Introduction to privacy

What is privacy?

- Privacy is the RIGHT of an individual to control how information about him/her is collected, stored, and shared
  - e.g., you can decide with whom you share your personal information
- As a concept, it emerged in 1890 due to the spread of sensationalist journalism and photography
  - initially defined as "the right to be let alone."
  - was exacerbated when telephones became widespread around the 1920s
  - government was identified as a potential privacy invader (see Holocaust)
- After the 1970s, new technologies emerged through the appearance of personal computers
  - they created new ways to gather/store/process personal information
- Today, a lot of information on individuals is collected and stored in many databases worldwide (BIG DATA)
  - control of that information by individuals became difficult

Why do companies collect data?

- Ads support the Internet ecosystem
  - You can read many websites for "FREE" just because the site has ads.
- The goal of companies is to build as detailed profile of you as possible
  - What are your interests?
  - Where do you do your shopping?
  - What do you eat, what do you like, when do you get up, etc?
- Why? To give you personalized (targeted) ads
  - If the ad is relevant to you, then you are more likely to buy the product or to be manipulated (see political ads)
- Is it a good business?
  - Online advertising revenue $ 225 billion in the US in 2023.

What can you lose?

- Companies can sell your data
  - to other companies who build profiles
- Companies can infer further sensitive information
  - religion, sexual orientation, financial status, medical status, etc.
- Your personal data can be stolen from the company
  - by a hacker or an employee who sells your data

How much do you share?

- Machine learning correctly infers whether someone is gay or straight
  - 81% of accuracy for men (vs. 61% of human judgement)
  - 74% of accuracy for women (vs. 54% of human judgement)
- Can you also predict other psychological conditions or even personality?
- Photos contain a wealth of personal data that can be used for profiling!

Companies monetize your data

- In 2006, AOL publicly released 20 million search queries for 650.000 users
- For anonymization, they removed the IP addresses of queries...
- User #4417749 was identified by its searches, like:
  - "*xy* sickness"
  - "*xy* years single man"
  - "*xy* animal with *z* sickness"
  - "landscapers in *xy* city"
- Queries of user #17556639
  - "how to kill your wife"
  - "pictures of dead people."
  - "car crash photo"
  - "photo of dead people."

How much do you share?

- Target is a software identified about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score
- More importantly, Target could also estimate her due date to be within a small window
- Target could (and does) send coupons timed to very specific stages of her pregnancy

How much do you share?

Human skin color on photos/videos

- Human skin color varies slightly with blood circulation
- heart rate can be extracted from a video based on the temporal variation of the skin color
  - this is invisible to the human eye

How much do you share?

Hearth rate on fitness wearables

- Fitness monitors reveal more information than most people realize
  - , E.g., when do you have sex?
- It may be possible to infer someone's religious beliefs from their heart rate data
  - Muslims pray at five prescribed times a day. Fitness data may reveal if someone is kneeling
  - Or detect when someone is singing every Sunday morning
  - Jews are inactive on Saturdays.
- These are NOT proofs, but EVIDENCES

Your data can be stolen

- Even if you trust the company, they can have security incidents when your data can be stolen and sold in the black market
- there always be such incidents (insider attacks by employees, attacks sponsored by foreign governments, etc.)

Why is it a problem?

- Sharing personal data can lead to discrimination, personal embarrassment, or damage to one's professional reputation
  - You can be stigmatized based on your religion, political affiliation, sexual orientation, or your physical condition!
- These can have serious consequences
  - from mere annoyance or fear to even death or physical disability!
  - Direct financial disadvantage. Health problems, risks, and dangerous lifestyles can lead to worse insurance or loan conditions.
- People cannot anticipate the future misuse of their data
  - tomorrow, one can develop a new machine-learning model to infer sexual orientation from photos that you share today. or infer your disease from the videos you upload today

Solution?

- Privacy is not only a technical issue; it needs legislation (laws) and law enforcement
- However, technology can also help
  - encryption and access control techniques (e.g., encrypted Cloud storage)
  - anonymous communication techniques (e.g., TOR)
  - Anonymization of corporal datasets
  - in general: Privacy Enhancing Technologies (PETS)

Why are privacy attacks so easy?

