ISSN 2281-5147

HOW TO COLLABORATIVELY USE STATISTICAL MODELS IN A SECURE WAY

Maciej Krzysztof Zuziak*

Abstract

The following articles compile research on the brink of privacy, federated learning and data governance to provide a reader with a basic understanding of the nuanced world of decentralised learning systems. It starts from simple notions of personal data and its connection to artifical intelligence. Afterwards, it goes into the realm of statistical learning to explain the basic technocratic lingo in a (hopefully) engaging way. With those topics covered, it proceeds to deliver on the basic notion of Data Collaborative and Decentralised Data Governance - an arcane term that the reader will be familiar with at the end of this lecture. Finally - it poses some open-ended remarks on the future of data analysis done in a way that benefits our communities. While the delivery of the article is relatively simple and straightforward, it also provides the curious reader with links and pointers that would allow them to go deeper into a well of data governance and large AI infrastructure.

Table of Contents

* Maciej joined LeADS in November of 2021 to work on the topic of privacy-enhanced Machine Learning and Personal Data Management at the Institute of Information Science and Technologies "Alessandro Faedo" at the National Research Council of Italy. He deals with the topics of Decentralised Machine Learning and Alternative Data Governance. During his three-years research at the Institute, he published a number of articles on both of those topics on a reputable venues such as IEEE Big Data or ACM FaaCT.
maciejkrzysztof.zuziak@isti.cnr.it, maciej.k.zuziak@protonmail.com
This work is supported by the European Union's funded project Legality Attentive Data Scientists (LeADS) under Grant Agreement no. 956562.

ISSN 2281-5147

Keywords

Federated Learning – Machine Learning – Data Governance – Alternative Data Governance – Data Collaboratives

1. Introduction to a world of big-data

The advent of highly capable intelligence systems is evident to everybody. From talk shows to popular news outlets, terms like *Artificial Intelligence*, *Machine Learning,* and *Data-Driven Economy* are being constantly called out and discussed, with little to no explanation of what is actually hiding behind them. Sometimes, you can hear that those systems *consume* (or perhaps – *require*) a tremendous amount of data to be *trained* – but what does this mean in practice? If you are one of those who encountered those terms in the wild – you may ask yourself a question: *but why should it actually concern me?* Suppose you have pondered on that matter a little bit more. In that case, you might actually rephrase the question a little into the following: *but what is the relationship between my personal information and the behaviour of those systems?*

ISSN 2281-5147

If you had indeed asked yourself one of those (or similar) questions, this text is addressed to dispel your doubts or instil new ones. If you haven't asked those questions yet, but the opening paragraph has sparked your interest, then there will be no better time to start seeking answers to those issues and this text may be a good beginning of that journey. I must disclose that this text is written clearly and succinctly and is addressed to a reader who wants to satisfy their curiosity rather than to an expert who has already spent a fair number of hours (perhaps weeks or maybe years) researching presented topics as it will simplify many steps and definitions to arrive at simple explanations of rather complex things. In Feynman's book titled QED: The Strange Theory of Light and Matter[1], the author states that explaining complex concepts in plain words is an art itself, and – I would like to add after him - as in art, one can fail miserably at it. However, I did my best to make this will digestible for someone who did not touch the matter presented in it previously or perhaps burnt their hand while trying to touch it, as it requires knowledge from a fair number of disciplines to comprehend it fully (and I do not pose myself as one of those who comprehended it. I somewhat believe that the most honest response would be to say that we all struggle to understand it, and the history of that struggle is what we present the world with).

2. Personal Data and Statistical Inference

It may be best to start the journey from the concept of *personal data*. For the sake of simplicity, let's assume that *personal data* is any information that relates to you as an identifiable living individual[2]. The GPS routes saved in your telephone, the photos with nametags on your iPhone, the health records your local doctor keeps…all this can constitute *personal data*. Of course, the reality may be more complex, but for the sake of this text (and for the sake of all other debates about *personal data* that you stumble across from time to time), it is safe to assume that this simplified definition is all you need (and indeed, this will not be that far from the truth).

