What every data scientist should know about data anonymization

Katharina Rasch

Data scientist | PhD Computer Science

kat@krasch.io | www.krasch.io

New York City taxi drop offs 2009 - 2015 [1]

Medical Data Released as Anonymous

SSN	Name	Race	DateOfBirth	Sex	ZIP	Marital Status	HealthProblem
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	....
Sue J. Carlson	900 Market St.	San Francisco	*94142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	....

Figure 1: Re-identifying anonymous data by linking to external data

Massachusetts medical data breach [2], figure from [3]

1. introduction
2. let’s anonymize a dataset
3. utility of anonymized data
4. alternative: the interactive model
5. practical tips and standards

sensitive information?

assumptions [4]

a) publish raw data, not statistics

b) want to minimize the risk for privacy breaches

c) all record owners have equal right to privacy

d) adversaries are determined, resourceful and technically competent

e) the de-anonymization algorithms that attackers will use are unknown

assume adversary has access to additional data

whether target is/is not in the dataset

external data sets (public or closed)

personal knowledge about target

in the US, 63% of inhabitants are likely to be uniquely identifiable by birthdate, sex, zip code [5]

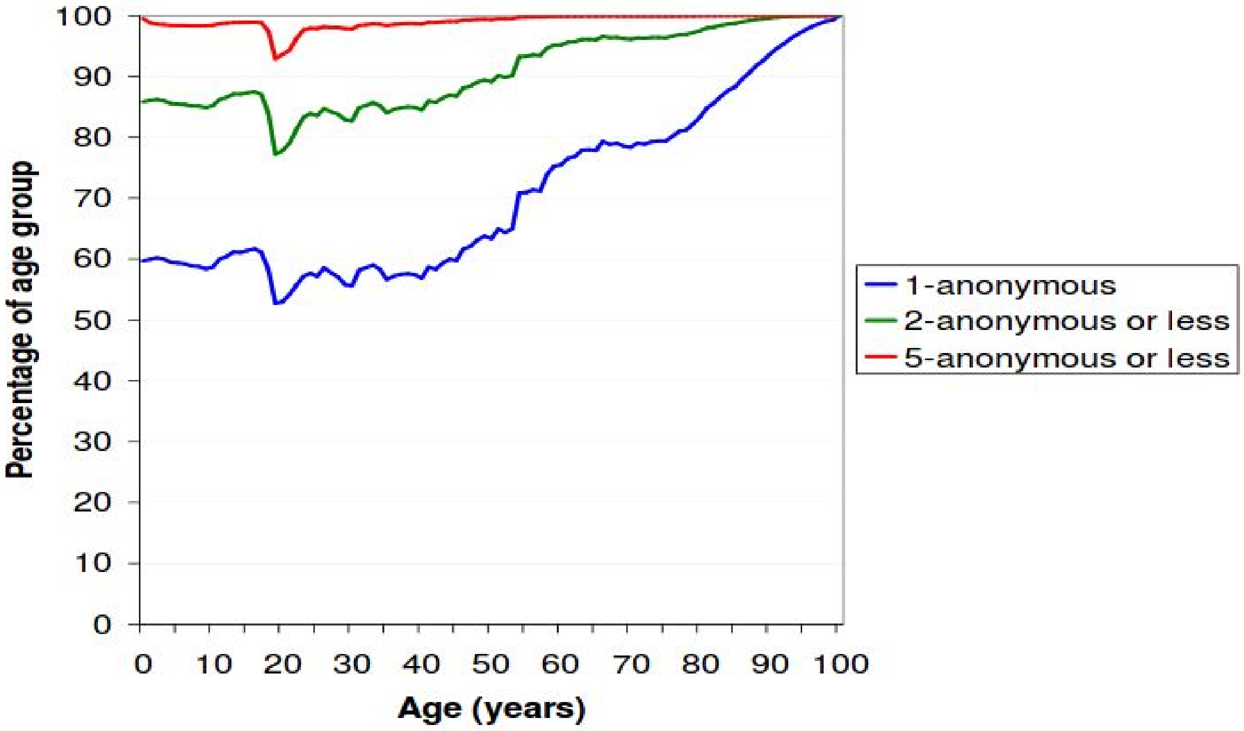


figure from [5]

age, sex, zip code, ethnicity, education, etc

are called

quasi identifiers

sex	age	location	profession	party affiliation
M	22	Dresden	sales	C
F	51	Heidelberg	software engineer	A
M	27	Leipzig	sales	A
F	23	Potsdam	nurse	B
F	54	Heidelberg	data scientist	A
F	62	Cologne	chef	C
M	43	Cologne	plumber	A

sex	age	location	profession	party affiliation
M	22	Dresden	sales	C
F	51	Heidelberg	software engineer	A
M	27	Leipzig	sales	A
F	23	Potsdam	nurse	B
F	54	Heidelberg	data scientist	A
F	62	Cologne	chef	C
M	43	Cologne	plumber	A

record linkage

target can be linked to one or very few records in the dataset

k-anonymity [2]

there must always be at least k records for each equivalence group present in the dataset

