/1965	F	0213
Black	10/1964	F	0213
Black	10/1964	F	0213
White	10/1964	M	0214
White	8/1965	F	0214
White	10/1964	M	0213
White	10/1964	M	0213
White	3/1967	M	0213
White	3/1967	M	0213

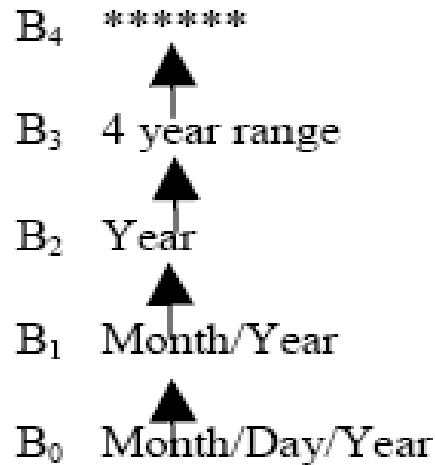
*PT table after Generalizing ZIP by both algorithms and BirthData column have most distinct value (3).*



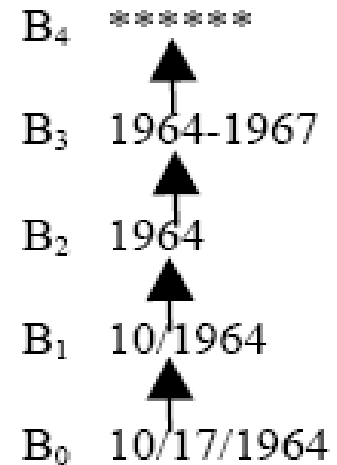
## Datafly vs Greedyrelease on a smaller data set

<i>Race</i>	<i>BirthDate</i>	<i>Gender</i>	<i>ZIP</i>
<i>Black</i>	1965	<i>M</i>	<i>0214</i>
<i>Black</i>	1965	<i>M</i>	<i>0214</i>
<i>Black</i>	1965	<i>F</i>	<i>0213</i>
<i>Black</i>	1965	<i>F</i>	<i>0213</i>
<i>Black</i>	1964	<i>F</i>	<i>0213</i>
<i>Black</i>	1964	<i>F</i>	<i>0213</i>
<i>White</i>	1964	<i>M</i>	<i>0214</i>
<i>White</i>	1965	<i>F</i>	<i>0214</i>
<i>White</i>	1964	<i>M</i>	<i>0213</i>
<i>White</i>	1964	<i>M</i>	<i>0213</i>
<i>White</i>	1967	<i>M</i>	<i>0213</i>
<i>White</i>	1967	<i>M</i>	<i>0213</i>

*PT table after Generalizing BirthDate by both algorithms and BirthDate have most distinct value (3).*



DGH<sub>Birthdate</sub>

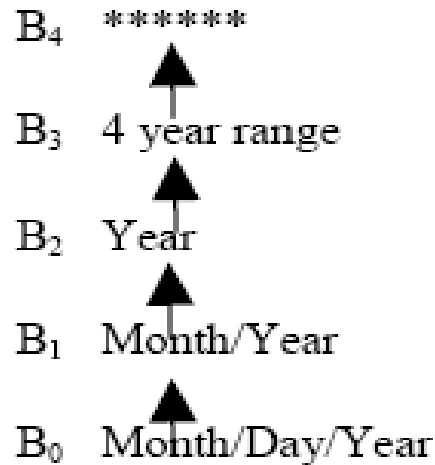


Value generalization example

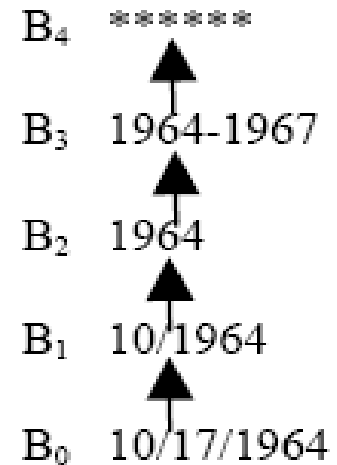
## Datafly vs Greedyrelease on a smaller data set

