Imperial College
London

Computational Privacy

The limits of anonymization and the future of privacy

by Ali Farzanehfar

Imperial College London

COMPUTATIONAL PRIVACY GROUP

2019.06.26

Shop	customerID	date	amount
Urban Outfitters	7abc1a23	09/23	$97.30
Market Basket	7abc1a23	09/23	$15.13
Whole Food	3092fc10	09/23	$43.78
Central Bakkery	7abc1a23	09/23	$4.33
MIT RecSport	4c7af72a	09/23	$12.29
Flour Cafe	89c0829c	09/24	$3.66
Border Cafe	7abc1a23	09/24	$35.81

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	www.newyorklawyersite.com	2006-03-18 08:03:09		
142	westchester.gov	2006-03-20 03:55:57	1	http://www.westchestergov.com
1326	budget truck rental	2006-03-24 18:27:07		
1326	holiday mansion houseboat	2006-03-29 17:14:01	5	http://www.everyboat.com
1326	back to the future	2006-04-01 17:59:28		

user1	action	direction	user2	timestamp	antID	lat	long
H6ycJQIv	call	in	sW4aFX	2014-03-02 07:13:30	210	42.366944	-71.083611
H6ycJQIv	call	out	5f0jX5G	2014-03-02 07:53:30	34	42.366944	-71.083611
H6ycJQIv	text	in	5f0jX5G	2014-03-02 08:22:30	1809	42.386722	-71.138778

Data is useful but sensitive

Berkeley SCHOOL OF INFORMATION

datascience@berkeley

About Admissions

Online Data Science Degree / Blog

How Cupid is Counting on Data... Find the Perfect Match

February 08, 2016 by datascience@berkeley Staff

DATAFLOQ

VENDORS ARTICLES JOBS EVENTS SERVICES ABOUT US

How Big Data Enabled Spotify To Change The Music Industry

Intelligent Business Performance

Sections

Forbes

47,635 views | Jan 20, 2016, 02:31am

How Big Data Is Disrupting Law Firms And The Legal Profession

Bernard Marr Contributor

There's a ton of information out th... put it to work.

Report foreca...

How the NFL uses Big Data in practice

Shop	customerID	date	amount
Urban Outfitters	7abc1a23	09/23	$97.30
Market Basket	7abc1a23	09/23	$15.13
Whole Food	3092fc10	09/23	$43.78
Central Bakkery	7abc1a23	09/23	$4.33
MIT RecSport	4c7af72a	09/23	$12.29
Flour Cafe	89c0829c	09/24	$3.66
Border Cafe	7abc1a23	09/24	$35.81

Income & location

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	www.newyorklawyersite.com	2006-03-18 08:03:09		
142	westchester.gov	2006-03-20 03:55:57	1	http://www.westchestergov.com
1326	budget truck rental	2006-03-24 18:37:02		
1326	holiday mansion houseboat	2006-03-29 17:14:01	5	http://www.everyboat.com
1326	back to the future	2006-04-01 17:59:28		

Interests & beliefs

user1	action	direction	user2	timestamp	antID	lat	long
H6ycJQIv	call	in	sW4aFX	2014-03-02 07:13:30	210	42.366944	-71.083611
H6ycJQIv	call	out	5f0jX5G	2014-03-02 07:33:30	34	42.366944	-71.083611
H6ycJQIv	text	in	5f0jX5G	2014-03-02 08:22:30	1809	42.386722	-71.138778

Location & social network

Anonymization:
The standard tool for protecting privacy

Example: yearly income of the rich

Name	DOB	Gender	Income [$/yr]
Katerine Enter	01/1936	F	100,000
Luella Perret	04/1960	F	35,678
Dong Rice	12/1982	M	45,000
Carl Stiner	03/1982	M	325,000
Ken Alamo	05/1988	M	125,000
Yulanda Parikh	11/1960	F	23,459
Janee Lundell	09/1935	F	75,008

Example: yearly income of the rich

Name	DOB	Gender	Income [$/yr]
vF0m6JGQ	01/1936	F	100,000
p0nYRG91	04/1960	F	35,678
LgRLdjaA	12/1982	M	45,000
uH4sUWLU	03/1982	M	325,000
zfyv9PRY	05/1988	M	125,000
qbu8Us1P	11/1960	F	23,459
SrQ4sonIn	09/1935	F	75,008