(equivalence group = all records with same combination of quasi-identifiers)

sex	age	location	profession	party affiliation
M	22	Dresden	sales	C
F	51	Heidelberg	software engineer	A
M	27	Leipzig	sales	A
F	23	Potsdam	nurse	B
F	54	Heidelberg	data scientist	A
F	62	Cologne	chef	C
M	43	Cologne	plumber	A

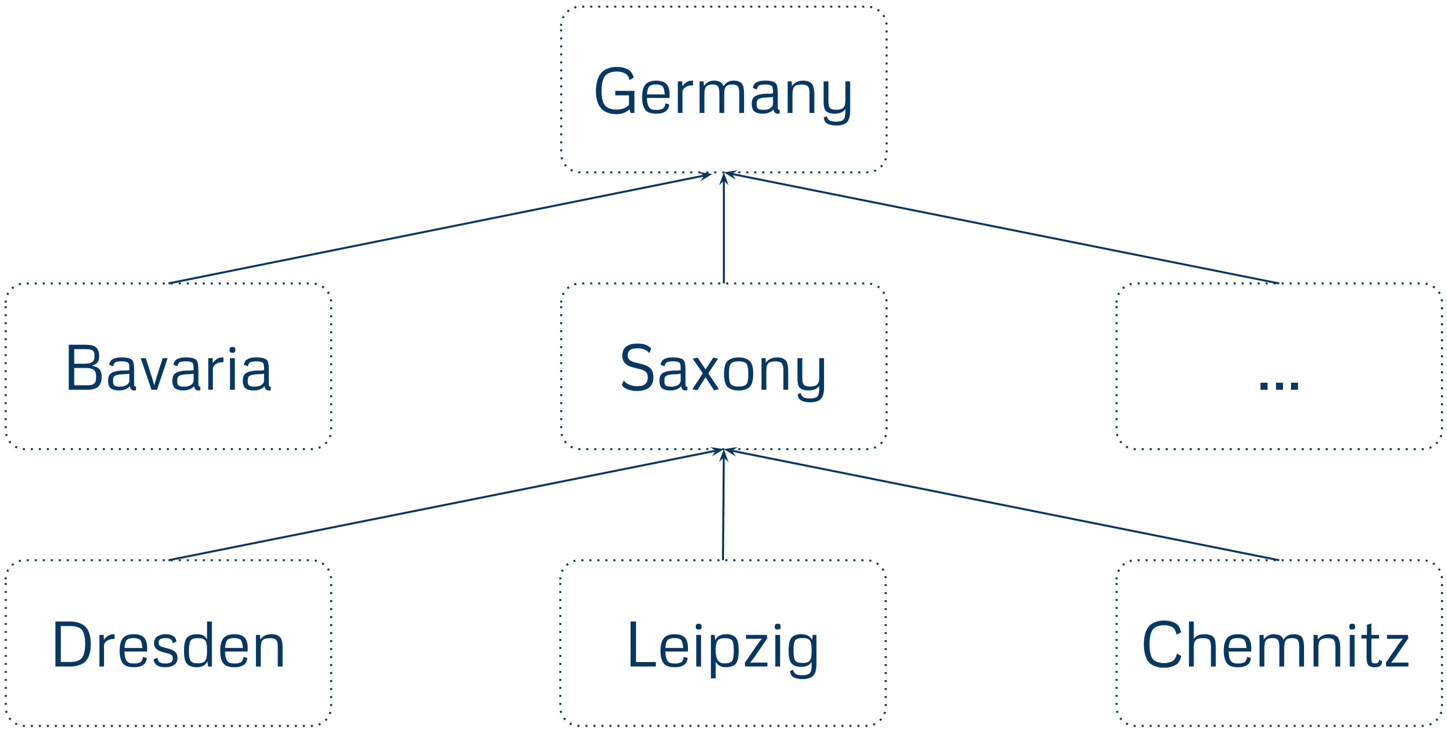
sex	age	location	profession	party affiliation
M	22	Dresden	sales	C
F	51	Heidelberg	software engineer	A
M	27	Leipzig	sales	A
F	23	Potsdam	nurse	B
F	54	Heidelberg	data scientist	A
F	62	Cologne	chef	C
M	43	Cologne	plumber	A