1. DATA CORRELATION
   - Almost all published data leak some sensitive information, even if it is not apparent at first sight
     - Pressure, power consumption $\rightarrow$ location
     - Electricity consumption $\rightarrow$ religion
     - Videos $\rightarrow$ heart rate $\rightarrow$ health status
     - List of installed apps $\rightarrow$ Religion, Sexual orientation
     - Facebook likes $\rightarrow$ personality
     - List of watched movies $\rightarrow$ Sexual orientation, Religion
     - Browsing history, search queries $\rightarrow$ Almost everything...
     - .
2. People share too much information about themselves consciously and unconsciously
   - they can forecast the misuse of their personal data!!!

PSYCHOLOGICAL PROFILING

Psychometric profiling

- Personality can be defined by the "Big Five"
  - **OCEAN**: **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism
- Such traits can be predicted pretty well from your Facebook likes:
  - Example: neurotic people tend to like Nirvana, while people who "liked" Lay Gaga are likely to be extroverts.
- How many likes are needed?
  - 70 likes provides almost as accurate estimation of your personality as your friend
  - 150 likes is similar to your parent's prediction
  - 300 likes is like your partner or spouse
- Typical Facebook user lists 227 likes.

Facebook "like prediction

FACEBOOK "LIKES" PREDICTIVE OF PERSONALITY

EXTRAVERSION

OUTGOING & ACTIVE

Snookie
SHOTS! SHOTS! SHOTS! SHOTS! SHOTS! SHOTS! SHOTS! SHOTS! SHOTS! EVERYBODY!!!
partying
Nikki Minaj
Mike The Situation
Lil' Wayne
LeBron James
Gucci
David Guetta
BEER PONG!
Tanning
meeting new people
Parties
Making People Laugh
Tiffany & Co.

SHY & RESERVED

The Matrix
Watching Anime
J-pop
The X-Files
Thinking
Gaming
Doctor Who
Programming
Neon Genesis Evangelion
Star Trek: The Next Generation
Wikipedia
Sonata Arctica
Mathematics
Minecraft
Kamelot
Star Trek: Voyager

CONSCIENTIOUSNESS

WELL ORGANIZED

The Bucket List
United States Navy
No Strings Attached
Kite Runner
Working
The Apprentice
Boxing
R&B
Mountain biking
Cheerleading
Italian Job
Studying
Working Out
Cycling
Motorcycles
The Guardian
Pearl Harbor

SPONTANEOUS

The Velvet Underground
The Pixies
vampires
procrastinating
Gay-Straight Alliances
Daydreaming
The Hitchhiker's Guide to the Galaxy
Being Lazy
I hate it when all other schools near you have a snow day and you don't.
League of Legends
Uncontrollable swearing after stubbing your toe in a dark room
Hunter S. Thompson
The Mighty Boosh
Blackadder
Watchmen
Douglas Adams

Psychometric profiling

- How is it used?
  - Targeted ads
    - Fear advertising is best suited for extroverts and agreeable
    - Conscientious trait individuals are generally more drawn to ads that evoke anger
  - Mass persuasion (Matz, Kosinski, Nave, Stillwell: `https://www.pnas.org/doi/10.1073/pnas.1710966114`)
- The Cambridge Analytica scandal: Trump's 2016 US presidential campaign
  - Collect personal data from 87 million Facebook users (most of which was collected without permission)
  - Used to affect the US election in 2016.
  - Facebook paid 725 million dollars to settle the lawsuits

DE-ANONYMIZATION AND ANONYMIZATION

Anonymization example: Netflix Prize (2010)

- Netflix has offered $\$1,000,000$ for a 10% improvement in its movie recommendation system
- A training set was released which contained the movie ratings of users w/o user IDs
- Researchers successfully linked this dataset with IMDB, where some user profiles are public
- What can they learn? E.g., sexual orientation

Anonymization example: NYC Taxi dataset (2014)

- It contains details about every taxi ride (yellow cabs) in NYC in 2013
  - pickup and drop off times, locations, fare, and tip amounts, as well as anonymized (hashed) versions of the taxi's license and medallion numbers
  - , but there are only about 22 million possible license numbers!!!
- What can we learn? "Jessica Alba got into her taxi outside her hotel, the Trump SoHo, and did not add a tip to her $9 fare."