<i>Race</i>	<i>BirthDate</i>	<i>Gender</i>	<i>ZIP</i>
<i>Black</i>	1964-1967	<i>M</i>	<i>0214</i>
<i>Black</i>	1964-1967	<i>M</i>	<i>0214</i>
<i>Black</i>	1964-1967	<i>F</i>	<i>0213</i>
<i>Black</i>	1964-1967	<i>F</i>	<i>0213</i>
<i>Black</i>	1964-1967	<i>F</i>	<i>0213</i>
<i>Black</i>	1964-1967	<i>F</i>	<i>0213</i>
<i>White</i>	1964-1967	<i>M</i>	<i>0214</i>
<i>White</i>	1964-1967	<i>F</i>	<i>0214</i>
<i>White</i>	1964-1967	<i>M</i>	<i>0213</i>
<i>White</i>	1964-1967	<i>M</i>	<i>0213</i>
<i>White</i>	1964-1967	<i>M</i>	<i>0213</i>
<i>White</i>	1964-1967	<i>M</i>	<i>0213</i>

*PT table after Generalizing BirthDate by both algorithms and Race, Gender, and Zip columns have most distinct value (2). Thus, Race is generalized.*



DGH<sub>Birthdate</sub>



Value generalization example

# Datafly vs Greedyrelease on a smaller data set

Already met K-anonymity requirement

Race	BirthDate	Gender	ZIP
Person	1964-1967	M	0214
Person	1964-1967	M	0214
Black	1964-1967	F	0213
Black	1964-1967	F	0213
Black	1964-1967	F	0213
Black	1964-1967	F	0213
Person	1964-1967	M	0214
Person	1964-1967	F	0214
White	1964-1967	M	0213
White	1964-1967	M	0213
White	1964-1967	M	0213
White	1964-1967	M	0213

PT table after Generalizing Race by GreedyRelease

$R_2 = \{*****\}$

$R_1 = \{\text{Person}\}$

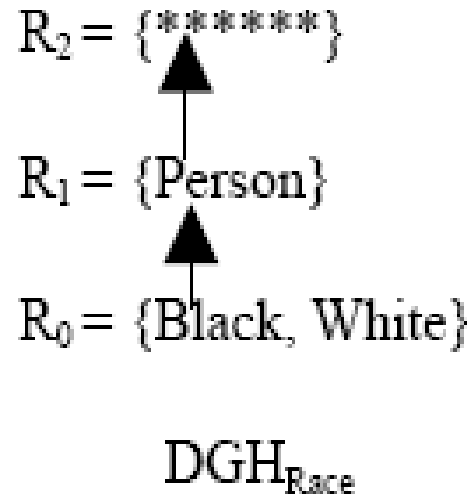
$R_0 = \{\text{Black, White}\}$

$DGH_{\text{Race}}$



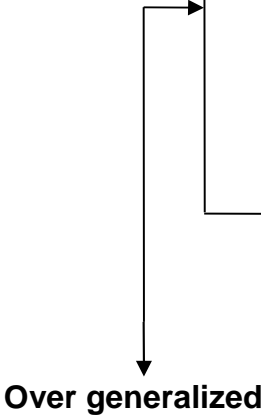
## Datafly vs Greedyrelease on a smaller data set

<i>Race</i>	<i>BirthDate</i>	<i>Gender</i>	<i>ZIP</i>
<i>Person</i>	<i>1964-1967</i>	<i>M</i>	<i>0214</i>
<i>Person</i>	<i>1964-1967</i>	<i>M</i>	<i>0214</i>
<i>Person</i>	<i>1964-1967</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-1967</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-1967</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-1967</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-1967</i>	<i>M</i>	<i>0214</i>
<i>Person</i>	<i>1964-1967</i>	<i>F</i>	<i>0214</i>
<i>Person</i>	<i>1964-1967</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-1967</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-1967</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-1967</i>	<i>M</i>	<i>0213</i>
<i>PT table after Generalizing Race by Datafly</i>			

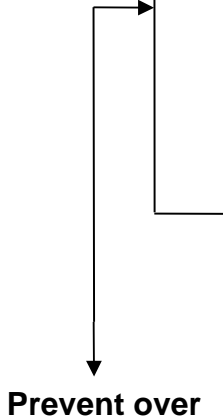


# Datafly vs Greedyrelease on a small data set

<i>Race</i>	<i>BirthDate</i>	<i>Gender</i>	<i>ZIP</i>
<i>Person</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0214</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0214</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0214</i>
<i>MT table produced by Datafly</i>			



<i>Race</i>	<i>BirthDate</i>	<i>Gender</i>	<i>ZIP</i>
<i>Black</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>Black</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>Black</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>Black</i>	<i>1964-67</i>	<i>F</i>	<i>0213</i>
<i>White</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>White</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>White</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>White</i>	<i>1964-67</i>	<i>M</i>	<i>0213</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0214</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0214</i>
<i>Person</i>	<i>1964-67</i>	<i>M</i>	<i>0214</i>
<i>MT table produced by GreedyRelease</i>			