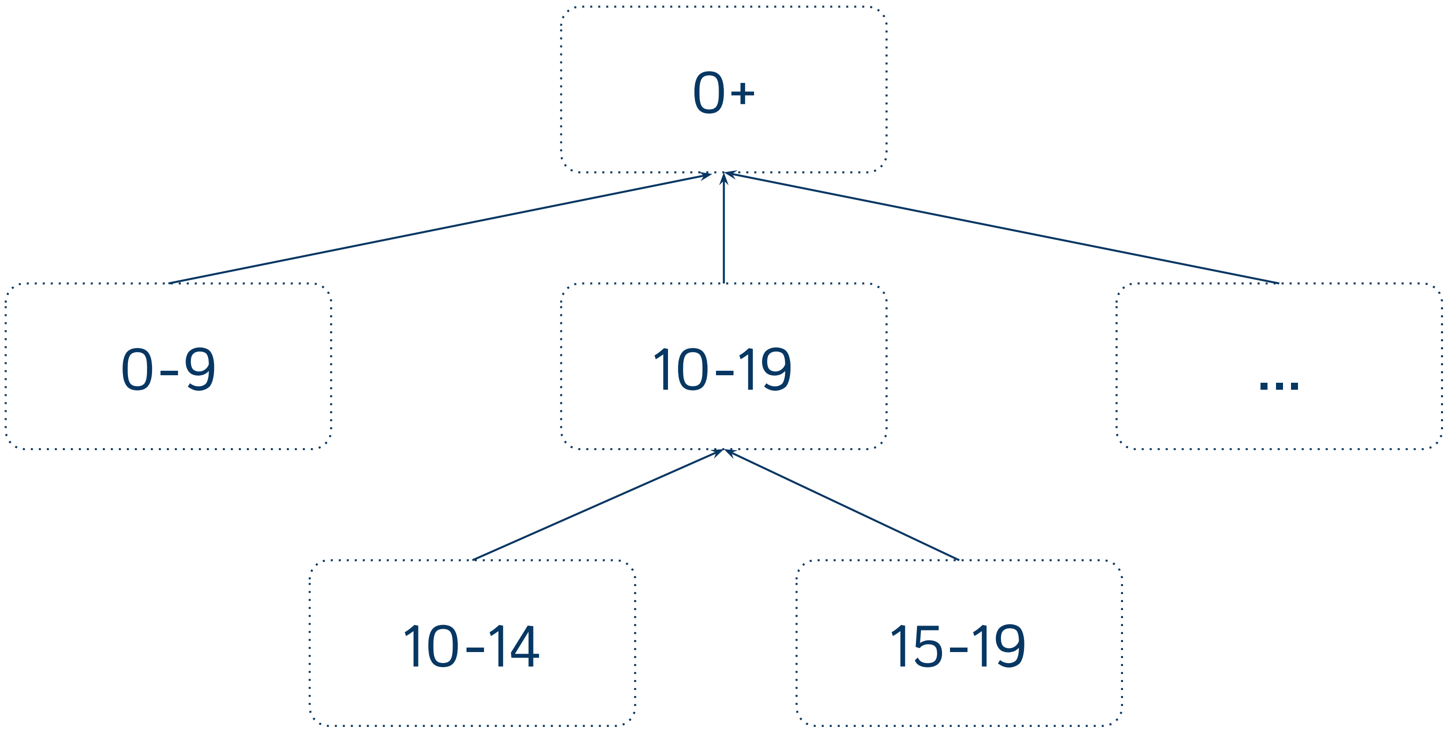
$k = 1$

let’s anonymize so that k=2

increase k through

generalization
suppression


Germany
Bavaria
Saxony
...
Dresden
Leipzig
Chemnitz


0+
0-9
10-19
...
10-14
15-19

sex	age	location	profession	party affiliation
M	21	Dresden	sales	C
F	51	Heidelberg	software engineer	A
M	26	Leipzig	sales	A
F	23	Potsdam	nurse	B
F	54	Heidelberg	data scientist	A
F	62	Cologne	electrician	C
M	43	Cologne	plumber	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	51	Heidelberg	software engineer	A
M	20-29	Saxony	sales	A
F	23	Potsdam	nurse	B
F	54	Heidelberg	data scientist	A
F	62	Cologne	electrician	C
M	43	Cologne	plumber	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	51	Heidelberg	software engineer	A
M	20-29	Saxony	sales	A
F	23	Potsdam	nurse	B
F	54	Heidelberg	data scientist	A
F	62	Cologne	electrician	C
M	43	Cologne	plumber	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
F	23	Potsdam	nurse	B
F	50-59	Heidelberg	engineer	A
F	62	Cologne	electrician	C
M	43	Cologne	plumber	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
F	23	Potsdam	nurse	B
F	50-59	Heidelberg	engineer	A
F	62	Cologne	electrician	C
M	43	Cologne	plumber	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
F	23	Potsdam	nurse	B
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	craftsperson	C
any	40-60	Cologne	craftsperson	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
F	23	Potsdam	nurse	B
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	craftsperson	C
any	40-60	Cologne	craftsperson	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
-	-	-	-	-
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	craftsperson	C
any	40-60	Cologne	craftsperson	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
-	-	-	-	-
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	blue collar	C
any	40-60	Cologne	blue collar	A

$k = 2$

success!

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
-	-	-	-	-
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	craftsperson	C
any	40-60	Cologne	craftsperson	A

attribute linkage

target is vulnerable because some sensitive values dominate target's equivalence group

l-diversity [6]

there must always be at least l distinct values for each sensitive attribute and equivalence group

(includes k-anonymity with $k \geq l$)

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
-	-	-	-	-
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	craftsperson	C
any	40-60	Cologne	craftsperson	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
-	-	-	-	-
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	craftsperson	C
any	40-60	Cologne	craftsperson	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
-	-	-	-	-
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	blue collar	C
any	40-60	Cologne	blue collar	A

l = 1

let's anonymize so that l=2

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
F	50-59	Heidelberg	engineer	A
M	20-29	Saxony	sales	A
-	-	-	-	-
F	50-59	Heidelberg	engineer	A
any	40-60	Cologne	craftsperson	C
any	40-60	Cologne	craftsperson	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
any	40-60	Germany	any	A
M	20-29	Saxony	sales	A
-	-	-	-	-
any	40-60	Germany	any	A
any	40-60	Germany	any	C
any	40-60	Germany	any	A

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
any	40-60	Germany	any	A
M	20-29	Saxony	sales	A
-	-	-	-	-
any	40-60	Germany	any	A
any	40-60	Germany	any	C
any	40-60	Germany	any	A

$l = 2$

success!

sex	age	location	profession	party affiliation
M	20-29	Saxony	sales	C
any	40-60	Germany	any	A
M	20-29	Saxony	sales	A
-	-	-	-	-
any	40-60	Germany	any	A
any	40-60	Germany	any	C
any	40-60	Germany	any	A

t-closeness [7]

distribution of a sensitive attribute in any equivalence group must be close to this attribute's distribution in whole dataset