PRIVACY OF AGGREGATED DATA

Example: Strava

Strava

Strava is a fitness app:

- Details of work-out (distance, speed, GPS trajectory) are recorded of runners/bikers
- GPS trajectories are uploaded to Strava
- Strava publishes a heat map of regions

What is the problem?

Example: Strava

Strava

Strava is a fitness app:

- Details of work-out (distance, speed, GPS trajectory) are recorded of runners/bikers
- GPS trajectories are uploaded to Strava
- Strava publishes a heat map of regions

What is the problem?

US Marines:

- Soldiers use the app like normal people and turn it on when they do exercise
- It reveals places that are hidden on Google Maps: U.S. military bases in Syria and Afghanistan and a Royal Navy base that contains the UK's nuclear arsenal.
- Users can now opt out of having their data aggregated

Other recent privacy concerns

- June 2022: security flow - identity and Tracking of security personnel working at military bases in Israel.
- June 2023: Strava heat map data could be used to identify the home addresses of highly active users in remote areas.
- August 2023: identify a person as the alleged arsonist who lit a Trump supporter's lawn sign on fire.
- July 2024: Runners for hire ("Strava jockeys") in Singapore

INTERDEPENDENT PRIVACY

Genomic privacy

- Genome sequencing is becoming less and less expensive
- 99.5% of the DNA sequences of two random human beings are identical, and this value is even larger for relatives
- Kin-genomic privacy: An individual revealing only his genome may not only damage only his own privacy but also that of his relatives who might not even publish their genome at all
  - an individual does not need consent from his relatives to share his genome with a 3rd party

Example: 23andme

- 23andme offers salvia-based genetic testing
  - one can send a sample of salvia, and 23andme returns ancestry composition and genetic pre-dispositions to certain diseases
  - They also offered a relative finder program, where they matched your DNA with other people's DNA
- One biologist participated in the relative finder program, and he found his paternal half-brother
- As a result, his parents got divorced

Why privacy matters?

- Everybody has something to hide
  - If you don't think so, would you publish your Google search queries? Or your web history?
- Profiling and surveillance can change your behavior
  - You may not search for certain things because
    - you don't want them to affect your future search results, ads, or recommendations
    - you may be monitored by a national agency
    - you don't want to be manipulated (Psychometric profiling)
- Nobody can see the future.
  - Your data can be stolen in the future from your "trusted" company

Why privacy is hard?

- User-tracking/de-anonymization are easy, but anonymization is hard, because
  - People share a lot of data about themselves on a daily basis
  - Data shared by people are often highly correlated with sensitive information
  - Interdepedent-privacy
    - my wife's location can reveal my own location even if my data is not collected
- Solution?
  - "Check your privacy settings!" is NOT a solution
    - Prefer opt-in vs. opt-out consent
  - Promote informed consent by transparency (as much as possible)
    - This is not easy: nobody knows what machine learning will be able to predict tomorrow

Location Tracking

Process of monitoring and recording the geographical location of an individual or a device.

Ways to collect location data

- GPS - satellites, precise location
- Wi-Fi Signals - estimate location by comparing the networks's MAC address with a database of known location
- Cell Tower Triangulation - measuring the strength and timing of signals from nearby cell towers
- Bluetooth Beacons - emit low-energy signals that can be detected by nearby smartphones

These methods enable the collection of location data (also timestamps and additional context about user activity) without the awareness or explicit consent of individuals.

Uses of location data

- Social networking and connectivity
- Tourism and travel
- Geotagging and memories
- Personal safety
- Fitness tracking

Risk of unauthorized access to location data

- Stalking and harassment
- Burglary and theft
- Personal safety
- Employment
- Identity theft
- Surveillance and discrimination
- Location profiling