veridu

Data

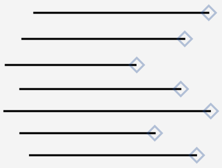
Veridu’s data analysis process

Analysing a digital footprint

Your digital footprint contains a staggering amount of information about you, and that’s what Veridu uses to verify your identity. But how do we make sense of a mass of unstructured data? How do we work out what’s important? And how do we know the data is credible?

How it works

1. Data scraping



A user signs in to their social or online accounts.
We gather the raw data.

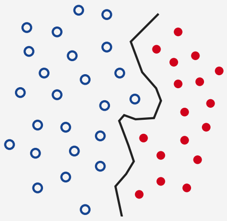
2. Data cleansing

We remove inconsistent or missing information, ensuring we start with high quality data.

3. Feature extraction

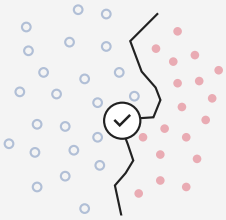
We compute features that serve as structured input to our machine learning models.

4. Model training



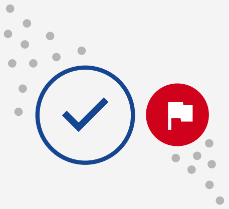
We build learning models to assess the credibility of a users identity.

5. Model evaluation



The learning model is put to the test to ensure it is learning as we expect.

6. Binary decision



Finally we calculate a credibility probability, which is then turned into a binary yes/no decision.

Why machine learning?

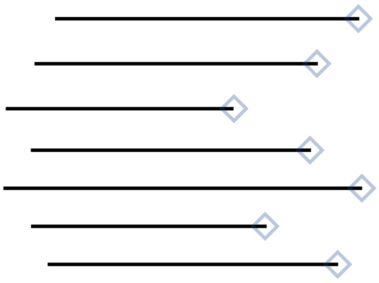
Machine learning offers significant benefits over a purely rules-based approach. Once our models are trained using controlled data, they continue to learn and evolve as they process user information in a real-life environment.

This continual evolvement of our models is especially critical given the increasing sophistication of fraudsters, who tirelessly develop new ways to circumvent existing technology. Rather than playing catch up, as is the case with a purely rules-based approach, our models remain several steps ahead by evolving as new fraudulent practices emerge.

Training the machine

When training our machine learning models with sample data we go through a five-step process:

1. Data scraping

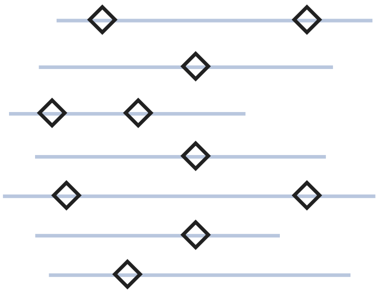


When a user verifies their identity with Veridu, they are asked to sign in to one or more of their social or online accounts and to grant us permission to access the data contained within these. We instantly scrape, or gather, this data using APIs.

> *Veridu puts the user in control of the data they share, providing us with access to data beyond what's available on public profiles.*

It's important to note that our consent-driven approach means that the information we collect goes much deeper than what appears on a user's public profile. Depending on the accounts a user has verified themselves with we collect information including their posts, comments, activities, playlists, location, and more.

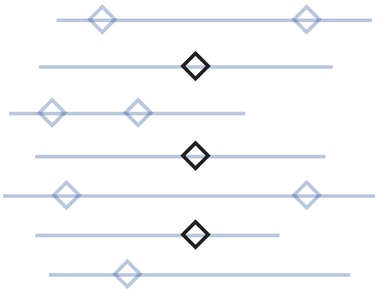
2. Data cleansing



Following the data scraping step we're left with a mass of raw, unstructured data that we need to make sense of. We do this by cleansing the data and structuring it into a standard format, turning the raw data into facts like 'first name', 'last name', 'gender', 'location', and 'age'.

When it comes to machine learning, if you put rubbish in you'll get rubbish out, and this step ensures we never get into that situation. By identifying and marking any missing facts (for example, perhaps none of the accounts used by a user to verify themselves contained a date of birth), we train our models using only the highest quality data.