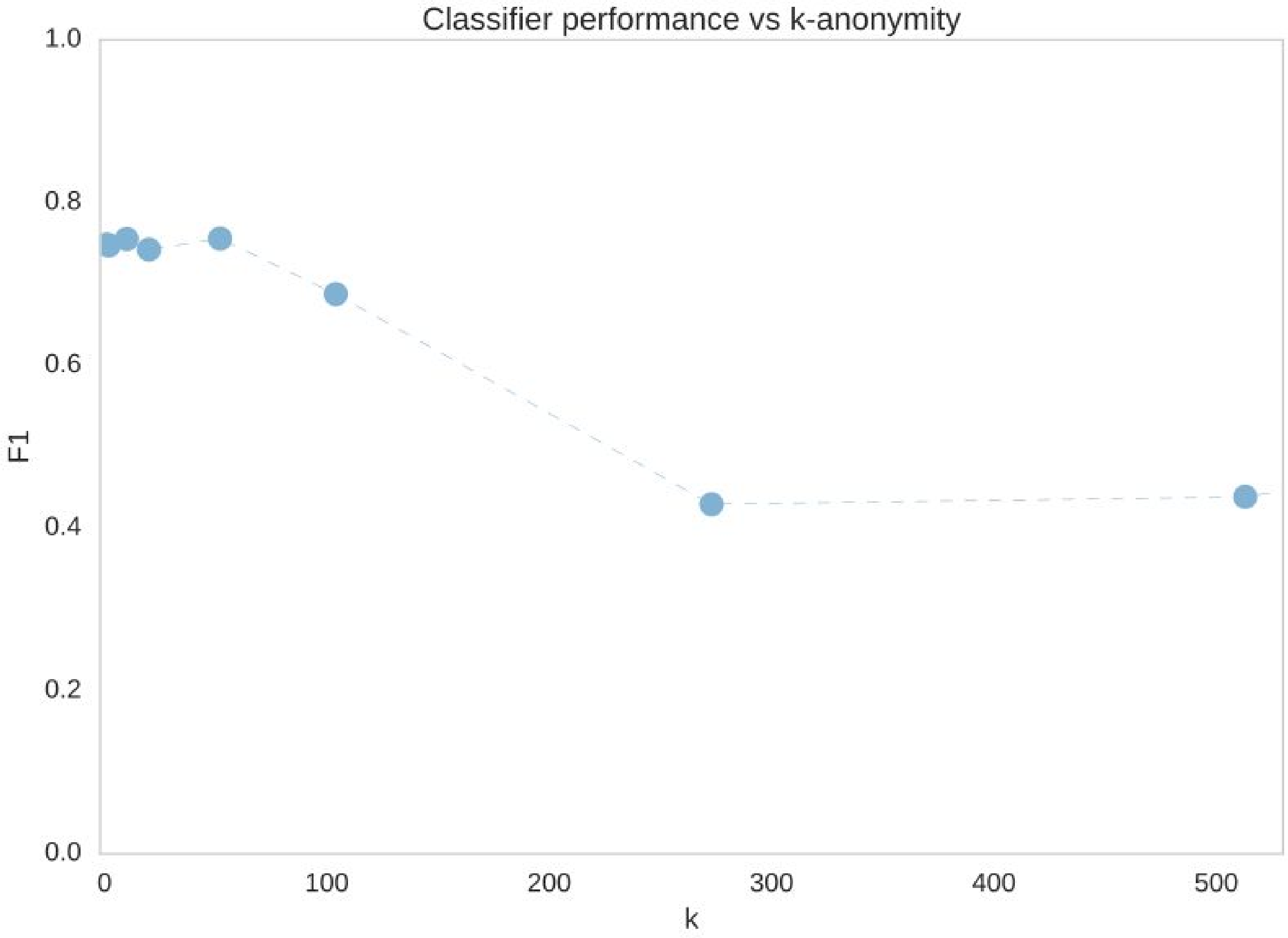
experiment: anonymization vs data utility

