In recent years, we have witnessed an explosion of successful applications of deep learning, including object extraction, speech recognition, machine translation, car driving, recommendation systems, etc. In many applications of machine learning, such as machine learning for medical diagnosis, we would like to have machine learning algorithms that do not memorize sensitive information about the training set, such as the specific medical histories of individual patients.

Machine learning algorithms work by studying a lot of data and updating their parameters to encode the relationships in that data. Ideally, we would like the parameters of these machine learning models to encode general patterns rather than specific facts about specific training examples. Unfortunately, machine learning algorithms do not learn to ignore these specifics by default. If we want to use machine learning to solve an important task, like making a cancer diagnosis model, then when we publish that machine learning
model, we might also inadvertently reveal information about the training set. A malicious attacker might be able to inspect the published model and learn private information.

Scientists have proposed many approaches to provide privacy when analyzing data. For instance, it is popular to anonymize the data before it is analyzed, by removing private details or replacing them with random values. Common examples of details often anonymized include phone numbers and zip codes. However, anonymizing data is not always sufficient and the privacy it provides quickly degrades as adversaries obtain auxiliary information about the individuals represented in the dataset.

Centrally kept data is subject to legal subpoenas and extra-judicial surveillance. Many personal data - for example, medical institutions that may want to apply deep learning methods to clinical records - are prevented, for reasons of privacy and confidentiality, from sharing data and thus benefiting from large-scale deep learning.

The objective of this thesis is to design, implement and evaluate a practical system that allows several parties to jointly learn a specific neural network model for a given purpose without sharing their input data. This is to exploit the fact that the optimization algorithms used in modern deep learning, namely those based on stochastic gradient descent, can be paralleled and executed asynchronously. One of the objectives would be for each participant to train independently on his or her own data set while selectively sharing with other participants a subset of the key parameters of his or her model. Each participant thus preserves the confidentiality of his data while taking advantage of the models of the other participants, which allows him to improve the accuracy of learning beyond what is only possible with his own data.

For example, one could imagine that a network of doctors or hospitals could exchange parts of the model in this way rather than building a single global system hosting all the data.

This project is obviously carried out in collaboration with the medical teams of the University Hospital and in particular with the Department of Medical Informatics (DIM) headed by Professor Pascal STACCINI.

This thesis is part of the IDB project funded by IDEX UCA IDEX. Anonymous data are available representing 10 years of EHRs for care and resuscitation suites (PMSI SSR) and for medicine, surgery and obstetrics (PMSI MCO).

**URL:** http://www.i3s.unice.fr/~riveill/subjects/privacy.html

**English version:**

In recent years, we have witnessed an explosion of successful applications of deep learning, including object extraction, speech recognition, machine translation, car driving, recommendation systems, etc. In many applications of machine learning, such as machine learning for medical diagnosis, we would like to have machine learning algorithms that do not memorize sensitive information about the training set, such as the specific medical
histories of individual patients.

Machine learning algorithms work by studying a lot of data and updating their parameters to encode the relationships in that data. Ideally, we would like the parameters of these machine learning models to encode general patterns rather than specific facts about specific training examples. Unfortunately, machine learning algorithms do not learn to ignore these specifics by default. If we want to use machine learning to solve an important task, like making a cancer diagnosis model, then when we publish that machine learning model, we might also inadvertently reveal information about the training set. A malicious attacker might be able to inspect the published model and learn private information.

Scientists have proposed many approaches to provide privacy when analyzing data. For instance, it is popular to anonymize the data before it is analyzed, by removing private details or replacing them with random values. Common examples of details often anonymized include phone numbers and zip codes. However, anonymizing data is not always sufficient and the privacy it provides quickly degrades as adversaries obtain auxiliary information about the individuals represented in the dataset.

Centrally kept data is subject to legal subpoenas and extra-judicial surveillance. Many personal data - for example, medical institutions that may want to apply deep learning methods to clinical records - are prevented, for reasons of privacy and confidentiality, from sharing data and thus benefiting from large-scale deep learning.

The objective of this thesis is to design, implement and evaluate a practical system that allows several parties to jointly learn a specific neural network model for a given purpose without sharing their input data. This is to exploit the fact that the optimization algorithms used in modern deep learning, namely those based on stochastic gradient descent, can be paralleled and executed asynchronously. One of the objectives would be for each participant to train independently on his or her own data set while selectively sharing with other participants a subset of the key parameters of his or her model. Each participant thus preserves the confidentiality of his data while taking advantage of the models of the other participants, which allows him to improve the accuracy of learning beyond what is only possible with his own data.

For example, one could imagine that a network of doctors or hospitals could exchange parts of the model in this way rather than building a single global system hosting all the data.

This project is obviously carried out in collaboration with the medical teams of the University Hospital and in particular with the Department of Medical Informatics (DIM) headed by Professor Pascal STACCINI.

This thesis is part of the IDB project funded by IDEX UCA IDEX. Anonymous data are available representing 10 years of EHRs for care and resuscitation suites (PMSI SSR) and for medicine, surgery and obstetrics (PMSI MCO).