Truffles and Champignons:
How to Find Cancer Cells with Machine Learning?

The white blood cells in our blood stream tell stories about how our bodies are doing. Hairy cells are cancerous cells indicating leukemia. How do we find such cells using contrastive machine learning?

You've heard about self-driving cars that can "see" the traffic. Did you know that the same kind of technology can be used to detect blood cancer cells, such as hairy cells, and help doctors diagnose and treat patients? This tool is called machine learning. To be specific, we compared a method called transfer learning and another called contrastive learning and got 88 % and 90 % correctly predicted cells, respectively. Now, machine learning might sound scary, but remember, we want to train machines to listen to us and to make our lives happier. So in a way, machines are like puppies (but much less fluffy and bark in 1's and 0's). For the time being, let's replace "machine learning" with "puppy learning".

Harry is a very cute puppy and you're eager to show off to your friends that Harry can find a truffle (the cells we want to find) among a bunch of champignons (other cells). How do you teach him? Puppies aren't inherently crazy about mushrooms, but they do love treats. You let Harry smell a truffle and a champignon. When Harry puts his little snoot on a truffle, you yell "Good boy!" and reward him with a yummy treat! The same thing happens inside a machine learning model. Every time it's correct, the behavior is rewarded, and if it makes the wrong decision, we'll guide it towards making a better decision next time. You continue with the next pair of fungi. As the training session progresses, Harry is more and more eager to boop the truffle.

Next day, you wonder, would it be easier for Harry if he got to see many fungi at once? You present Harry with 5 champignons and 5 truffle at once. As the goodest boy in the world, Harry doesn't disappoint you and quickly separates all the truffles from the set. This is the essence of contrastive learning. The similar smells of the truffles make them more recognizable as a group.

Harry gets better at the game each day, but during one session, he picked the wrong mushrooms. No treat was served and Harry looks at you with pleading puppy eyes. Is there anything wrong with the mushrooms? Suddenly, it struck you that when packing the mushrooms in a hurry yesterday, you put both kinds in the same plastic bag. Their scents have gotten mixed up! Dogs don't see as good as humans do and Harry has been relying on his nose to tell them apart. Of course he can't find the right ones now!

What happened here is similar to using wrongly labeled data for training. Harry's nose can't tell a truffle from a champignon if they smell the same. Likewise, a machine learning network also cannot make the right choices if it has been fed poor quality data.

After discovering the mistake you store the mushrooms apart and, to your relief, Harry excels at the fungi game the next day. Now your training can continue. Hopefully you and Harry can go and win truffle searching competitions together! Oh, and for us? We'll continue finding new ways to make data and algorithms work for us and help doctors discover illnesses.