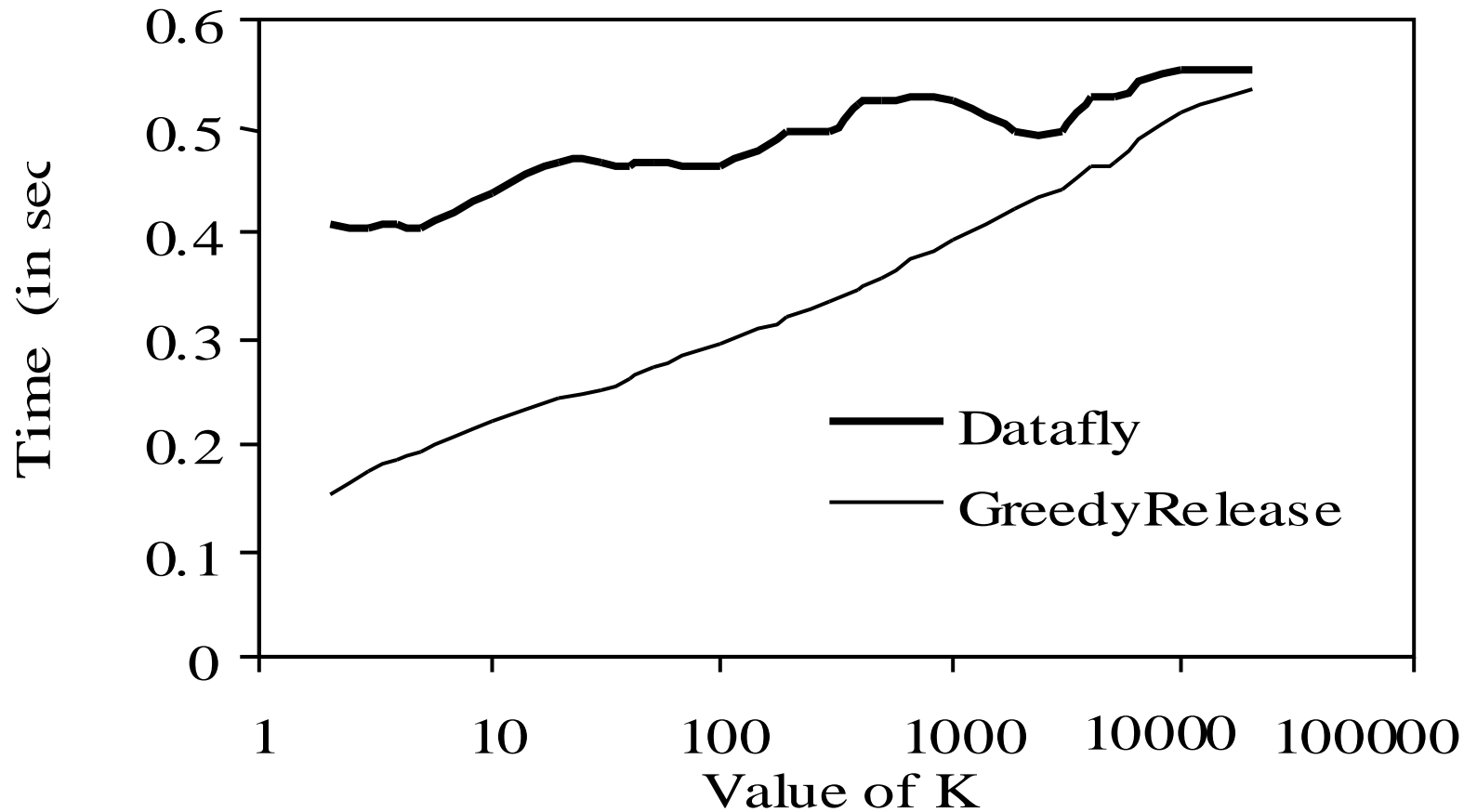
Over generalized

Prevent over generalization



