



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE



Information Security and Privacy (COM-402)

Part 4: Privacy enhancing technologies

Carmela Troncoso

SPRING Lab

carmela.troncoso@epfl.ch

PETs for Data anonymization

Scenario:

You have a set of data that contains personal data and you would like to anonymize it to:

- not be subject to data protection while processing
- make it public for profit
- make it public for researchers

Goal:

Produce a dataset that **preserves the utility** of the original dataset **without leaking information** about individuals

PETs for Data anonymization

Scenario:

You have a set of data that contains personal data and you would like to anonymize it to:

- not be subject to data protection while processing
- make it public for profit
- make it public for researchers

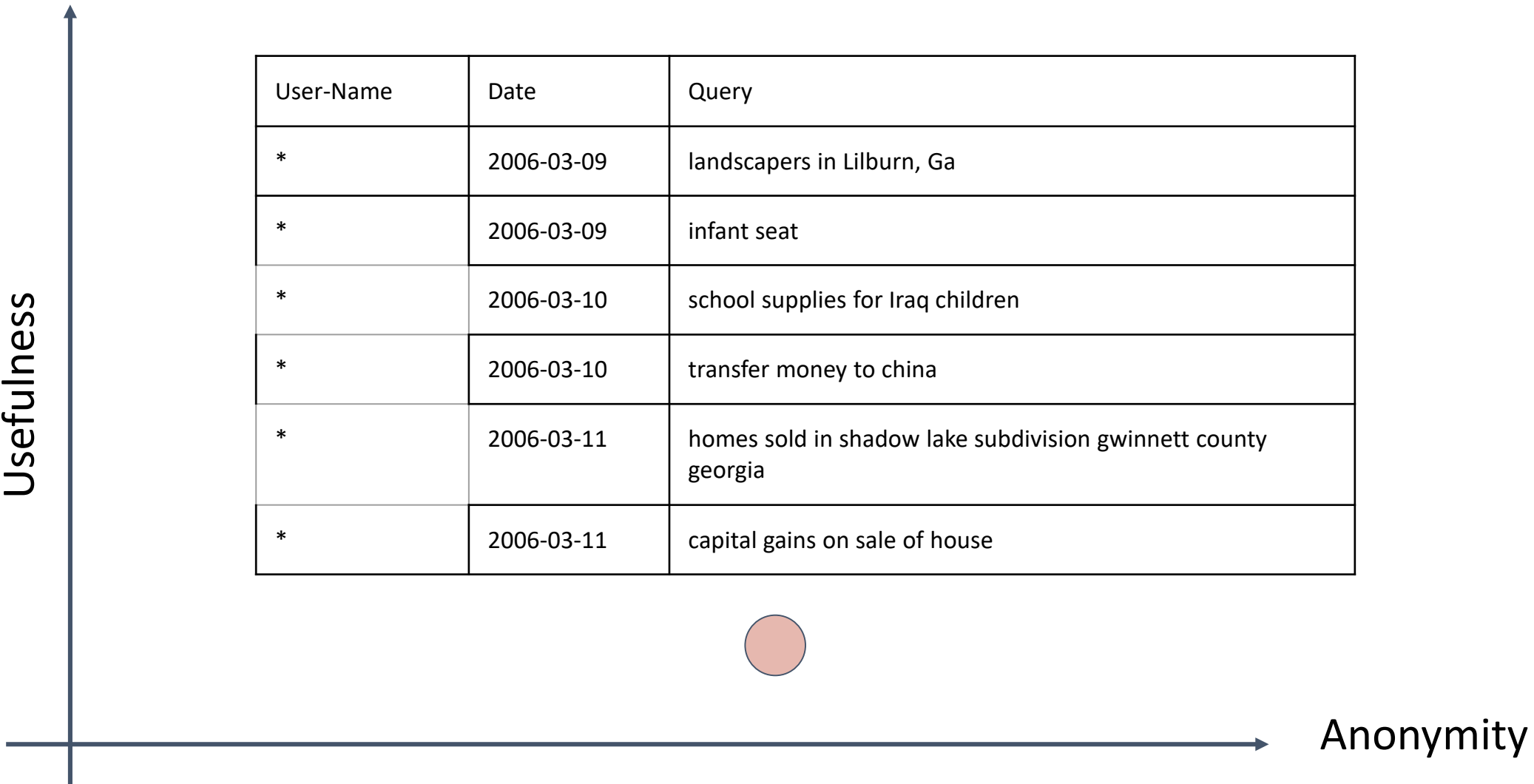
Goal:

Produce a dataset that **preserves the utility** of the original dataset **without leaking information** about individuals

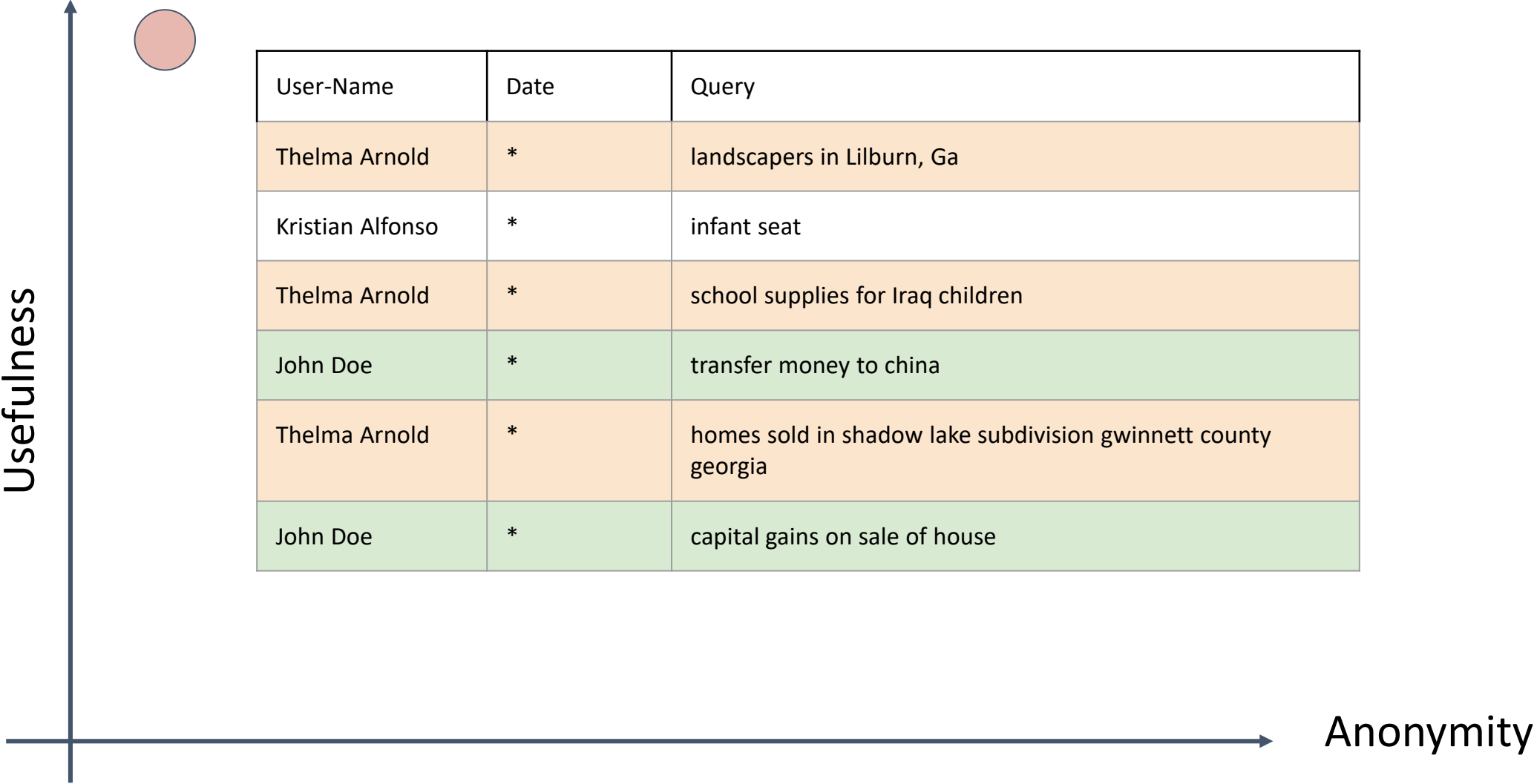
Part of Original AOL Dataset

User-ID	Date	Query
Thelma Arnold	2006-03-09	landscapers in Lilburn, Ga
Kristian Alfonso	2006-03-09	infant seat
Thelma Arnold	2006-03-10	school supplies for Iraq children
John Doe	2006-03-10	transfer money to china
Thelma Arnold	2006-03-11	homes sold in shadow lake subdivision gwinnett county georgia
John Doe	2006-03-11	capital gains on sale of house