[1] (Richard P. Feynman & Anthony Zee ([introduction], 2024)

[2] The fairly simple explanation of the term personal data provided on the European Commission website states that *Personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data*. While the concept of personal data is more nuanced in practice, this definition is a fundamental building block for all more subtle interpretation regarding personal data (European Commission, 2024).

ISSN 2281-5147

Now, let's focus on the concept of *statistical model*. According to Wikipedia – which closely follows D.R. Cox's book on Principles of Statistical Inference (Cox, 2006) – *a statistical model is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population). A statistical model represents, often in a considerably idealised form, the data-generating process*. However, this definition may bear little to no meaning to a person not acquainted with statistics (and the one who is acquainted with statistics probably would not require this explanation in the first place), so let's take a step back. Most of us would probably agree that the number of ways a specific phenomenon can be measured is limited. Political polls generally survey only a limited number of participants since surveying all of those entitled to cast a vote would be nearly impossible in the first place. Explaining consumers' trends and turnovers operates on (limited) historical data and is constrained to events that can be (effectively) recorded. Drawing conclusions about the height and size of a specific type of penguin is based on the measurements of penguins in a limited sample since catching and measuring all of them would be considered highly impractical – and so on. It is obvious that while making conclusions about a population at large, the best we can often do is to take a minimal glance over the window of our own reality. On an intrinsic and highly intuitive level – the statistical models are the windows through which we can deduct information about the reality surrounding us. The models themselves are wooden frames, while the model's parameters are the frame's dimensions. Once we choose a frame that has caught our attention (and one which we believe will give us a good outlook on the world around us), we must tune the width and height of that frame – so that it captures exactly what we want it to. *Training a statistical model* means no more than adjusting those parameters with the help of prior information we have obtained.

3. Training Your First Statistical Model

Let's take a look at an illustrative example. Let us assume you want to identify patients with a high risk of a particular genetic disease. Due to a high number of patients and a complicated diagnosis process, you want to automate this process to foster the preventive measures that can prevent the disease's development if applied early. Then, you would need to develop a classification model, i.e., classifier – a hypothesis

ISSN 2281-5147

function that will return 'true' if a specific genetic marker[3] is associated with a higher risk of disease occurrence and 'false' otherwise. Assume that the *hypothesis function* is in the form of a black box. It is hardly a metaphor, just imagine a black steel box arriving at your desk. The black box has two modes: *train* and *infer*. When the black box is in *train* mode, it tries to distinguish between the markers potentially associated with a higher chance of occurrence and those risk-free. The first step would be to switch our black box to train mode and teach it the distinction. As with children, the simplified learning process relies on correlating the features of the object with a corresponding label. While human beings' generalising capabilities give us a remarkable ability to learn a pattern by observing just a few data instances, machines tend to require more examples to familiarise themselves with a specific pattern. On the other hand, their main strength is strictly connected to the ability to detect much more complex patterns that we could potentially detect, making them a perfect tool for identification based on genetic markers.[4]

[3] A genetic marker is a gene with known location on a chromosome that can be used to infer properties about the individual or species. A beautiful comparison of a genetic marker to a *landmark* is presented on the site of National Human Genome Research Institute. In this narration, a genetic marker can be compared to a marker (characteristic location) that can help you navigate through a city that you are not yet familiar with (National Human Genome Research Institute (NIH), 2024)

[4] The human learning process is – of course – more nuanced than that and relies on many different skills developed throughout our upbringing. Interestingly enough, the machines also benefit from learning some fundamental knowledge about the domain. A fairly accessible article by M.G. Levy exploring this topic can be found under the link: https://www.quantamagazine.org/machines-learn-better-if-we-teach-them-the-basics-20230201/ (Levy, 2023).