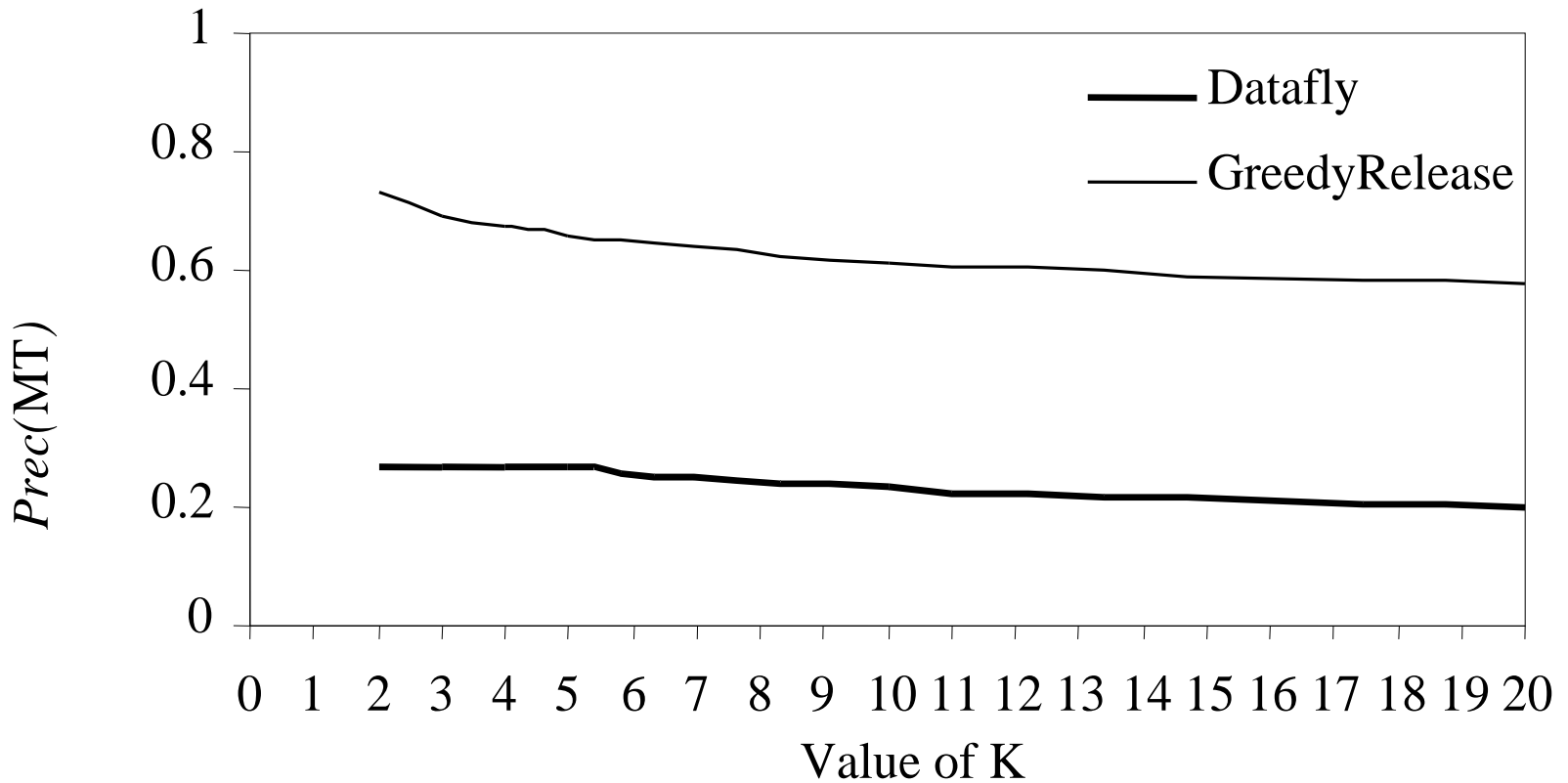
# ***Datafly vs Greedyrelease on Real-World Data***