Example: yearly income of the rich

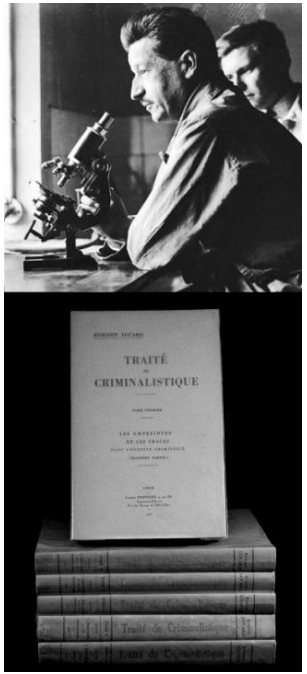
Name	DOB	Gender	Income [$/yr]
vF0m6JGQ	1930	F	100,000
p0nYRG91	1960	F	35,678
LgRLdjaA	1980	M	45,000
uH4sUWLU	1980	M	325,000
zfyv9PRY	1980	M	125,000
qbu8Us1P	1960	F	23,459
SrQ4sonIn	1930	F	75,008

Data protection regulation does not apply to anonymous data

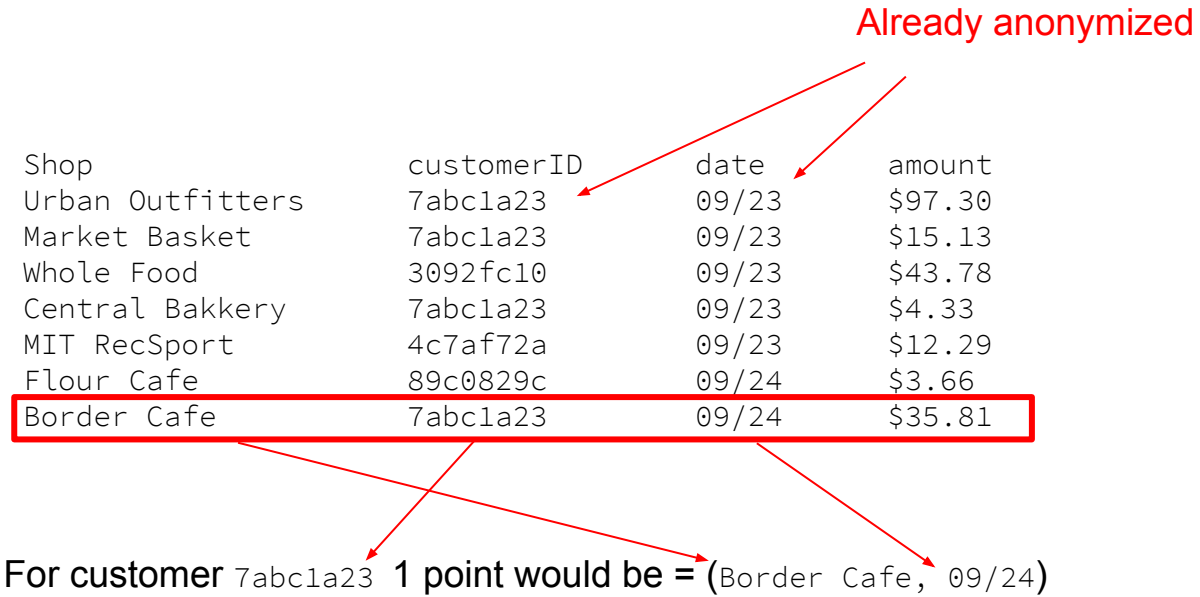
Limits of anonymization

Fingerprints: 12 points are needed to identify you

- Fingerprints are natural identifiers
- To identify someone 12 points are required
- “Points” are distances between ridges
- Parallel to “points” in modern high dimensional data?