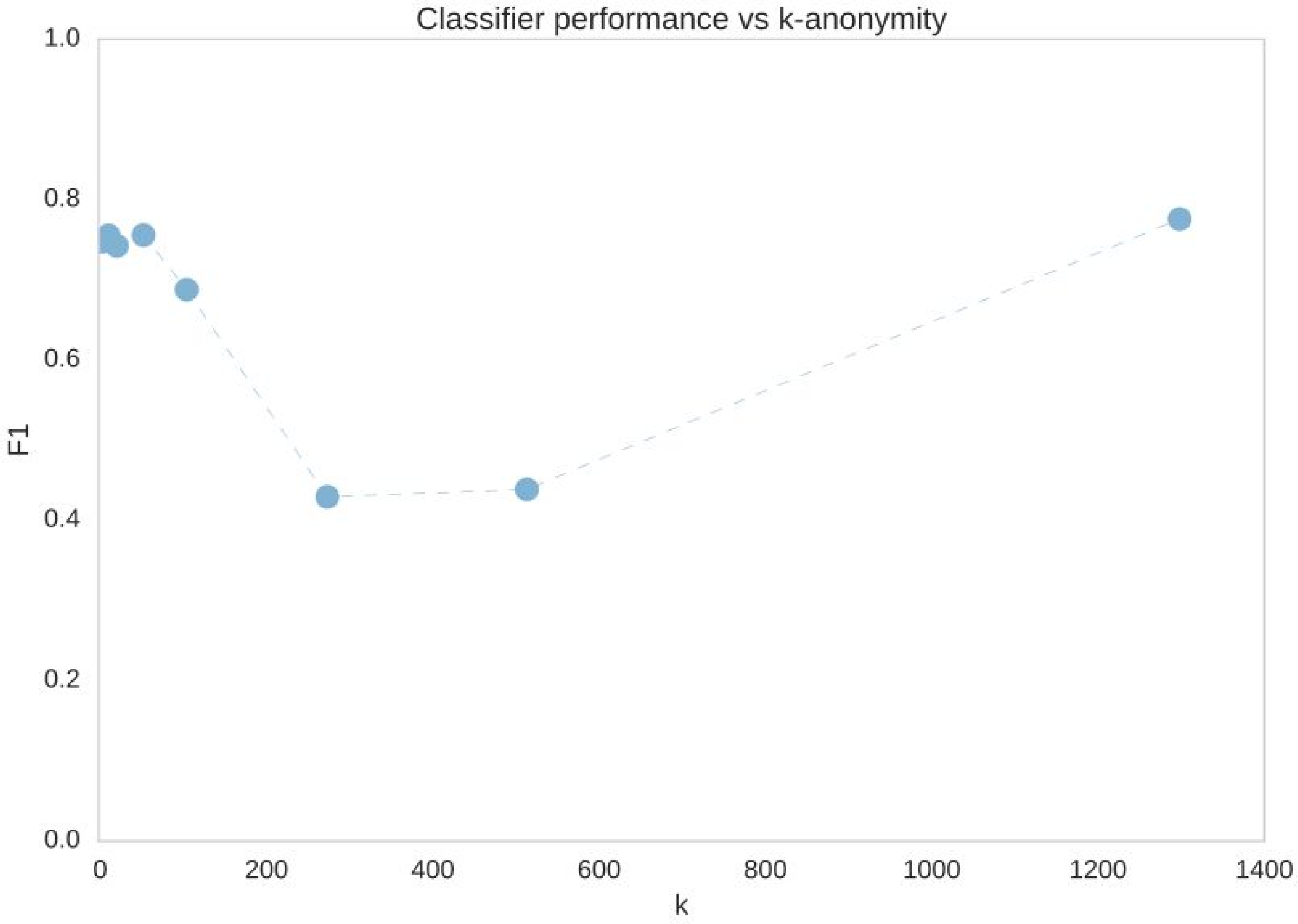
'adult' dataset