ISSN 2281-5147

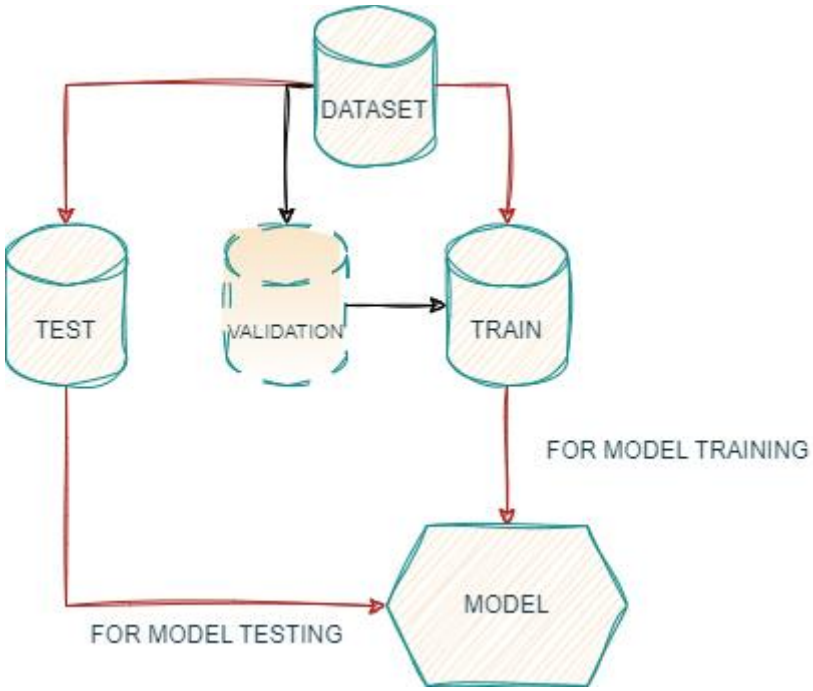

Fig.1: Before the training, we split the dataset into two sets: one for training purposes and one for testing purposes. Sometimes we also distinguish a third set for validation purposes (checking the progress of the training without using the external test part that is reserved only for a final evaluation).

However, to begin the training, we will require a number of pre-classified genetic markers with corresponding labels ('risk-associated and 'risk-free'). The sample of pre-classified markers will constitute our *training set*. Another sample of pre-classified objects is necessary to evaluate the general performance of our model – this will be called our *test set*. The distinction between training and testing datasets is crucial, as we do not want to evaluate the performance of our model based on the same things that we trained it on. Using education as an example, teachers and professors seldom provide students with answers before the test.

ISSN 2281-5147

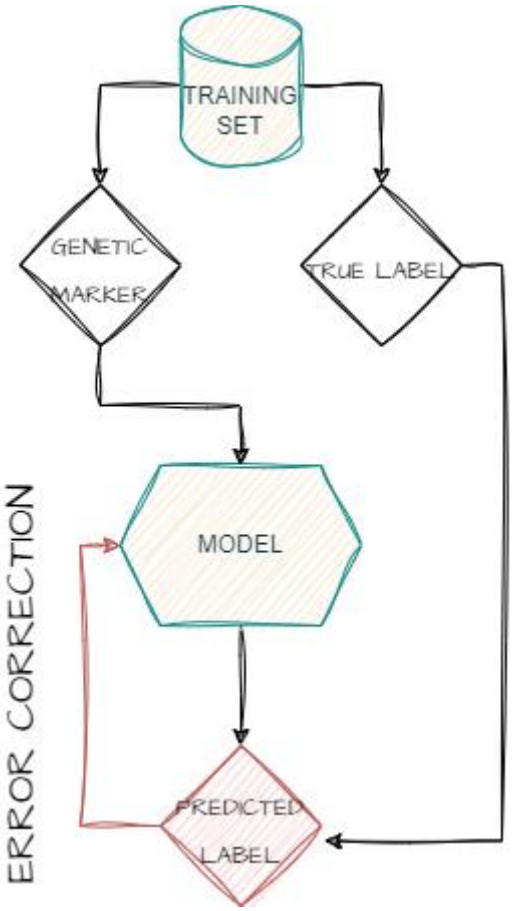

Fig.2: Training phase. The model associates genetic markers with corresponding label. The difference between the predicted and true label is used to correct the model.

