# Data Housekeeping Checklist

Welcome to the era of artificial intelligence (AI), where the way you do business is reliant on data-intensive technologies like machine learning and deep learning. To take advantage of these new AI tools, you need to make sure your organization's data "house" is in order.

Here's a checklist to get you started on your path towards a clean data house, broken down into the two key phases of housekeeping — training and inference.

For more information on how to implement AI, visit *https://www.ibm.com/it-infrastructure/solutions/ai*

## Training

In the training side of preparing for AI, you'll be developing algorithms to understand a data set. Your main concern will be gathering existing data and utilizing AI to learn a new capability.

- [ ] Figure out the specific business problem you'll want to solve using AI (start with smaller projects to help you learn)
- [ ] Locate the data that can solve that problem from relevant sources (it most likely won't all be located in a single place)
- [ ] Prep your data with metadata tags to significantly reduce the time required to find pertinent data
- [ ] Ensure your data is properly synched and linked across all the data sets you'll be using (including time synchronized)
- [ ] Flag any customer-sensitive and other private data to make sure you're keeping it absolutely secure and abiding by all appropriate governance and regulation (the metadata tagging process can help with this)
- [ ] Choose the right development environment for the type of data you're using and the way it will be formatted (i.e., images, video, free-form text and audio each typically have one kind of environment)
- [ ] Pull data sets from your repository and bring them into your development environment
- [ ] Split your data into two groups to help improve your model development process (keep one set in a folder called "train" and another set in a folder called "test")
- [ ] Maintain data traceability by keeping track of where/what source your data has come from (consider using tools that can help automate the process)
- [ ] Perform basic data hygiene tasks to prep the data for building a model (for example, include filling in missing data entries and removing null entries)
- [ ] Use a subset sample of data for which you already know the answer to the prediction activity (this is called a "training set") and identify all the pre-processing steps needed to prepare the data to make a prediction
- [ ] Use your knowledge of this training set to compute accuracy scores that may give you confidence to apply the same model to new data for which the model has never been explicitly trained

## Inference

Once you've developed a model that works to solve your business problem, you'll move from training to inference. In this phase, you're taking that successful model and applying it to new data, which requires some ongoing data housekeeping as well.

- [ ] Locate your AI model close to your data to reduce latency, reduce bandwidth requirements and improve overall model performance
- [ ] Develop an efficient data pipeline process and apply metadata labeling to your data as it comes in, so that new data can be gathered and used to enhance the model going forward
- [ ] Tag data in a way that is linked and synchronized (for example, if the data is time sequenced, you can synchronize across data sets or link by picking one field — like a customer's name — across all data that comes in)
- [ ] Develop a long-term data life cycle storage plan for how you will manage the volume and velocity of the data as it comes in and as you archive it
- [ ] Consider hiring a Chief Data Officer to maintain your organization's data management for future AI, deep learning and other data-driven projects



IBM