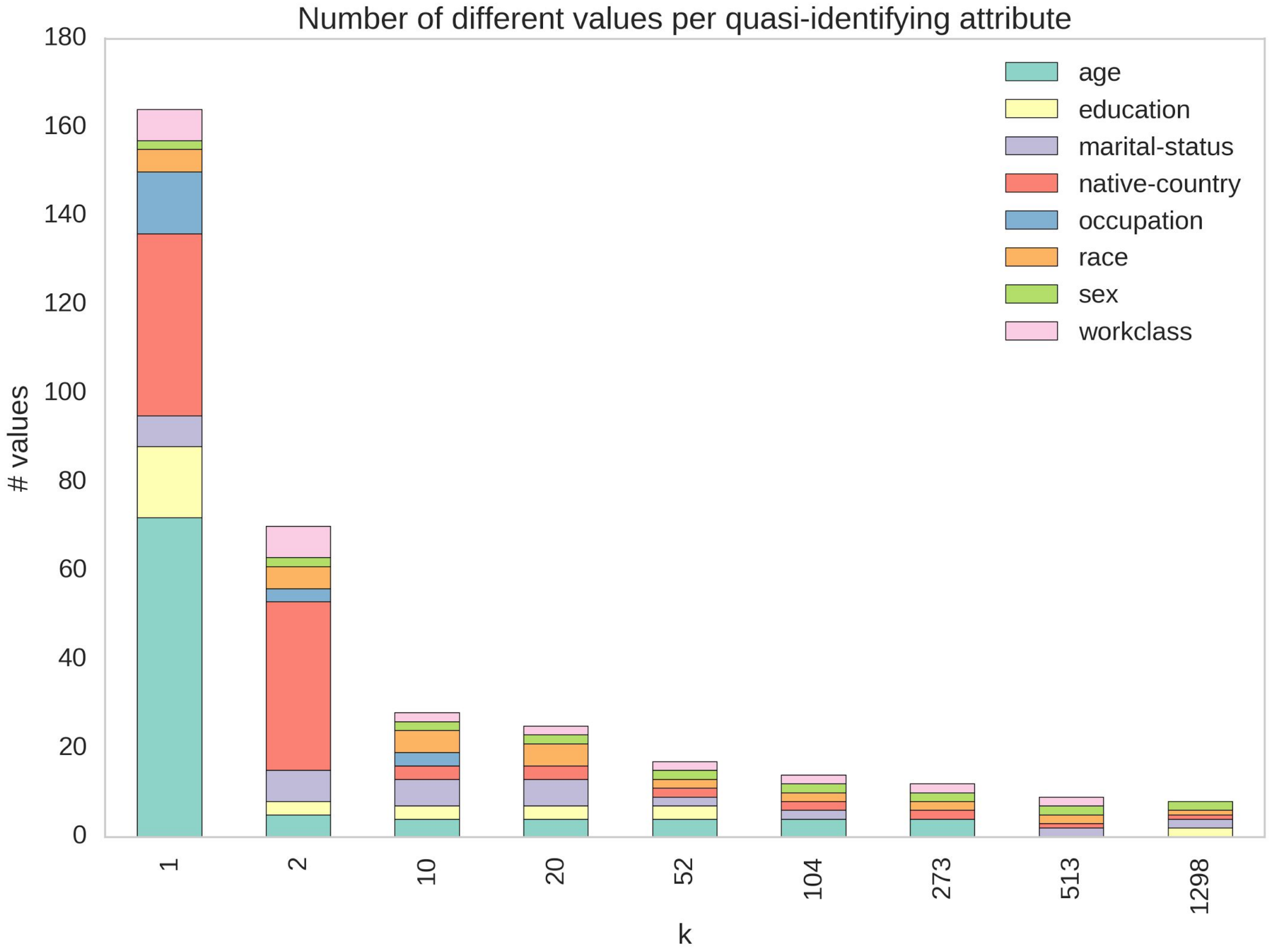
arx de-identification software [8]