The training will consist of putting a pre-classified marker within a black box, scanning, and then picking up another data sample. Depending on the task complexity, it may require a few hundred to a few hundred thousand different data points to be deployed. After the training, when you switch the black-boxin to *infer* mode, it will allow you to make an automatic classification of the data sample. You do that by placing a marker inside the device and then pressing the big red button on the top of the black box. It either returns 'true' or 'false' as mentioned above, where *true* value would imply that the patient is at risk of occurrence of a particular genetic disease. Now, let us imagine that the black box has memory cards that are used to distinguish between markers. Those memory cards preserve the information about what types of markers are associated with a high risk of occurrence. The knowledge is based on markers that they have seen so far. In the machine learning lingo, the memory cards are called *parameters*, while the black box is often called *architecture*. When we use the word *model*, it often implies a specific black-box architecture with a pre-determined set of *parameters*. The

ISSN 2281-5147

markers we use to *train* the black box are part of our *sample*, while the black box will be used to make predictions about the *population*.

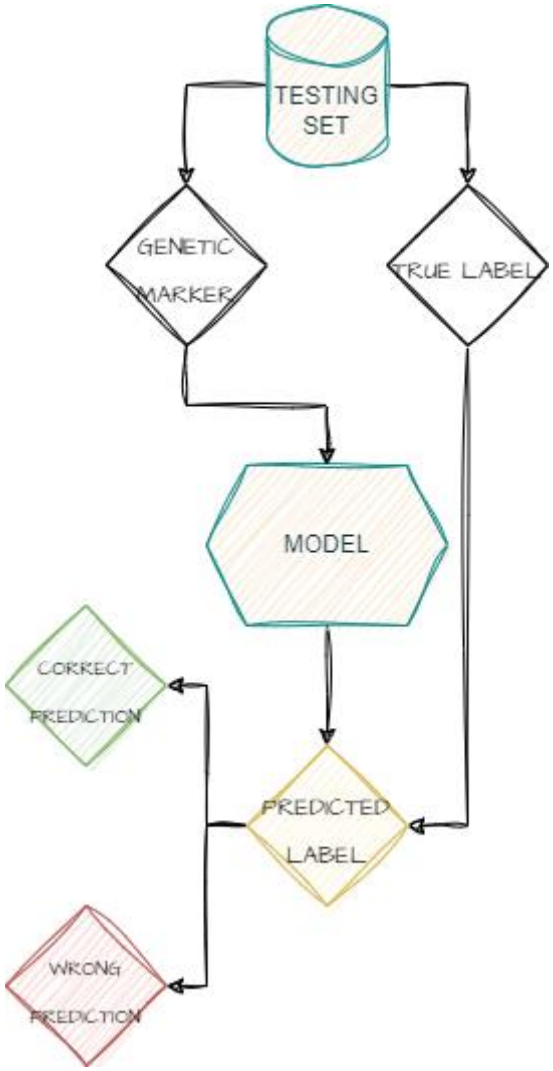

Fig.3: Testing phase of the model. This time, we count the correct and wrong predictions to quantify how well our model performs on unseen data. In contrast to the training phase, this time, we do not use the difference between the true and predicted labels to correct our model.

4. Between Statistical Inference and Personal Data

As can be understood from the last paragraphs, training the model requires some amount of pre-classified data that can be used as a basis for future pattern recognition. The source of this data depends on the particular task we aim to solve. In many

ISSN 2281-5147

instances, the data will be registered during carefully designed experiments. In other instances, the samples will be provided by us directly, as is in the case of digital marketing, where our commercial activities are the basis of sentiment analysis and consumer-type classification. I want to focus on the second case, as it involved the previously introduced concept of personal data.

Genetic markers may – or may not – constitute personal data. Suppose the genetic marker can be uniquely attributed to an individual person. In that case, it is considered a special category of personal data that is guarded by a higher regime of legal protection under the General Data Protection Regulation. As with many cases of human-related data, the final answer to whether it constitutes personal data will be highly context-specific and based mainly on the notion of identifiability (whether a specific sample can or cannot be attributed to a particular person).

