## **Data privacy**

Data Science in a Box datasciencebox.org



## Case study: AOL search data leak



### The New York Times

# A Face Is Exposed for AOL Searcher No. 4417749

Ms. [Thelma] Arnold, who agreed to discuss her searches with a reporter, said she was shocked to hear that AOL had saved and published three months' worth of them. "My goodness, it's my whole personal life," she said. "I had no idea somebody was looking over my shoulder."

In the privacy of her four-bedroom home, Ms. Arnold searched for the answers to scores of life's questions, big and small. How could she buy "school supplies for Iraq children"? What is the "safest place to live"? What is "the best season to visit Italy"?

Michael Barbaro and Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times. 9 August 2006.

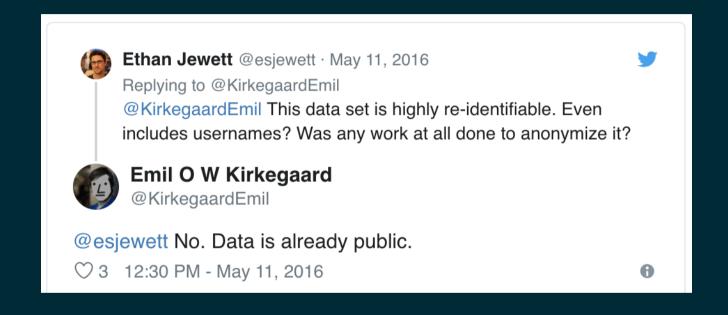
# Case study: OK Cupid

## **OK Cupid data breach**

- In 2016, researchers published data of 70,000 OkCupid users—including usernames, political leanings, drug usage, and intimate sexual details
- Researchers didn't release the real names and pictures of OKCupid users, but their identities could easily be uncovered from the details provided, e.g. usernames

Some may object to the ethics of gathering and releasing this data. However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form.

Researchers Emil Kirkegaard and Julius Daugbjerg Bjerrekær In analysis of data that individuals willingly shared publicly on a given platform (e.g. social media), how do you make sure you don't violate reasonable expectations of privacy?



## Case study: Facebook & Cambridge Analytica

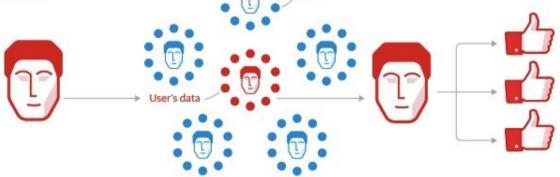
#### Cambridge Analytica: how 50m Facebook records were hijacked

Approx. 320,000 US voters ('seeders') were paid \$2-5 to take a detailed personality/ political test that required them to log in with their Facebook account

The app also collected data such as likes and personal information from the testtaker's Facebook account ...

The personality quiz results were paired with their Facebook data - such as likes - to seek out psychological patterns

Algorithms combined the data with other sources such as voter records to create a superior set of records (initially 2m people in 11 key states\*), with hundreds of data points per person



... as well their friends' data, amounting to over 50m people's

raw Facebook data

Friends'

These individuals could then be targeted with highly personalised advertising based on their personality data

Guardian graphic. \*Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

Carole Cadwalladr and Emma Graham-Harrison. How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool. The Guardian. 17 March 2018.