standard settings (minimize information loss)

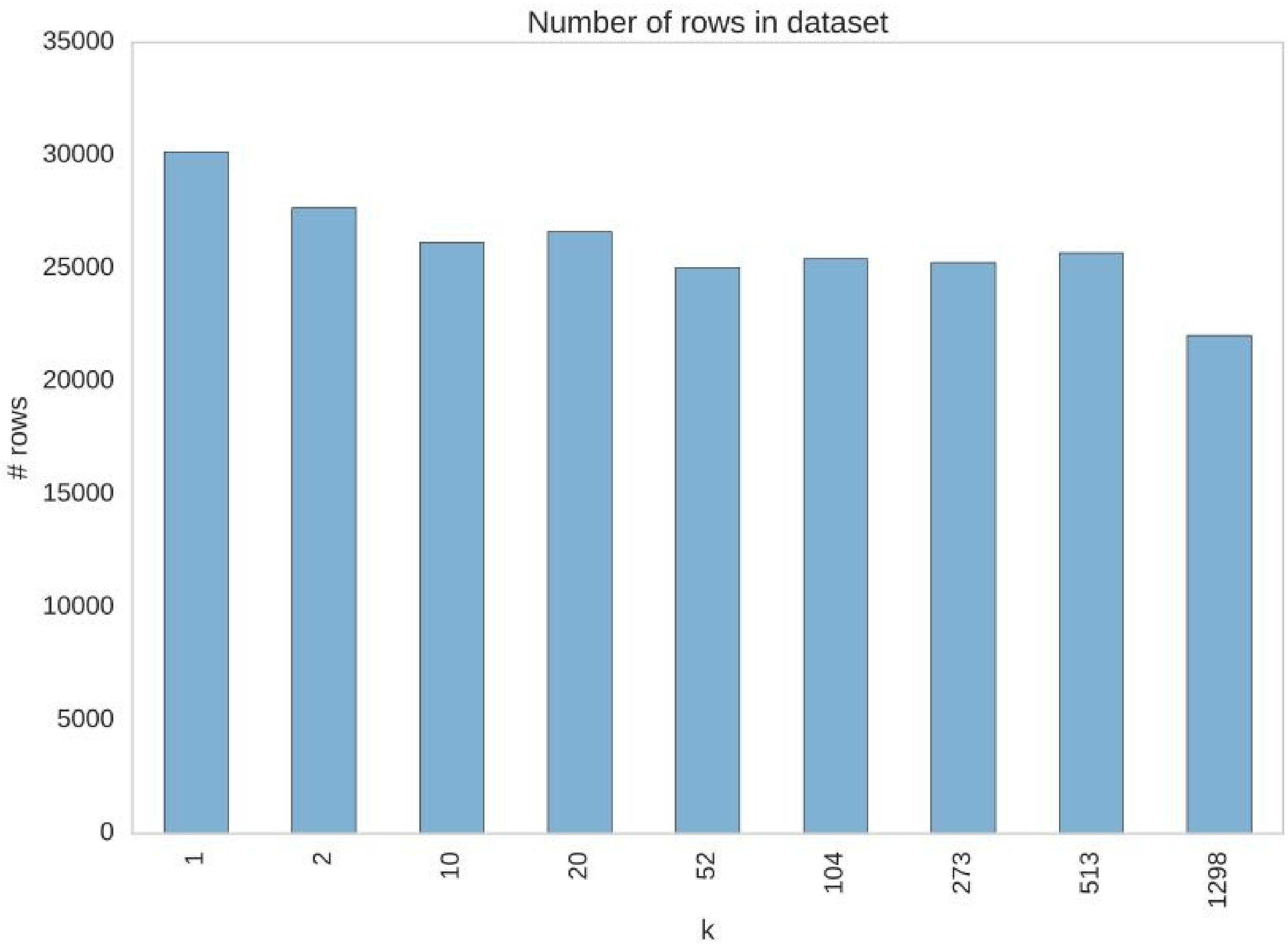
logistic regression


Classifier performance vs k-anonymity
1.0
0.8
0.6
0.4
0.2
0.0
F1
0
100
200
300
400
500
k