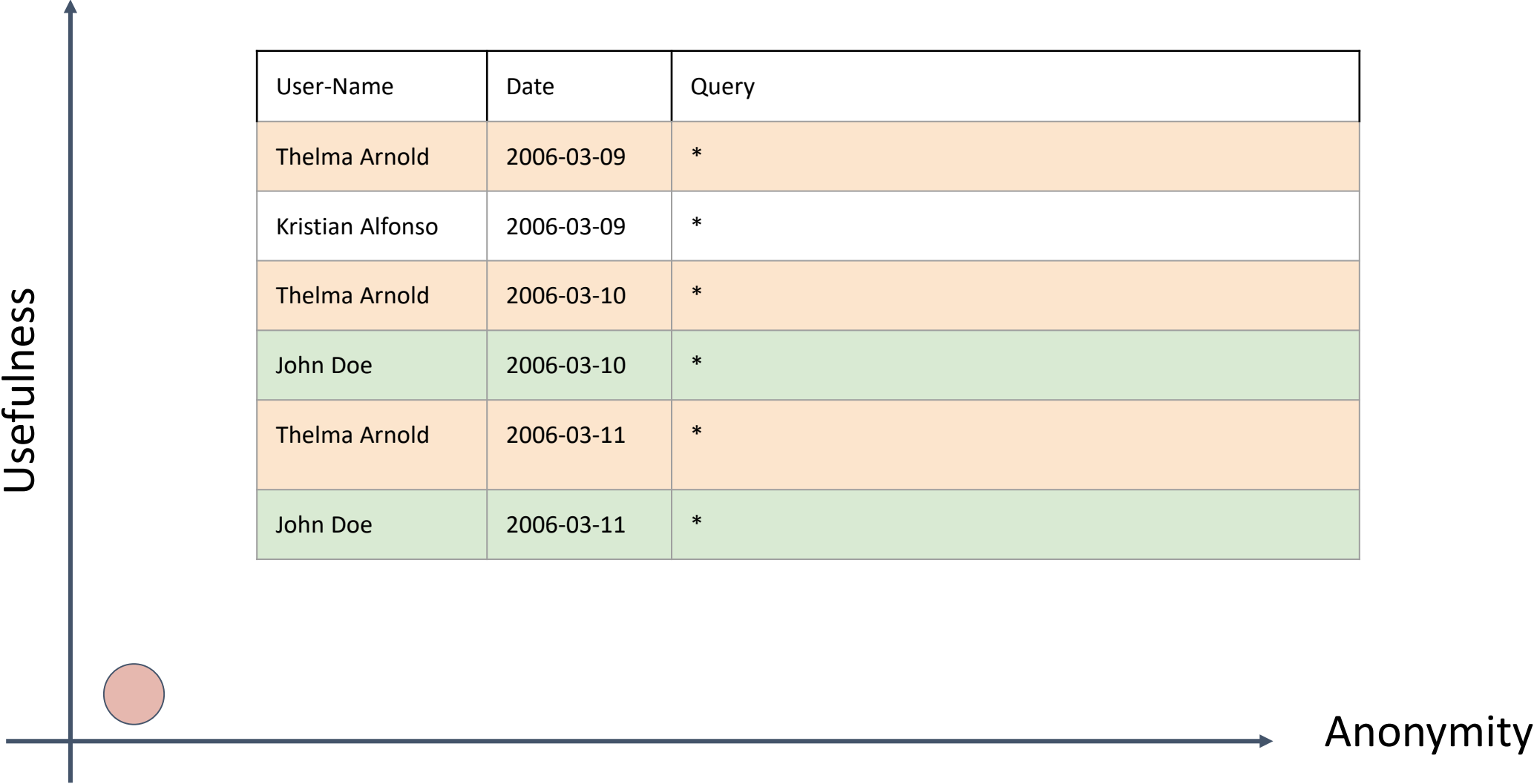
Remove User-names



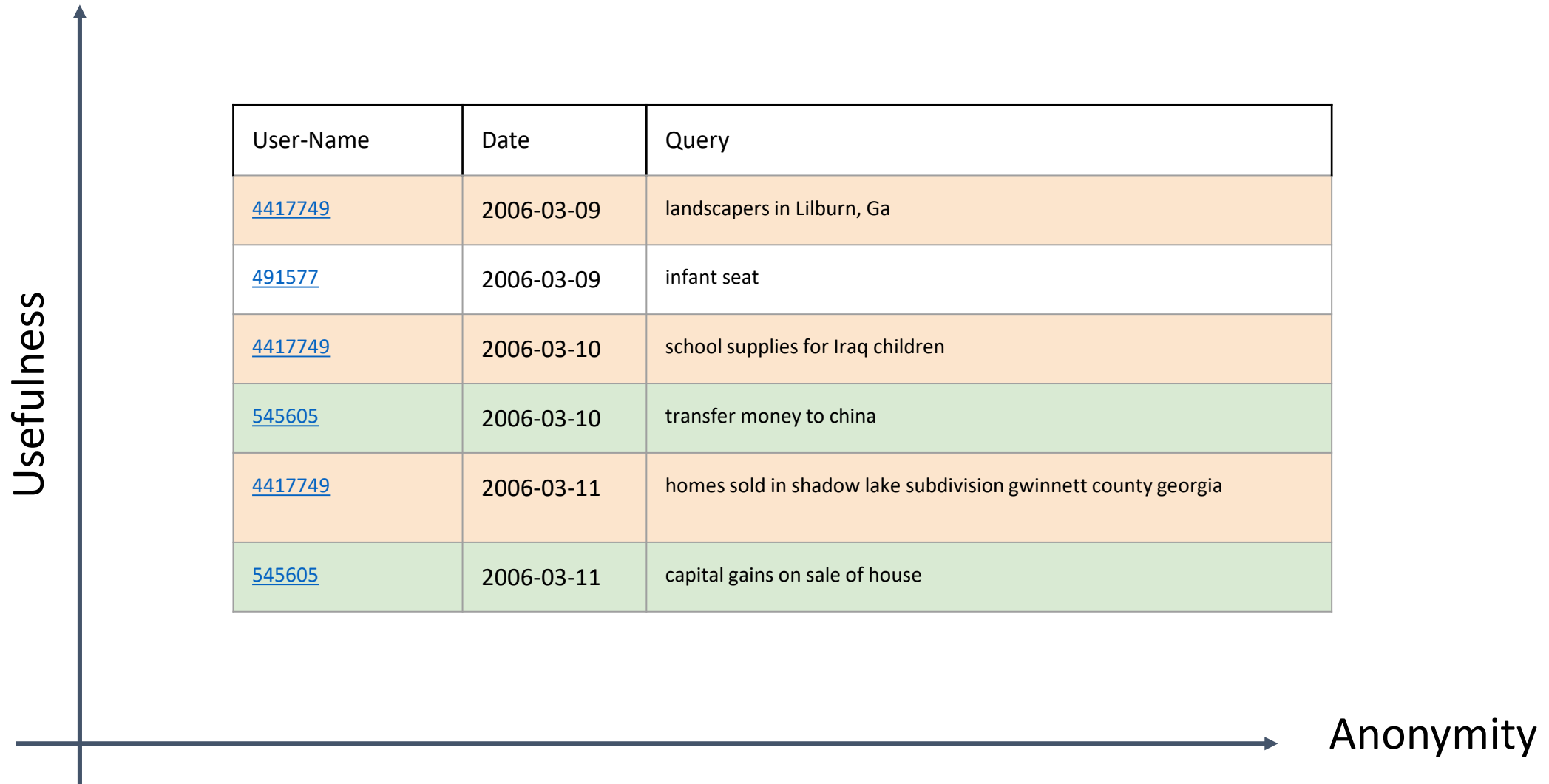
Remove Date



Remove Query



Replace User-Names with Random ID



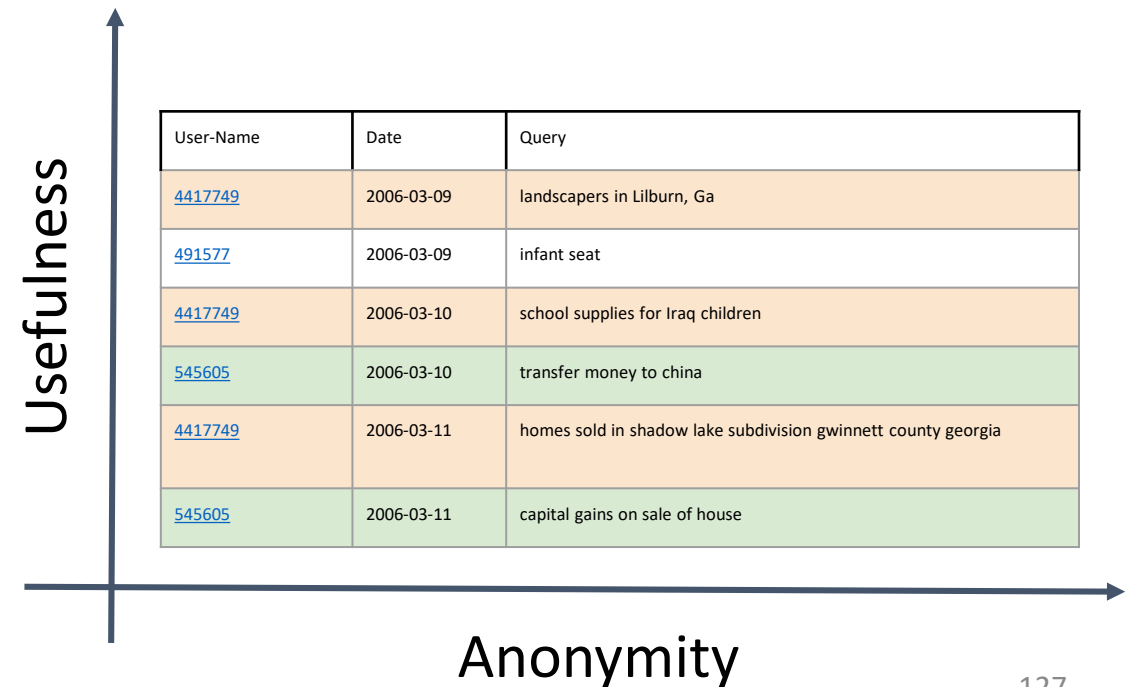
User-Name	Date	Query
4417749	2006-03-09	landscapers in Lilburn, Ga
491577	2006-03-09	infant seat
4417749	2006-03-10	school supplies for Iraq children
545605	2006-03-10	transfer money to china
4417749	2006-03-11	homes sold in shadow lake subdivision gwinnett county georgia
545605	2006-03-11	capital gains on sale of house



Is this enough to say the dataset is anonymized?

A. Yes

B. No



AOL Proudly Releases Massive Amounts of Private Data

Posted Aug 6, 2006 by [Michael Arrington](#) (@arrington)



Yet Another Update: AOL: "This was a screw up"

Further Update: Sometime after 7 pm the download link went down as well, but there is at least one [mirror site](#). AOL is in damage control mode – the fact that they took the data down shows that someone there had the sense to realize how destructive this was, but it is also an admission of wrongdoing of sorts. Either way, the data is now out there for anyone that wants to use (or abuse) it.

Update: Sometime around 7 pm PST on Sunday, the [AOL site](#) referred to below was taken down. The direct link to the data is still live. A cached copy of the page is [here](#).