Example: Points in credit card data = (shop, date)

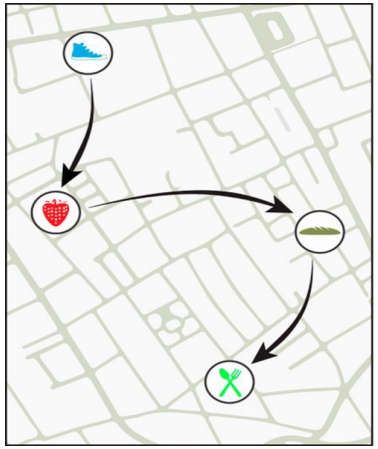


How many points are needed to uniquely identify a person in a big location data set?

4 points: 90% of individuals are uniquely identifiable

$$\varepsilon_4 = .90$$

- Credit card data from an OECD* country
- Data containing histories of 1.1M people
- Collected over 3 months
- Points = (shop, date)
- With 4 (randomly picked) points, 90% of traces are uniquely identifiable
- Then the **whole** trace is available
- Study[△] performed on anonymized data



shop	user_id	time	price
	7abc1a23	09/23	$97.30
	7abc1a23	09/23	$15.13
	3092fc10	09/23	$43.78
	7abc1a23	09/23	$4.33
	4c7af72a	09/23	$12.29
	89c0829c	09/24	$3.66
	7abc1a23	09/24	$35.81

So where can we find these 4 points?

* The Organisation for Economic Co-operation and Development

△ de Montjoye Y.-A., Radaelli L., Singh V. K., Pentland A. S., Unique in the shopping mall: On the reidentifiability of credit card metadata. Science 347 (6221), 536-539. (2015).

Auxiliary information (points) are publicly available

- We leave these points online constantly
- In a targeted attack you might know some information already
  (e.g. place of work / home)
- You could obtain a few points through more traditional means (e.g. by following people)

List of previous successful re-identification instances

- **Anonymous movie ratings**: Narayanan, A., Shmatikov, V., 2008. Robust De-anonymization of Large Sparse Datasets. IEEE, pp. 111–125. https://doi.org/10.1109/SP.2008.33
- **Anonymous apps on our phones**: Achara, J.P., Acs, G., Castelluccia, C., 2015. On the Unicity of Smartphone Applications. ACM Press, pp. 27–36. https://doi.org/10.1145/2808138.2808146
- **Anonymous location data**
  - **From credit cards**: de Montjoye, Y.-A., Radaelli, L., Singh, V.K., Pentland, A.S., 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. Science 347, 536–539. https://doi.org/10.1126/science.1256297
  - **From mobile phones**: de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports 3. https://doi.org/10.1038/srep01376
  - **From public transport**: Lavrenovs, A., Podins, K., 2016. Privacy violations in Riga open data public transport system. IEEE, pp. 1–6. https://doi.org/10.1109/AIEEE.2016.7821808
  - **From GPS**: Naini, F.M., Unnikrishnan, J., Thiran, P., Vetterli, M., 2016. Where You Are Is Who You Are: User Identification by Matching Statistics. IEEE Transactions on Information Forensics and Security 11, 358–372. https://doi.org/10.1109/TIFS.2015.2498131
  - **From taxi rides**: Pandurangan, V., 2014. On Taxis and rainbows.
- **Anonymous medical data**: Sweeney, L., 2002. K-Anonymity:A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 557–570. https://doi.org/10.1142/S0218488502001648
- **Many more . . .**

Anonymization does not always work for privacy

- New data sets are often high dimensional (thousands of points per person)
- This often means that each person is very unique in the data set
- By knowing only a few points, a person can become uniquely identifiable
- **Anonymization is less and less effective against this type of attack**

What can be done with re-identified data

Sensitive attributes: Discovery from anonymous data

- **Predicting personality traits** (e.g. extraversion, openness, etc.) from mobile phone data
  - de Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. S. (2013). Predicting personality using novel mobile phone-based metrics. In Social Computing, Behavioral-Cultural Modeling and Prediction (pp. 48-55). Springer

- **Gender** (78% accuracy) and **age** (60% accuracy) inferred from mobile phone metadata
  - Felbo, B., Sundsøy, P., Pentland, A. 'Sandy,' Lehmann, S., Montjoye, Y.-A. de, 2017. Modeling the Temporal Nature of Human Behavior for Demographics Prediction, in: Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science. Presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Cham, pp. 140–152. https://doi.org/10.1007/978-3-319-71273-4_12