There is a catch – acquiring (any kind, not only personal) data is expensive. It requires a lot of infrastructure. Firstly, to collect it, then to store it – not even mentioning any methods of processing it to obtain any meaningful information of our interest. Large technological companies have a natural advantage here. They provide and utilise a large-scale architecture that is mainly based on interconnected social networks – while they still need to abide by the consent-based regulatory framework that we've described earlier, the staggering scale of their enterprises allows them easily to obtain the consent of their users, as those users are often engaging with their services daily (think about online marketplace, video-on-demand services, social networks providers and other entities that monetise heavy user-flow of their platforms). Of course, public infrastructure and individual citizens seldom can deploy at such a scale. It naturally explains the benefit of the size, which can make or break in terms of becoming a champion of the digital race.

5. And The Weak Suffer What They Must?

The pioneers of the digital race, the same that utilise the large-scale infrastructure mentioned earlier, do that in their own name and for their own benefit. It should hardly come as a surprise since those entities are enterprises that main aim is to generate revenue and their investment in said infrastructure was calculated as a long-term investment. How they benefit from it may depend on the particular case

ISSN 2281-5147

scenario. Netflix – subscription video-on-demand service – uses data analysis to train recommendation models that can deliver their clients a suitable set of new titles to watch every time they visit the main homepage (which, in turn – can allow them to retain their customers by engaging them in yet another series). Amazon – a multinational technology company – uses a stream of data from various services they offer to create multi-modal systems predicting behaviours and trends of their customer base. Since the time the media has expressed enormous interest in the Large Language Models (LLM) – models that are capable of (among other) language generation – a number of companies have tried to jump on the bandwagon with their products. Because those models are trained on vast amounts of text and require enormous monetary expenditure – the race was briskly overtaken by industrial champions, with little space left for smaller entities.

You may ask – what about the individual users? Can they also benefit from the blessings of big data advent? Well, that is the question that I try to answer with my research. Undeniably, an individual has a very limited ability to acquire the data, infrastructure and workforce necessary to achieve such a task. There exist open source pre-trained models that are available on the market. Pre-train – in that context – means that most of the hard work has already been carried out, and you only need to adjust the model to your needs. This sounds tempting, but like always – there is a catch. Firstly, those models may not suit our particular needs. In drastic simplification, a model (and an architecture) that was reserved for an image classification task (let's say – distinguishing between cars and pedestrians) would not be a viable choice for a language translation – and vice versa. Secondly, the 'pre-trained' does not mean 'fully-trained'. For the model to suit your desired purpose (whether it will be digital marketing, medical research or language generation) – you will still need to have access to some (possibly personal) data that will be used in the course of pre-training the model. Hence, we have come a full circle.

6. Benefits Through the Collaboration

When individual efforts do not yield a significant result, it is a natural behaviour to look upon the concept of collaboration. In fact, some scholars have noticed that data-driven endeavours do not have a monopolistic nature by default. In fact, they may be

ISSN 2281-5147

more suited towards collaborative effort than they seem at first glance. This observation has led to the development of several concepts, such as *Data Trusts, Data Collaboratives or Data Cooperatives* (Mozilla Insights et al., 2020). While they all come with some nuanced differences, I do not want to dwell on that topic too much. However, I would like to refer an interested reader to the studies on the difference between collaboration-based concepts performed by Mozilla Insights together with Jonathan von Geuns and Ana Bradusescu since their findings are fully open to the public[5]. What matters from the perspective of this essay is the basic understanding of the *Data Collaborative* concept as defined by our research.