AOL must have missed the [uproar](#) over the DOJ's demand for "anonymized" search data last year that caused all sorts of pain for Microsoft and Google. That's the only way to explain their [release of data](#) that includes 20 million web queries from 650,000 AOL users.

The data includes all searches from those users for a three month period this year, as well as whether they clicked on a result, what that result was and where it appeared on the result page. It's a 439 MB compressed download, expanded to just over 2 gigs. The data is available [here](#) (this link is directly to the file) and the output is in ten text files, tab delineated.

The utter stupidity of this is staggering. AOL has released very private data about its users without their permission. While the AOL username has been changed to a random ID number, the ability to analyze all searches by a single user will often lead people to easily determine who the user is, and what they are up to. The data includes personal names, addresses, social security numbers and everything else someone might type into a search box.

Crunchbase

AOL

FOUNDED
1985

OVERVIEW

AOL Lifestream is a web-based application that enables users to keep track of all their comments on social networking sites. Integrated with AIM Express, AIM 7, and AIM for Mac, users can publish their statuses, reply to comments on networking sites from their Lifestream tab, and more. AOL Lifestream is a product of [AOL]
(<https://www.crunchbase.com/organization/aol#/entity>)

LOCATION
New York, NY

CATEGORIES

Digital Media, Advertising Platforms, Content Creators, News

WEBSITE
<http://www.aol.com>

[Full profile for AOL](#)

NEWSLETTER SUBSCRIPTIONS

☐ **The Daily Crunch**

Get the top tech stories of the day delivered

Naive anonymization can easily be broken

In 2006, AOL released search queries of 500,000 pseudonymous users

- Days later, New York Times reveals the identity of user 4417749
- CTO and researchers responsible of sharing data fired

The same year, Netflix revealed over 100 million movie ratings made by 500,000 users after removing personal details

- Researchers de-anonymize dataset by comparing it with publicly available ratings on Internet Movie Database (IMDB)



[Credit: New York Times]

Re-identification by Linking

Medical Data

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Re-identification by Linking

Medical Data

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Takeaways

Naïve anonymization does NOT work



Attributes themselves are a pseudo-identifier



Pseudo-identifiers allow to link databases



Linking allows to break anonymity