- Predicting **income** levels (0.81 AUC) from mobile phone metadata
  - Blumenstock, J., Cadamuro, G., On, R., 2015. Predicting poverty and wealth from mobile phone metadata. Science 350, 1073–1076. https://doi.org/10.1126/science.aac4420

- Discovery of **political beliefs** of people from their Netflix history
  - Narayanan, A., Shmatikov, V., 2008. Robust De-anonymization of Large Sparse Datasets. IEEE, pp. 111–125. https://doi.org/10.1109/SP.2008.33

- **Lawsuit against Netflix by in-the-closet lesbian mother for fear of outing (Netflix settled)**
  - NetFlix Cancels Recommendation Contest After Privacy Lawsuit [WWW Document], n.d. . WIRED. URL https://www.wired.com/2010/03/netflix-cancels-contest/ (accessed 6.27.18).

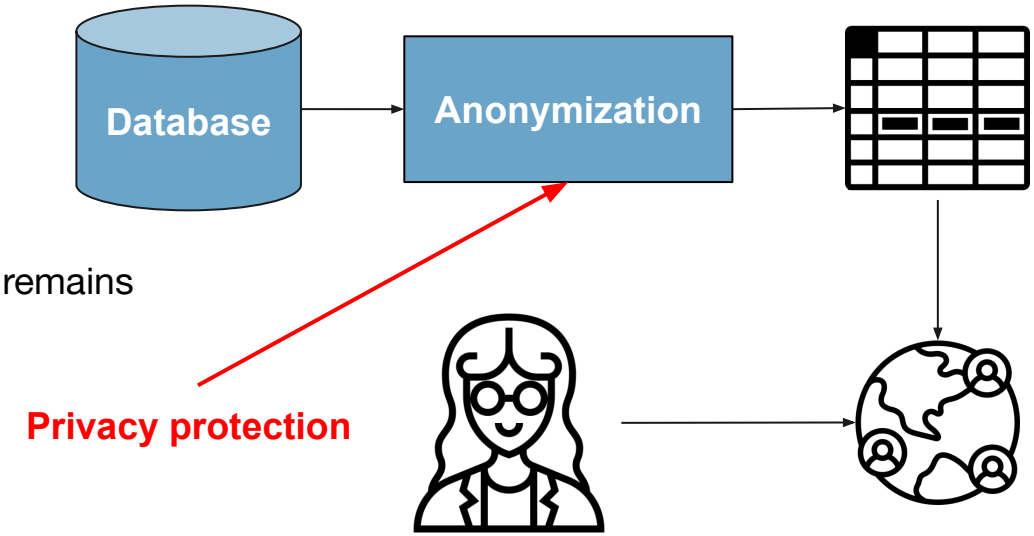
- Many more . . .

The solution:

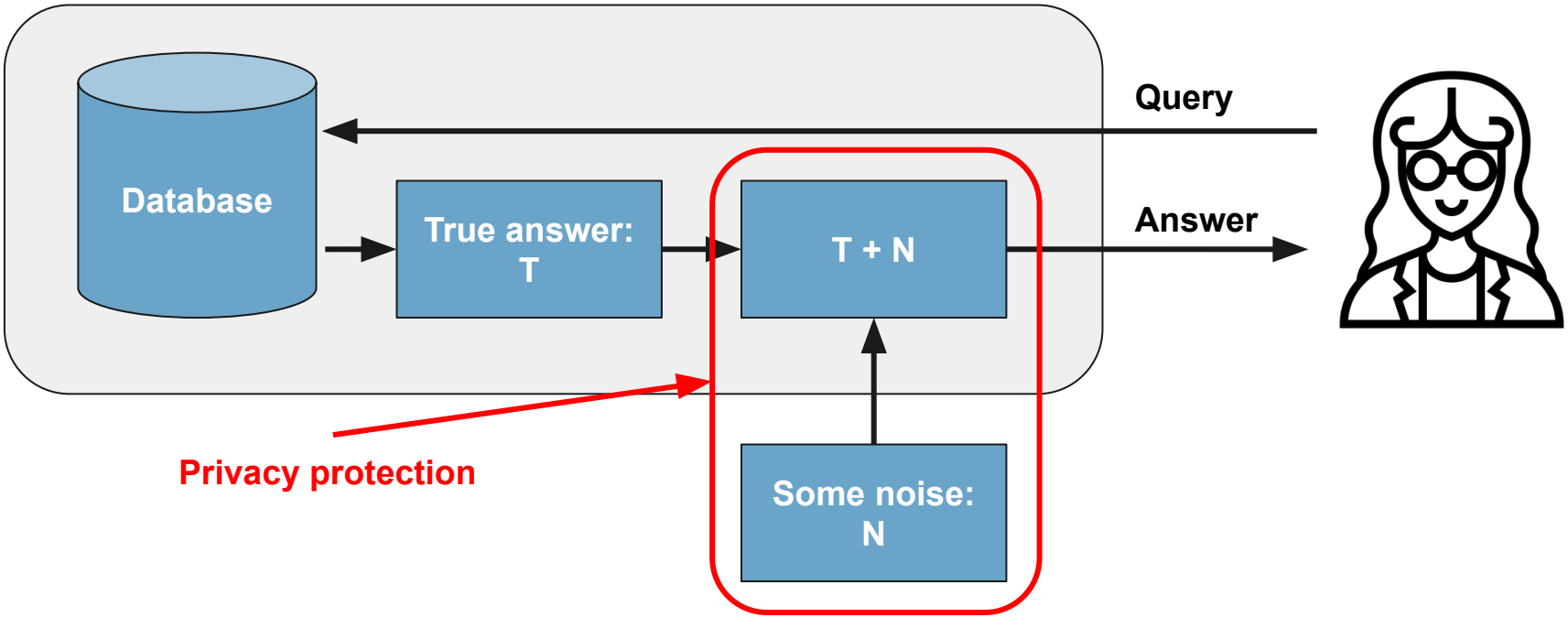
Privacy-Enhancing Technologies (PETs)

Publishing a data set is forever: we cannot *unpublish* data

- In the past data is “anonymized”
- Data is shared (usually online where it remains forever)
- Even if today this is safe, new machine learning techniques tomorrow could change things



A solution: query based systems



OPAL

Imperial College London

mit media lab

orange™

WORLD ECONOMIC FORUM

What is OPAL?

- OPAL (Open Algorithms) is a query based system
- Currently deployed in Senegal
- Uses location data of close to 10M people
- Can be used for many good applications (e.g. national statistics)

Privacy of OPAL

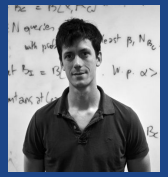
- Set of queries is limited
- The queries are designed to be privacy preserving
- Queries are logged
- The code is open source
- **Many other protective layers …**

Key takeaways

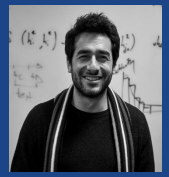
- Data protection regulation does not apply to anonymized data
- Anonymization is ineffective for modern big data sets -> rethink policy
- Privacy enhancing technologies are the future of privacy protection -> invest

Thanks!

Yves-Alexandre de Montjoye
Head of group

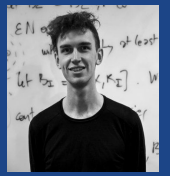


Thibaut Leinart
Postdoc

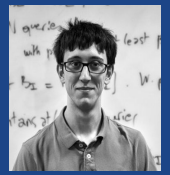


Ali Farzanehfar
PhD student

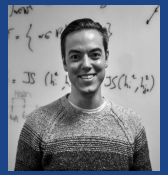
Andrea Gadotti
PhD student



Luc Rocher
PhD student



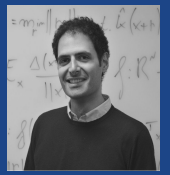
Florimond Houssiau
PhD student



Arnaud Tournier
PhD student

Ana-Maria Cretu
PhD student

Shubham Jain
PhD student



Stefano Marronne
Visiting PhD student

Ali Farzanehfar - ali.farzanehfar@imperial.ac.uk

cpg.doc.ic.ac.uk

COMPUTATIONAL PRIVACY GROUP