In the simplest terms, the Data Collaborative is a common undertaking of independent actors to train a shared model (or a number of shared models) throughout the lifecycle of the collaboration. The cornerstone of each Data Collaborative is its ability to train one statistical model by a number of participants collectively. For example, let's assume that we have five hospitals in Tuscany. All of them possess highly specific data about a number of patients with pulmonological diseases of a certain type. Let's further assume that those diseases can be identified based on X-ray imaging. However, their occurrence is somewhat region-specific, and the symptoms visible on the X-ray image will vary from region to region. Since the disease is highly problematic, those five hospitals want to train a model for its early detection to aid doctors in analysing the X-ray imagining. Seldom will it be the case that those hospitals are able to use some pre-trained model since such a study could not have been conducted yet or the results of such a study (hence, access to a model) are not publicly available. Neither often is it a case that those hospitals can train one model individually. Hence, they create a structure, a *Data Collaborative*.

The *Data Collaborative* is a very wide definition of every structure that fulfils some pre-defined criteria and allows for shared model training. In our article, *Data Collaboratives with the Use of Decentralised Learning,* we have developed four fundamental principles that could characterise such a structure (Zuziak et al., 2023).[6] *Firstly, the data collaboratives should provide an accessible infrastructure for performing various analytical tasks without the necessity to transfer raw data beyond the participants' devices [**Decentralised Data Storage**]*. It means that the hospitals will not make their x-ray images public due to

[5] Link to the studies form September 2020: https://assets.mofoprod.net/network/documents/ShiftingPower.pdf.
[6] Link to the article from June 2023: https://dl.acm.org/doi/10.1145/3593013.3594029

ISSN 2281-5147

privacy reasons, nor will they create a shared data storage that will hold all of their data in one place. It means that the very data that is necessary to train the model will not be transferred beyond the (local) datacentres belonging to the hospitals. You could now be asking: *How is that even possible? Isn't the data absolutely necessary to train the model?* Indeed, it is a perfectly reasonable question that I will try to answer soon. *Once the model is trained (or once an analytical task is accomplished), it should be governed by all the members of the collaborative (in proportion to their marginal contribution). Shared governance is a key guarantee that all the members will benefit from joint participation in the analytical tasks* by collectively making decisions about the future of the model. *Collective-choice arrangements could be realised by allowing participants to collaborate to create their own rules and governance conditions. [**Shared Model Governance**].* Since the model is trained by a consortium of hospitals, it will be governed by the joint board of their representatives. It guarantees that no member of the collective can, without prior consent from others, remove, modify or make the model publicly available. Since all of the hospitals are equally participating in the training, using their own data and incurring infrastructure costs, they must have some common way of making a decision regarding the model governance. In terms of hospitals, it can manifest in a set of rules regarding how the model can be made public and under what circumstances (for example, after unanimous consent of all the members of the collaborative). *The structure of a data collaborative should be mostly implementation-agnostic. This is because any structure or implementation that satisfies the baseline definition and the four essential principles can be treated as a data collaborative - irrespective of the implementation details [**Universality**].* This is mostly due to the fact that such a partnership can be implemented with multiple technologies that are available. While – in the next section – this article overviews only one of them (namely, Federated Learning). By no means can the use of other methods preclude someone from using the term *Data Collaborative*. Since the pool of available technologies is evolving all the time, this ensures that the concept will stay universal long after its first presentation. *Finally, data collaboratives can be established and executed in many different ways, combining the available technology with local needs. However, each data collaborative should be able to perform at least one analytical operation in a distributed environment [**Minimal Utility – Collaborative Computation].*** Since the original goal of the Collaborative was to train one statistical model, it is a natural consequence that it should serve its goal. While the hospitals may opt-in for training multiple models, one common undertaking is an essential criterium for the existence of the collaborative.

ISSN 2281-5147

7. Data Collaborative in Brief

In very brief terms, the ***Data Collaborative*** is a vehicle for training one or more statistical models in collaboration with other members – where they share resources and data to come up with a shared result. But how can the model be learnt ***collaboratively*** without firstly gathering the raw data in one (central) place? Well – and here perhaps you will need to take it at face value – the model is learnt through sharing the intermediate values that are learnt locally. More precisely, each participant is learning their own local model based on their own data. In a subsequent step, a global model is created using a mixture of local models. The technique that we used for the sake of this research is called ***Federated Learning,*** and it was proposed as the more privacy-centered manner of statistical learning some years ago. Referring to an example of hospitals that were made before, each of the hospitals will train its own model for classifying the X-ray images based on the locally available data. The next step will involve merging those models into one global model that can accumulate the knowledge of all its local counterparts. The procedure can continue until a satisfactory result is reached.

