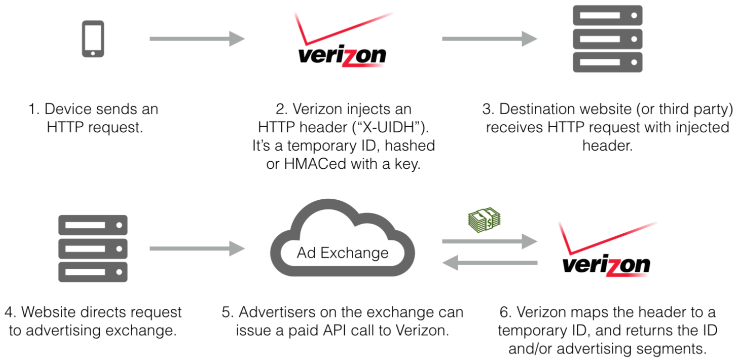
Handout 7 (Privacy)

The first motor car was invented around 1886. For ten years, until 1896, the law in the UK and elsewhere required a person to walk in front of any moving car waving a red flag. Cars were such a novelty that most people did not know what to make of them. The person with the red flag was intended to warn the public, for example horse owners, about the impending novelty—a car. In my humble opinion, we are at the same stage of development with privacy. Nobody really knows what it is about or what it is good for. All seems very hazy. There are a few laws (cookie law, right-to-be-forgotten) which address problems with privacy, but even if they are well intentioned, they either backfire or are already obsolete because of newer technologies. The result is that the world of "privacy" looks a little bit like the old Wild West.

For example, UCAS, a charity set up to help students to apply to universities, has a commercial unit that happily sells your email addresses to anybody who forks out enough money in order to be able to bombard you with spam. Yes, you can opt out very often in such "schemes", but in case of UCAS any opt-out will limit also legit emails you might actually be interested in.[1]

Another example: Verizon, an ISP who is supposed to provide you just with connectivity, has found a "nice" side-business too: When you have enabled all privacy guards in your browser (the few you have at your disposal) Verizon happily adds a kind of cookie to your HTTP-requests.[2] As shown in the picture below, this cookie will be sent to every web-site you visit. The web-sites then can forward the cookie to advertisers who in turn pay Verizon to tell them everything they want to know about the person who just made this request, that is you.


verizon
1. Device sends an HTTP request.
2. Verizon injects an HTTP header ("X-UIDH"). It's a temporary ID, hashed or HMACed with a key.
3. Destination website (or third party) receives HTTP request with injected header.
Ad Exchange
verizon
4. Website directs request to advertising exchange.
5. Advertisers on the exchange can issue a paid API call to Verizon.
6. Verizon maps the header to a temporary ID, and returns the ID and/or advertising segments.

[1] The main objectionable point, in my opinion, is that the *charity* everybody has to use for HE applications has actually very honourable goals (e.g. assist applicants in gaining access to universities), but in their small print (or better under the link "About us") reveals they set up their organisation so that they can also shamelessly sell email addresses they "harvest". Everything is of course very legal…moral?…well that is in the eye of the beholder. See:
http://www.ucas.com/about-us/inside-ucas/advertising-opportunities or http://www.theguardian.com/uk-news/2014/mar/12/ucas-sells-marketing-access-student-data-advertisers

[2] http://webpolicy.org/2014/10/24/how-verizons-advertising-header-works/

How disgusting? Even worse, Verizon is not known for being the cheapest ISP on the planet (completely the contrary), and also not known for providing the fastest possible speeds, but rather for being among the few ISPs in the US with a quasi-monopolistic "market distribution".

Well, we could go on and on…and that has not even started us yet with all the naughty things NSA & Friends are up to. Why does privacy actually matter? Nobody, I think, has a conclusive answer to this question yet. Maybe the following four notions help with clarifying the overall picture somewhat:

- **Secrecy** is the mechanism used to limit the number of principals with access to information (e.g., cryptography or access controls). For example I better keep my password secret, otherwise people from the wrong side of the law might impersonate me.

- **Confidentiality** is the obligation to protect the secrets of other people or organisations (secrecy for the benefit of an organisation). For example as a staff member at King's I have access to data, even private data, I am allowed to use in my work but not allowed to disclose to anyone else.

- **Anonymity** is the ability to leave no evidence of an activity (e.g., sharing a secret). This is not equal with privacy—anonymity is required in many circumstances, for example for whistle-blowers, voting, exam marking and so on.

- **Privacy** is the ability or right to protect your personal secrets (secrecy for the benefit of an individual). For example, in a job interview, I might not like to disclose that I am pregnant, if I were a woman, or that I am a father. Similarly, I might not like to disclose my location data, because thieves might break into my house if they know I am away at work. Privacy is essentially everything which 'shouldn't be anybody's business'.

While this might provide us with some rough definitions, the problem with privacy is that it is an extremely fine line what should stay private and what should not. For example, since I am working in academia, I am every so often very happy to be a digital exhibitionist: I am very happy to disclose all 'trivia' related to my work on my personal web-page. This is a kind of bragging that is normal in academia (at least in the field of CS), even expected if you look for a job. I am even happy that Google maintains a profile about all my academic papers and their citations.

On the other hand I would be very irritated if anybody I do not know had a too close look on my private live—it shouldn't be anybody's business. The reason is that knowledge about my private life usually is used against me. As mentioned above, public location data might mean I get robbed. If supermarkets build a profile of my shopping habits, they will use it to *their* advantage—surely not to *my* advantage. Also whatever might be collected about my life will always be an incomplete, or even misleading, picture—for example I am sure my creditworthiness score was temporarily(?) destroyed by not having a regular income in this country (before coming to King's I worked in Munich for five years). To correct such incomplete or flawed credit history data there is, since recently, a law that allows you to check what information is held about you for determining your creditworthiness. But this concerns only a very small part of the data that is held about me/you.