Classifier performance vs k-anonymity
F1
k
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
1400



Number of different values per quasi-identifying attribute
values
180
160
140
120
100
80
60
40
20
0
age
education
marital-status
native-country
occupation
race
sex
workclass
1
2
10
20
52
104
273
513
1298
k


Number of rows in dataset
35000
30000
25000
20000
15000
10000
5000
0
rows
1
2
10
20
52
104
273
513
1298
k

'Higher education', 'spouse not present', 'United-States', 'White', 'Female'	<=50K	3367
'Higher education', 'spouse not present', 'United-States', 'White', 'Male'	<=50K	2754
'Higher education', 'spouse present', 'United-States', 'White', 'Male'	>50K	3725
	<=50K	2448
'Secondary education', 'spouse not present', 'United-States', 'White', 'Female'	<=50K	2494
'Secondary education', 'spouse not present', 'United-States', 'White', 'Male'	<=50K	2752
'Secondary education', 'spouse present', 'United-States', 'White', 'Male'	<=50K	3199
	>50K	1298

most anonymization methods fail for

high-dimensional, sparse datasets [9]

e.g. purchase histories

(curse of dimensionality, everything is a quasi-identifier)

netflix prize privacy breach [10]

68% of records can be uniquely identified
based on rating + date (±3 days) of two movies

mobile phone location data [11]

four random spatio-temporal points are enough to uniquely characterize 95% of the traces amongst 1.5 million users

differential privacy [12]

the risk to one's privacy should not substantially increase as a result of participating in a statistical database

interactive query model instead of releasing data

what is the average age of people voting for ‘B’?

how many people in Potsdam vote for ‘B’?

answer = actual result plus appropriate amount of noise

rephrase machine learning for interactive queries

[13] PCA, k-means, perceptron, ID3 classifiers

there is no one-size-fits-all manual

what data do you want to protect?

what is impact if you fail to protect?

what kind of knowledge / methods could adversary use?

what is the use case of the people using dataset?

is enough utility retained after anonymization?

some red flags [14]

names, addresses, phone numbers

locations (coordinates, references to home/work)

members of small populations

untranslated text, slang

HIPAA Safe Harbour (US, medical data) [15]

at most first three digits of zip (fewer if less than 20k people)
for all dates use only years, group ages 90+
no names, phone numbers, social security numbers etc

(unique for approximately 0.04% of US residents)

SAFE (Germany, census 2011, statistics) [16]

there must always be at least three records for each equivalence group present in the dataset, suppress smaller groups

[1] http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance
[2] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002: 557-570
[3] P. Samarati. Protecting Respondents' Identities in Microdata Release. IEEE Transactions on Knowledge and Data Engineering, 13 (6), 2001: 1010-1027
[4] M. Jändel. Decision support for releasing anonymised data, Computers and security, 46, 2014: 48-61
[5] P. Golle. Revisiting the Uniqueness of Simple Demographics in the US Population. 5th ACM workshop on Privacy in electronic society, 2006: 77-80
[6] A. Machanavajjhala et al. L-diversity: Privacy beyond k-anonymity. ACM Transactions of Knowledge Discovery from Data, 1 (1), 2007
[7] N. Li, T. Li. t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity. 23rd International Conference on Data Engineering, 2007
[7] F. Prasser et al. ARX - A Comprehensive Tool for Anonymizing Biomedical Data, AMIA Annu Symp Proc. 2014: 984–993. http://arx.deidentifier.org/
[8] C. Aggarwal. On k-anonymity and the curse of dimensionality. 31st international conference on Very large data bases, 2005: 901 - 909
[9] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy, 2008: 111-125
[10] YA. de Montjoye. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports 3, 2013
[11] C. Dwork. Differential Privacy. International Colloquium on Automata, Languages and Programming, 2006: 1-12
[12] A. Blum et al. Practical privacy: the SuLQ framework. ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2005: 128-138
[13] https://responsibledata.io/summary-of-our-discussion-on-the-risks-and-mitigations-of-releasing-data/ (in video)
[14] http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
[15] https://www.statistik-berlin-brandenburg.de/zensus/themenblaetter/07_Geheimhaltungsverfahren_SAFE.pdf

insensitive for you $\neq$ insensitive for everybody else

best method dependent on your use case and threat model

there is a tradeoff between anonymity and utility

many methods break down for high-dimensional data

get expert help if necessary