There are a few more technical issues that we experiment with in that setting. Firstly, there is of course, an issue of fault detection. Since the number of collaborators can be semi-trustworthy, it is crucial to detect potential intruders and free-riders. Although it can be difficult to imagine it in terms of hospitals (that are public institutions of a high reputation), let us assume that instead of hospitals, the Data Collaborative is formed by a number of industry partners that want to train one model for consumer classification. In such a case, it may be reasonable to expect that some of them may be either interested in obtaining a model for free (without really contributing their own knowledge) or jeopardising the global model (by including false information). In one of our papers called *Amplified Contribution Analysis for Federated Learning* in 2024, we presented an intuitive way of detecting participants that could potentially harm the global model (Zuziak & Rinzivillo, 2024)[7]. Another open issue would be that of *personalisation*. Since the models can be learnt on different data-generating distributions, there may not be a suitable mixture of models that suits all

[7] Link to the paper form April 2024: https://link.springer.com/chapter/10.1007/978-3-031-58553-1_6

ISSN 2281-5147

the participants of the learning. In this case, we would like to make some splits between the members of the collaborative to allow them to learn individualised models. If the hospitals are located in places of vastly different population characteristics, one diagnostic model could not necessarily be effective locally, as it could fail to focus on lcoally relevant traits of the population. In this case, we would like to automatically detect such variations in needs and allow the participants to *personalise* their models. Lastly, there are some issues with privacy that may not be so obvious to spot. In the introduction to statistical models, I have said that memory cards that preserve information about what types of genomes are associated with the higher chance of genetic disease occurrence are called *parameters*. Well, those parameters can also store personal information about objects they were trained on. For example, an attacker may be able to infer whether the genome of a certain person was included in the model training and its corresponding label. This would imply that they will obtain knowledge about whether a certain individual is associated with a higher chance of genetic disease occurrence.