Running time Datafly v. GreedyRelease

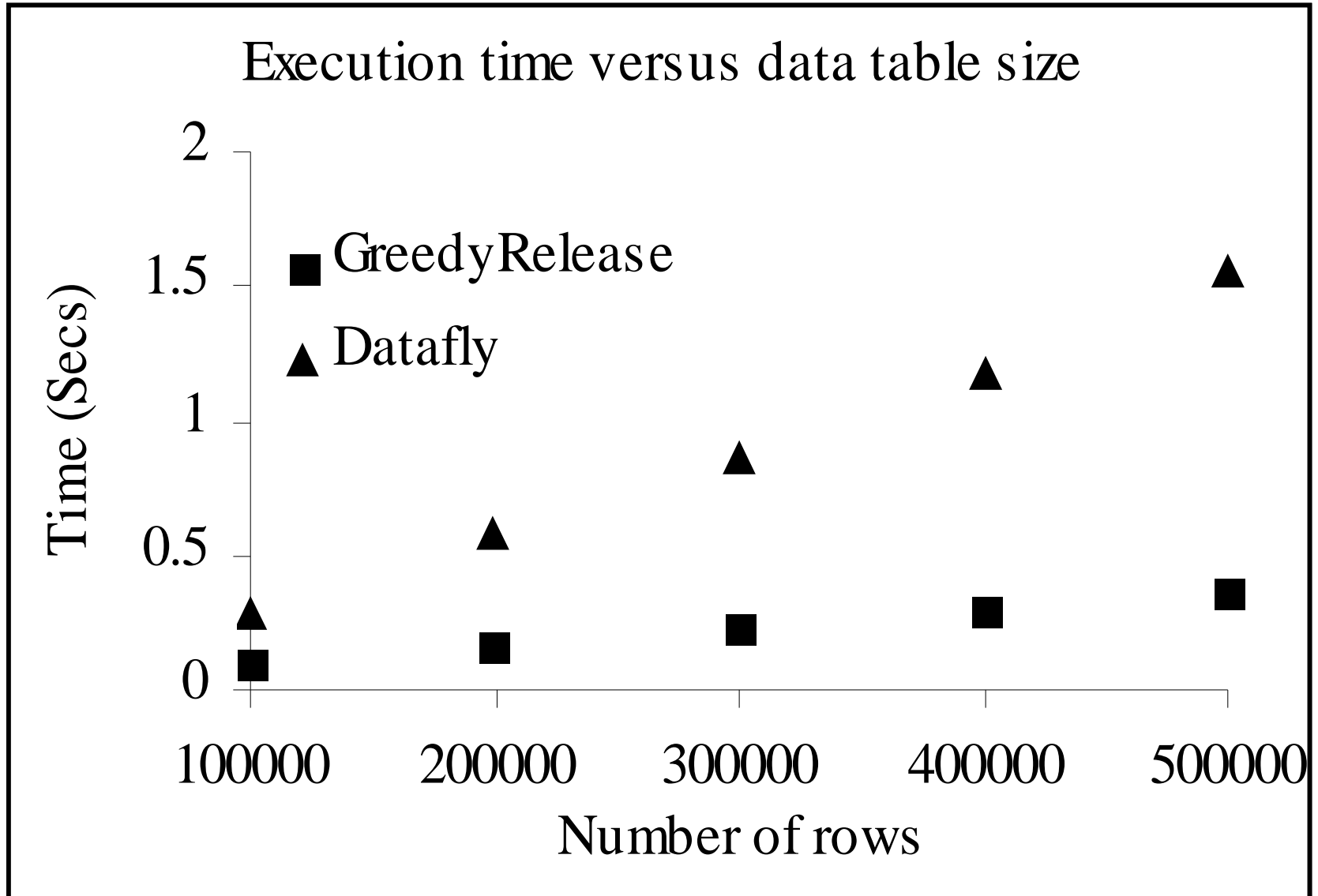


# ***Datafly vs Greedyrelease on Real-World Data***

Quality of output for different Algorithms



## ***Datafly vs Greedyrelease on Synthetic Data***



## ***k-anonymity results***

**Improvement in terms of execution time and output results.**

**Mixed domain data values**

# Distributed Data

- data can be partitioned horizontally (different attributes/columns at different sites) or partitioned vertically (different records/rows at different sites)
- cryptographic approaches work well
  - but may have to collect information from multiple sites to extract aggregate information
- k-anonymity approach works well for horizontal partitions
  - each site can maintain its own k-anonymous data
  - aggregating data will preserve minimum k-value
  - future research: adding redundant values to a site to increase anonymity

# Problems with the model(s)

- balance between anonymity and value of data sets
- non-identifier fields (similarity among fields)
- no modelling of actual population
  - assumed the data records reflect actual population
  - accessibility of “real-world” demographics
- dynamic or multiple data releases
- alternative inferential attacks
  - temporal based inference attacks
  - inter-record attacks
  - probabilistic inference (i.e. probabilistic results)
- attacks are assumed to be looking for individual records
- consent required for use of personal data (not personally identifiable information)

# Future directions

- distributed algorithms
- dynamic anonymization
- eliminate generalization hierarchies
- parameterize for specific applications
- temporal attacks

# Bibliography

- L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression”. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; pp. 571-588.
- G. Agrawal, et al. “Anonymizing Tables”, *International Conference on Database Theory*, Edinburg, Scotland, 2005, pp 246-258.
- P. Samarati: “Protecting Respondents' Identities in Microdata Release”, *IEEE Transactions on Knowledge and Data Engineering*,. 13(6), 2001 pp 1010-1027.
- K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient Full-Domain K-Anonymity”, *ACM SIGMOD '2005*, Baltimore, Maryland, June 14-16, 2005.
- K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Multidimensional K-Anonymity”, *Technical Report 1521*, Department of Computer Science, University of Wisconsin, Madison, June 22, 2005.
- V. Verykios et al, “State of the Art in Privacy preserving Data Mining”, *SIGMOD Record*, 33 (1), 2004.