To see how private matter can lead really to the wrong conclusions, take the example of Stephen Hawking: When he was diagnosed with his disease, he was given a life expectancy of two years. If employers would know about such problems, would they have employed Hawking? Now, he is enjoying his 70+ birthday. Clearly personal medical data needs to stay private. A movie which has this topic as its main focus is Gattaca from 1997.[3]

To cut a long story short, I let you ponder about the two statements that often voiced in discussions about privacy:

- *"You have zero privacy anyway. Get over it."*

  by Scott Mcnealy (CEO of Sun)

- *"If you have nothing to hide, you have nothing to fear."*

If you want to read up further on this topic, I can recommend the following article that appeared in 2011 in the Chronicle of Higher Education

http://chronicle.com/article/Why-Privacy-Matters-Even-if/127461/

Funnily, or maybe not so funnily, the author of this article carefully tries to construct an argument that does not only attack the nothing-to-hide statement in cases where governments & Co collect people's deepest secrets, or pictures of people's naked bodies, but an argument that applies also in cases where governments "only" collect data relevant to, say, preventing terrorism. The fun is of course that in 2011 we could just not imagine that respected governments would do such infantile things as intercepting people's nude photos. Well, since Snowden we know some people at the NSA did exactly that and then shared such photos among colleagues as "fringe benefit".

Re-Identification Attacks

Apart from philosophical musings, there are fortunately also some real technical problems with privacy. The problem I want to focus on in this handout is how to safely disclose datasets containing very potentially private data, say health data. What can go wrong with such disclosures can be illustrated with four well-known examples:

- In 2006, a then young company called Netflix offered a 1 Mio $ prize to anybody who could improve their movie rating algorithm. For this they disclosed a dataset containing 10% of all Netflix users at the time

[3] http://www.imdb.com/title/tt0119177/

(appr. 500K). They removed names, but included numerical ratings of movies as well as times of ratings. Though some information was perturbed (i.e., slightly modified).

  Two researchers had a closer look at this anonymised data and compared it with public data available from the International Movie Database (IMDb). They found that 98% of the entries could be re-identified in the Netflix dataset: either by their ratings or by the dates the ratings were uploaded. The result was a class-action suit against Netflix, which was only recently resolved involving a lot of money.

- In the 1990ies, medical datasets were often made public for research purposes. This was done in anonymised form with names removed, but birth dates, gender and ZIP-code were retained. In one case where such data about hospital visits of state employees in Massachusetts was made public, the then governor assured the public that the released dataset protected patient privacy by deleting identifiers.

  A graduate student could not resist cross-referencing public voter data with the released data including birth dates, gender and ZIP-code. The result was that she could send the governor his own hospital record. It turns out that birth dates, gender and ZIP-code uniquely identify 87% of people in the US. This work resulted in a number of laws prescribing which private data cannot be released in such datasets.

- In 2006, AOL published 20 million Web search queries collected from 650,000 users (names had been deleted). This was again done for research purposes. However, within days an old lady, Thelma Arnold, from Lilburn, Georgia, (11,596 inhabitants) was identified as user No. 4417749 in this dataset. It turned out that search engine queries are deep windows into people's private lives.

- Genome-Wide Association Studies (GWAS) was a public database of gene-frequency studies linked to diseases. It would essentially record that people who have a disease, say diabetes, have also these genes. In order to maintain privacy, the dataset would only include aggregate information. In case of DNA data this was achieved by mixing the DNA of many individuals (having a disease) into a single solution. Then this mixture was sequenced and included in the dataset. The idea was that the agregate information would still be helpful to researchers, but would protect the DNA data of individuals.

  In 2007 a forensic computer scientist showed that individuals can be still identified. For this he used the DNA data from a comparison group (people from the general public) and "subtracted" this data from the published data. He was left with data that included all "special" DNA-markers of the individuals present in the original mixture. He essentially deleted the "background noise". Now the problem with DNA data is that it is of such a high resolution that even if the mixture contained maybe 100 individuals, you can now detect whether an individual was included in the mixture or not.

This result changed completely how DNA data is nowadays published for research purposes. After the success of the human-genome project with a very open culture of exchanging data, it became much more difficult to anonymise datasuch that patient's privacy is preserved. The public GWAS database was taken offline in 2008.

There are many lessons that can be learned from these examples. One is that when making data public in anonymised form you want to achieve *forward privacy*. This means, no matter of what other data that is also available or will be released later, the data does not compromise an individual's privacy. This principle was violated by the data in the Netflix and governor of Massachusetts cases. There additional data allowed one to re-identify individuals in the dataset. In case of GWAS a new technique of re-identification compromised the privacy of people on the list. The case of the AOL dataset shows clearly how incomplete such data can be: Although the queries uniquely identified the old lady, she also looked up diseases that her friends had, which had nothing to do with her. Any rational analysis of her query data must have concluded, the lady is on her deathbed, while she was actually very much alive and kicking.

Differential Privacy

Differential privacy is one of the few methods, that tries to achieve forward privacy with large datasets. The basic idea is to add appropriate noise, or errors, to any query of the dataset. The intention is to make the result of a query insensitive to individual entries in the database. The hope is that the added error does not eliminate the "signal" one is looking for by querying the dataset.

$$\begin{array}{ccc} \text{User} & \begin{array}{c} \text{tell me } f(x) \Rightarrow \\ \Leftarrow f(x) + \text{noise} \end{array} & \begin{array}{l} \text{Database} \\ x_1, \ldots, x_n \end{array} \end{array}$$