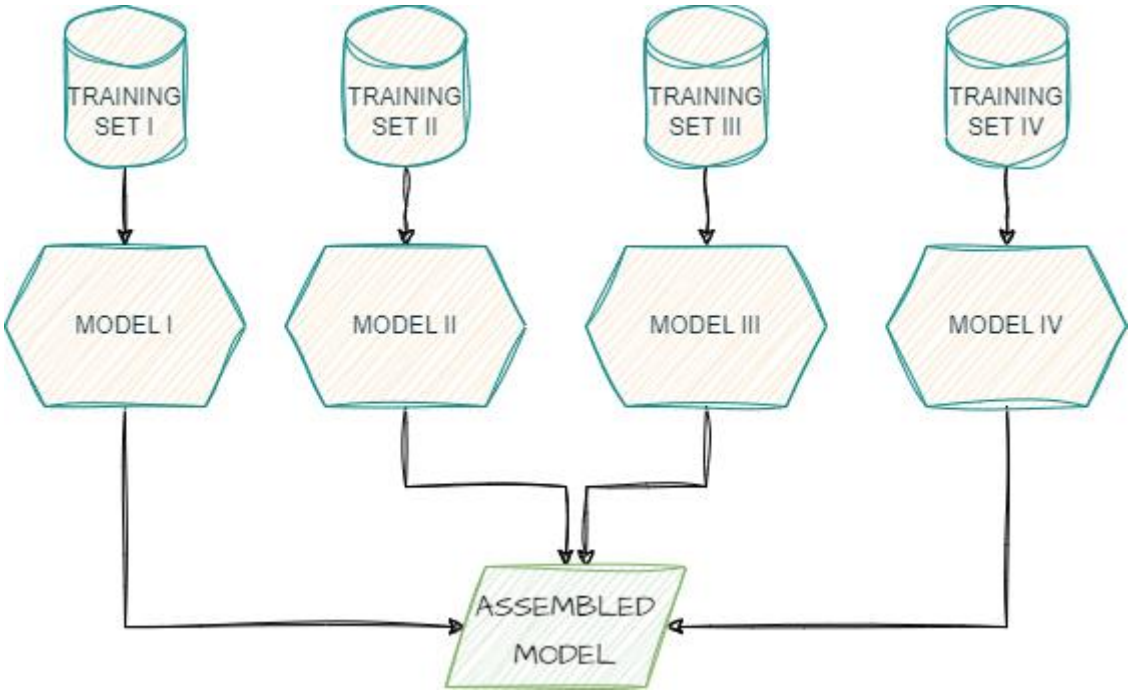

Fig.4: Model assembly. By combining a multiple model into one, a number of different organizations can create can shared model with far great capabilities, without the need to transfer the data directly into hands of one of the organizations.

ISSN 2281-5147

8. The Farewell Note

There are many concepts and nuances that were not included in this text – as they go beyond the brief description that could be delivered in one digestible piece of writing. However – I have – as an author - promised to deliver a set of pointers to a curious reader who would like to expand their knowledge upon this lecture. Throughout this text, I have included numerous links to articles that served as building blocks for this research – and I can firmly reassure every curious soul – that those links provide their own sources, upon which you can further expand your understanding of the topics delivered here.

In the last paragraphs, I have tried to explain the notion of *statistical model* in the most accessible manner possible. I have also related the concept of a model to one of personal data and why – in some cases – the availability of training samples may be highly dependent on the size of the business infrastructure (and – in turn – how this benefits larger players). This all ties back to the concept of *Data Collaborative* – a formal association of individual members that joints efforts in training one common statistical model. In the afterwords, I want to express my belief that we all value our personal data. Hence, we should not turn away from the difficult discussion on how we all may benefit from it when it comes to statistical inference. As the examples were simplified in order to explain the described concepts, they are by no means restrictive to the number of ways that the *Data Collaborative* can be used. From real-time traffic jam inference to research on genomes, common efforts in statistical analysis may impact our lives in a number of ways. It only depends on us and our communities how we will use it to benefit us and those around us.

9. Selected Readings

Cox, D. R. (2006). Principles of Statistical Inference. Cambridge University Press. https://doi.org/10.1017/CBO9780511813559

European Commission. (2024, September 1). What is personal data? [Govermental Website]. European Commission Website. https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en

ISSN 2281-5147

Levy, M. G. (2023, February 1). Machines Learn Better if We Teach Them the Basics. Quanta Magazine. https://www.quantamagazine.org/machines-learn-better-if-we-teach-them-the-basics-20230201/

Mozilla Insights, Jonathan van Geuns, & Ana Brandusescu. (2020). Shifting Power Through Data Governance (p. 22) [White Paper]. Mozilla Foundation. https://assets.mofoprod.net/network/documents/ShiftingPower.pdf

National Human Genome Research Institute (NIH). (2024, September 1). Genetic Marker (Glossary Entry) [Govermental Website]. National Human Genome Research Institute (NIH). https://www.genome.gov/genetics-glossary/Genetic-Marker

Richard P. Feynman & Anthony Zee (introduction_. (2024). QED: The Strange Theory of Light and Matter. Princeton University Press. https://press.princeton.edu/books/paperback/9780691164090/qed?srsltid=AfmBOorY57KBeFZCtddUvUs0yZLLSUlSGOtGFAlXQnIZKNrQM9MNIFhM

Zuziak, M. K., Hinrichs, O., Abdrassulova, A., & Rinzivillo, S. (2023). Data Collaboratives with the Use of Decentralised Learning. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 615–625. https://doi.org/10.1145/3593013.3594029

Zuziak, M. K., & Rinzivillo, S. (2024). Amplified Contribution Analysis for Federated Learning. In I. Miliou, N. Piatkowski, & P. Papapetrou (Eds.), Advances in Intelligent Data Analysis XXII (pp. 68–79). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-58553-1_6

ISSN 2281-5147