9 DATA SCIENCE AND AI

PJD's Opening Remarks

In the past 8 lectures, we've looked at four categories of AI machines – rule-based, supervised learning, unsupervised learning, and human-machine teaming. We also looked at possible future machines AI researchers are working to build.

For the next 7 lectures, we're going to look at important application areas and see what has been accomplished. We'll begin today with data science.

You have already noticed that the existing machines require a LOT of data to make them work. Expert systems may have millions of rules in their databases. Image classifiers may need 10 million labeled images to train them. Game machines may need to simulate 100 million games to beat the best human players.

Only in recent years have we had the sensors and the storage to gather and hold that much data. And the computing power to process them into the learning function of the machine.

But there is a big problem. How can we trust the data we're using to train our machines? Where do we get 10 million labeled images from? Who labels them? How do we know that the labelers have the required expertise to make proper labels?

We also know that biases in the training data are likely to show up as biases in the function of the machine after it is trained. Is there anything we can do to remove the bias?

Unsupervised learning tries to get around the trust issue by learning from the data themselves without external supervision. Thirty years ago, one of my research colleagues, Peter Cheeseman, invented a program called AUTOCLASS that grouped data into similarity classes – without any external input other than the data. Peter put it to the test by classifying the 5400 objects detected in the NASA infrared sky survey. AUTOCLASS found exactly the classes already found by astronomers – plus one new class. AUTOCLASS not only agreed with the expertise of astronomers, it made a new discovery. AUTOCLASS relied on an advanced statistical method called Bayesian Learning. What else can advanced statistical methods do for us?

Answering questions like these – trusting large data sets, using them to train machines, evaluating the reliability of the results, extracting patterns, employing advanced statistical methods, and building models – are among the main concerns of data science.

Today's speaker is Major Ross Schuchard of the US Army. He has degrees in economics and social science and just got his PhD from George Mason University in computational social science. He was an Army aviation officer for 7 years and operations research officer for 8. He established the first data science cell at Army Cyber Command and worked recently on data science within DOD's new Army Futures Command. He just arrived at NPS this quarter and is a member of the Data Science Advisory Group.