# Main Insights

## Scenario 1: Sharing Your (Tabular) Data:

* **Anonymization can work:** Although challenging, research data can sometimes be effectively anonymized, for example by applying [k-anonymity](https://en.wikipedia.org/wiki/K-anonymity) and [l-diversity](https://en.wikipedia.org/wiki/L-diversity), which is a relatively simple approach. This can be implemented using the tool “[ARX](https://arx.deidentifier.org/)”, for example.
* **Anonymization has important limitations:** However, the above approach requires strong assumptions about what attributes (columns) can potentially be used to identify an individual, which can lead to the publication of unsafe data if the assumptions are too optimistic. Alternatively, a conservative approach may only allow the publication of data that is of little use. The increasing availability of personal data on commercial data markets means that for an increasing number of attributes there exists a real possibility that they can be used to identify individuals by [matching attributes across data sets](https://arxiv.org/ftp/arxiv/papers/1307/1307.1370.pdf). This makes anonymization increasingly difficult.
* **Aggregated data can be unsafe too:** An alternative approach to publishing anonymized data is to publish only data that is aggregated across individuals. However, this may still communicate sensitive information about individuals, as parts of the private data may be inferable from the published aggregations through a [database reconstruction attack](https://dl.acm.org/doi/pdf/10.1145/3287287). This is a problem especially when one publishes many different aggregations in order to support a wide range of use cases.
* **Making your data publicly accessible may not be necessary:** **Instead, for example, only selected researchers can be given access. Alternatively, it may be feasible that researchers can analyze your data without ever directly looking at it!**
  + This should in principle be possible, because researchers are interested in **generalizable patterns**, not in the attributes of particular individuals.
    - The concept “[differential privacy](https://en.wikipedia.org/wiki/Differential_privacy)” formalizes this idea. With differential privacy, patterns within a population can be discovered, but the information that can be inferred about any particular individual is strictly limited.
  + Relevant techniques include:
    - **Calibrated Noise** (can be used to implement differential privacy)
    - **Access Control** (providing different access based on sensitivity and trust in the analyst)
    - **Remote Execution** (bringing the query to the data)
    - **Secure Data Hubs** (storing the data in a secure system which offers a range of access possibilities)
* **Solutions for safe sharing are currently being developed:** Here are four example systems that implement some of the above techniques:
  + **OpenMined Duet:** Connects two Jupyter notebooks over the internet to allow one analyst to remotely analyze the data that is loaded in the other notebook.See the [blog article](https://blog.openmined.org/duet-opengrid-infrastructure-for-easy-remote-data-science/) and [demo video](https://www.youtube.com/watch?v=9qYDtt2XM2o).
  + **Aircloak:** A business-oriented [software product](https://aircloak.com/) that allows privacy-preserving analysis of data stored in a SQL database, using a technique similar to differential privacy, but which trades off some formal guarantees of privacy for more accurate analysis.
  + **Leonhard Med:** A cluster at ETH for processing privacy-sensitive data.   
    *NOTE (NOT IN THE PRESENTATION): Leonhard Med is part of the following system that aims to provide a comprehensive solution:*
    - **BioMedIT, part of the Swiss Personalized Health Network (SPHN):** See the [paper](https://pubmed.ncbi.nlm.nih.gov/32570566/) and [webpage](https://sphn.ch/network/projects/biomedit/). Supports remote desktop analysis. Based on infrastructure in Zürich, Basel, and Lausanne. Planned features include (federated) remote execution.
  + **Linkhub.ch**: An early stage initiative at the Swiss national level, focused on social science (see [website](https://linkhub.ch/)).

## Scenario 2: Federated Learning

An early example for **federated learning** (first defined [here](https://arxiv.org/abs/1602.05629)) is the Google Keyboard. Federated learning enables the following benefits when predicting the next word that the user wants to type:

* Keeps typing data local and private
* Maintains a personalized model
* Can learn from many users, for example to quickly adapt to changes in patterns seen across many users

## Scenario 3: Federated Learning on Media Data

With media data, typically media type-specific approaches need to be applied to protect privacy, but some techniques are also more generally applicable. A recent example applying privacy-preserving federated learning, including in combination with differential privacy, is the [PriMIA](https://g-k.ai/PriMIA/) system, described in [this paper](https://www.nature.com/articles/s42256-021-00337-8). This approach uses deep learning to analyze medical images, but should be generalizable to other media types, such as audio and text.

## Scenario 4: Time Series Data (NOT IN THE PRESENTATION)

Time series data is difficult to anonymize. It is often both sensitive (e.g. could contain information about a medical condition) and can be used to match individuals across data sets. The latter can be done in at least two important ways:

* If the time frames for the two time series overlap, a high correlation (over time) between two related variables can indicate a match.
* If the time frames do not overlap, patterns that are somewhat stable over time can constitute a “fingerprint” that can be matched across datasets.

A conceptually simple way to anonymize time series data is to replace them with summary statistics or latent representations, thus bringing the data into a form that is suitable for applying k-anonymity.

Another approach is to try to measure the similarity between two time series instances, and to modify the time series in a way that ensures that similar time series are sufficiently likely to be from different individuals.

# Recommendations

* When in doubt about how to publish your data, **ask an expert at your institution** (e.g. at ETH, contact the [ETH Library](https://library.ethz.ch/en/publishing-and-archiving/publishing-and-registering/publishing-research-data.html)).
* Keep in mind that ensuring **information security** is a prerequisite for privacy – there is little benefit in anonymizing your data if you accidentally leave the raw data for everyone to download, or if your system can easily be hacked.
* Note that **differential privacy** so far seems to be suitable only for a limited set of use cases in practice. It may only be useful if the number of individuals in your data set is far above what is typically achievable.
* Note that **new possibilities** for publishing sensitive data **are currently emerging** and may be relevant for you.

# Further Material

* An excellent resource for further learning as well as (emerging) relevant tools is the [OpenMined](https://www.openmined.org/) open-source community. They are currently offering two free [online courses](https://courses.openmined.org/), and more are in the pipeline.
* Additional interesting material/insights not mentioned in the presentation:
  + **Privacy Harms**: [This paper](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3782222) describes the different ways that failure to protect privacy can result in people being harmed (both obvious harms such as stalking, credit card abuse, defamation, coercion, and more subtle harms such as emotional impact, discrimination, manipulation and loss of control).
  + **Research Privilege in privacy law**: Legislation such as the EU’s [GDPR](https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/) and the Swiss [data protection law](https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/) allow the use of personal data for research even when no explicit consent has been given for this use, under the condition that there is sufficient public interest in the research.
  + **Legal requirements for anonymization are surprisingly vague**. For example it is usually required that “subjects must not be identifiable”, but it is not specified whether a subject should be prevented from identifying themselves in a published data set, or whether this applies to people who know the subject well. An exception to this vagueness is the U.S.’ HIPAA law, which defines a concrete method ([Safe Harbor Method](https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance)) for anonymization.