What is data?

Role of data in an AI system

In the general digital sense, data corresponds to the information that is used, processed and generated by software in a computer system.

There is no AI without data. Data plays a central role in all machine learning processes as it is used for both training and for testing. It also comes in the form of the parameters used for managing the training processes. Finally, the AI system is a combination of a certain software architecture with all the trained parameters, the so-called model, which is also data.

Understanding the role of data in AI systems together with the way it is selected, documented and disseminated is essential for being able to assess the behavior of an AI system. This is important in terms of reproducibility or for comparing two different AI systems.

In the natural language processing world for instance, the availability of large quantities of spoken and written data is central to the performance of spell checkers, prediction in search engines, or of course machine translation. They are used to build up so-called language models, which feed further processes with statistical representations of word or sentence combinations.

The sustainability of AI systems is thus closely dependent upon the data management methods that are deployed in their design.

Data for supervised and unsupervised AI system

As we have already seen, AI systems come in two flavors depending on the way data is used to train them. Supervised systems rely on the provision of inputs together with the corresponding intended outputs. The training thus consists in teaching the system to generate the most probable output out of unknown inputs. There can be various ways to obtain such data. For instance, an image database where each image is associated with keywords or a collection of digitized documents which have been transcribed by annotators (cf. figure below).

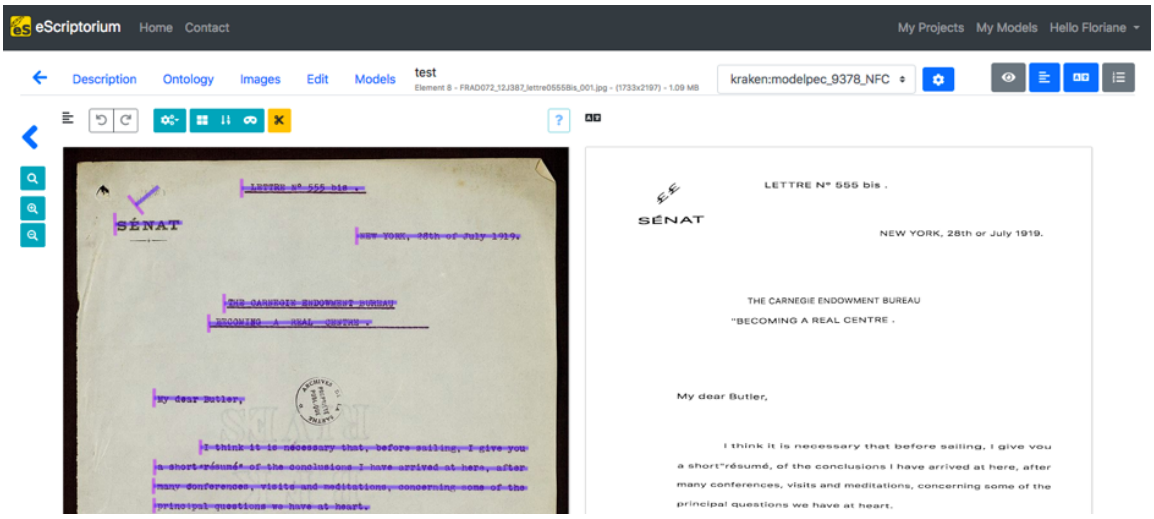

Figure: Automatic transcription of a letter by Paul D'Estournelles (courtesy of F. Chiffoleau, Coll. Archives de la Sarthe)

AI systems based on unsupervised training will not be designed for a specific behavior but designed to bear the statistical property of the training data. This is the case for instance of language models such as BERT which tend to associate similar positions in a mathematical space to words having the same syntactic or semantic behavior, observed from the provision of a large number of sample sentences for each word. Such models are for instance very good at predicting synonyms or the follow-up words of a given sequence.

Sources - selection, documentation, preparation, annotation

The design of an AI system is essentially based on the proper design of the data set that is used to train it. Among the various factors that come into place we can cite the relevance of the data for the task at hand, the size of the data that should match the complexity of the AI software architecture --- the more mathematical parameters you have to train, the more data you need --- and the variety of samples that should reflect the complexity of the task.

Depending on the data sources, data must be selected and often cleaned up before it is fed into the training process. If we take for instance the training of a language model on the basis of web content, the various samples must be sorted according to the actual language, depleted from the accompanying web code (HTML, Javascript etc.) and possibly shuffled to prevent copyright infringements. A good example of such data preparation is the design of the OSCAR corpus[1].

The design of annotated data for supervised AI systems is more complex as it involves the design of an annotation scheme, the organization of annotation campaigns and the control of the quality of the annotated data for instance by assessing the agreement between annotators on the same data.

All in all, it is essential that the design process be well documented to be able to track back to the source of possible failed behaviors in the resulting trained system.

Biased AI systems

As just alluded to, the behavior of an AI system closely reflects the nature of the data which has been used to train it. This in turn may induce a whole range of possible biases that may result from the underlying selection of data sets. For instance, a language model which is only trained on newspaper articles will cover completely different types of expressions and topics than one for which literature or social networks content has been chosen. In the same way, image generation systems will reflect the size and variety of the source image databases (e.g. artistic works) that have been considered.

In the case of supervised systems, a specific bias may come from the way the annotation labels are designed as well as the way the annotators, with their own cultural background, will interpret the data. If for instance you want to identify hate speech on social networks, the way sentiments are interpreted by annotators may vary according to the age,nculture, and personal feelings of the annotators vis-à-vis the material to be annotated.

All in all, one should always keep in mind that AI systems are by construction very conservative with regards to their training data and thus with regards to existing observables. One cannot expect any kind of real creativity from an AI system.

Hosting, pooling, distributing data

Because of the size and possible complexity of the training data in AI systems as well as the resulting models, various initiatives have been set up to allow for their hosting and distribution.

Open data sets and models can be hosted in specialized repositories (ex. the image data resource[2]) or in generic national or international (e.g. Zenodo[3]) repositories. Such repositories usually provide the necessary infrastructure for managing authorship, licensing, versioning and archiving of their content.

In the case of complex tasks, where various teams work in parallel on annotating a variety of data samples, some initiatives act as catalogs for the corresponding data sources. This is the case for instance for the HTR United[4] initiative which gathers metadata of annotated documents for (handwritten) text recognition.

---

1. OSCAR corpus: https://oscar-corpus.com/ ↩
2. Image data resource: https://idr.openmicroscopy.org/ ↩
3. Zenodo: https://zenodo.org/ ↩
4. HTR United: https://htr-united.github.io ↩