3. Feature extraction



This is one of the most critical aspects of a machine learning pipeline. In our scenario, we obtain raw data from a user's social and online accounts, which is made up of a mass of unstructured textual and image data, such as a user's posts, comments, likes, images, tags, and more.

> *Using online activity including social media posts, images and comments Veridu can instantly verify a user's identity.*

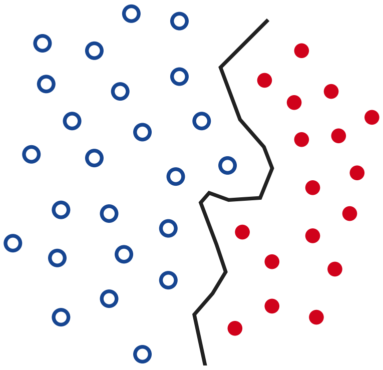
During this learning step we compute features that serve as structured input to our models. These features can be composed of numerical, categorical and binary values, and can be described as an individual measurable property of a phenomenon being observed[1].

Let's take the example of a user's name. First, we scrape a user's online accounts, before cleansing and structuring the data into granular facts, such as 'First name on Facebook' and 'First name on Google'. At this point, however, we have no way of knowing if the user is actually a fraudster using fake details to commit their crime.

To assess the likelihood that this is actually the user's real name, we train our model using the set of features we have computed. One example of a binary feature is "does the first name on Facebook match the first name on Google?". An example of a numeric-based feature around name is the number of comments a user has received on his posts which mention his name.

These are just simple examples of features. Veridu computes hundreds of other features from the raw data we collect, such as how active a profile is, how information from different sources correlate with each other, and many more. All of these features are then used during our model training step.

4. Model training

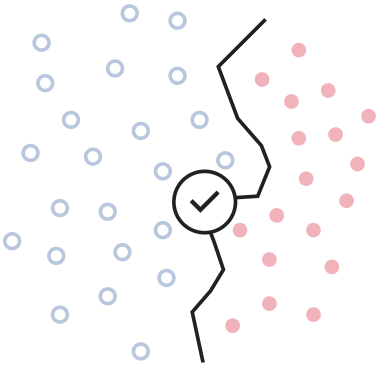


This is where things get really interesting. We now build a model to assess our confidence in the user's credibility, based on the known real and fake data samples we have collected. We use a set of learning models to do this, one of the main ones being neural networks.

> *A neural network is a model inspired by the way our brains work, with a set of connected neurons that fire when a particular input is received.*

Neural networks explore this idea by building a set of connected neurons which are trained via a mathematical optimisation technique called gradient descent. By using a set of training data containing known real users and known fake users, we can tell the model when it's right, and when it's wrong, allowing it to continually learn to reduce any output error and become highly accurate.

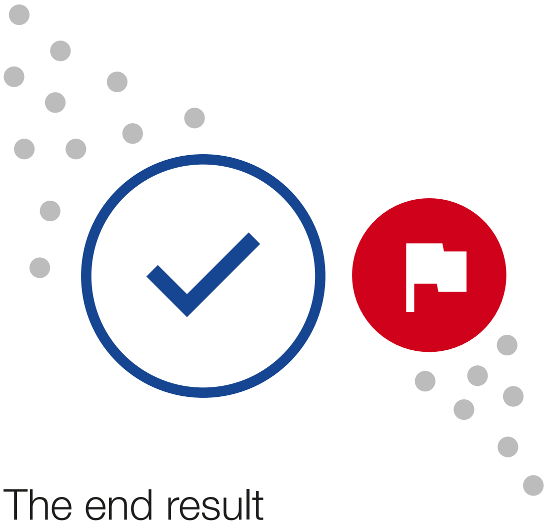
[1] *https://en.wikipedia.org/wiki/Feature_(machine_learning)*



5. Model evaluation

The final step is to put our model to the test to ensure it is learning as we expect. We do this using a statistical technique called cross validation.

Cross validation is the process of training our model on one subset of data, and then testing it using a different subset. This process is repeated many times using different subsets of data for training and testing. By doing this, we can be sure that our models are smart enough to accurately analyse all new data coming in. If we're happy with the results of the evaluation step, we have our final model.



The end result

When a user verifies with Veridu, we use our trained models to calculate a credibility probability, which is then turned into a binary decision - "is this profile fake or real"?

Learn more
www.veridu.com
sales@veridu.com