bbc.com

# Machine learning 'causing science crisis'

*By Pallab Ghosh Science correspondent, BBC News, Washington*

5-6 minutes

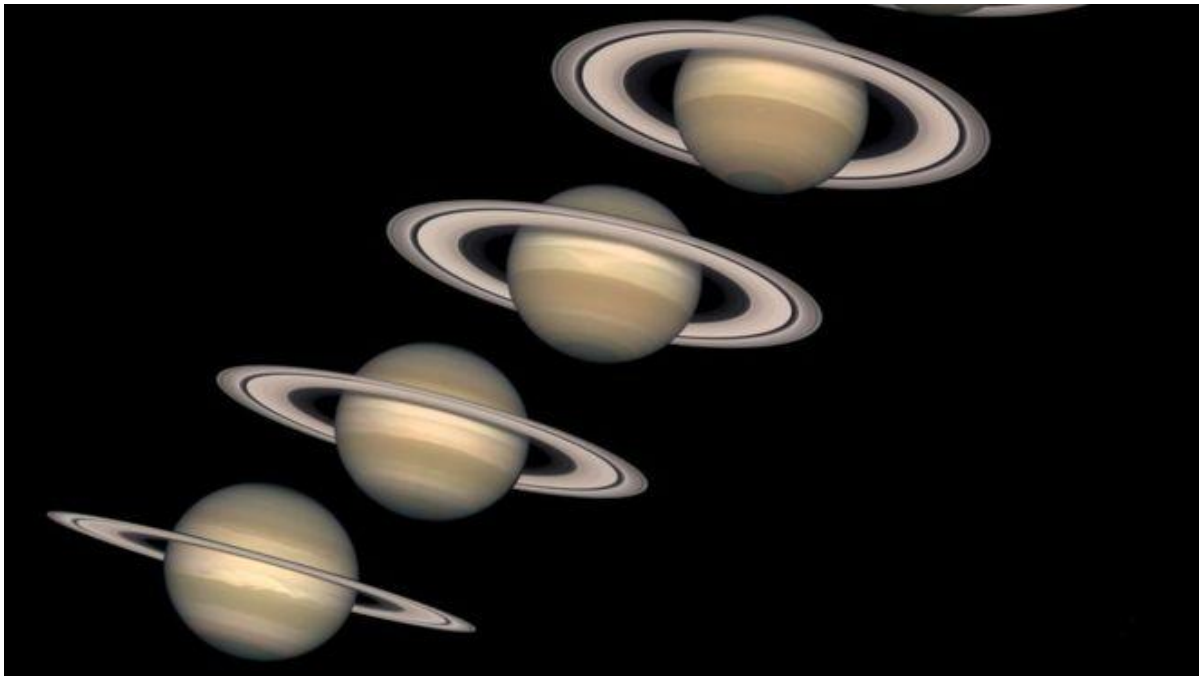

Image copyright Reuters

Image caption Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

Dr Genevera Allen from Rice University in Houston said that the increased use of such systems was contributing to a “crisis in science”.

She warned scientists that if they didn't improve their techniques they would be wasting both time and money. Her research was presented at the American Association for the Advancement of Science in Washington.

A growing amount of scientific research involves using machine learning software to analyse data that has already been collected. This happens across many subject areas ranging from biomedical research to astronomy. The data sets are very large and expensive.

## 'Reproducibility crisis'

But, according to Dr Allen, the answers they come up with are likely to be inaccurate or wrong because the software is identifying patterns that exist only in that data set and not the real world.

Image copyright Rice University
Image caption Dr Allen says flawed machine learning is producing a "crisis in science"

"Often these studies are not found out to be inaccurate until there's another real big dataset that someone applies these techniques to

and says ‘oh my goodness, the results of these two studies don’t overlap‘," she said.

“There is general recognition of a reproducibility crisis in science right now. I would venture to argue that a huge part of that does come from the use of machine learning techniques in science.”

The “reproducibility crisis” in science refers to the alarming number of research results that are not repeated when another group of scientists tries the same experiment. It can mean that the initial results were wrong. One analysis suggested that up to 85% of all biomedical research carried out in the world is wasted effort.

## More from Pallab at the AAAS:

It is a crisis that has been growing for two decades and has come about because experiments are not designed well enough to ensure that the scientists don’t fool themselves and see what they want to see in the results.

## Flawed patterns

Machine learning systems and the use of big data sets has accelerated the crisis, according to Dr Allen. That is because machine learning algorithms have been developed specifically to find interesting things in datasets and so when they search through huge amounts of data they will inevitably find a pattern.

“The challenge is can we really trust those findings?” she told BBC News.

“Are those really true discoveries that really represent science? Are they reproducible? If we had an additional dataset would we see

the same scientific discovery or principle on the same dataset? And unfortunately the answer is often probably not."


Image copyright Reuters
Image caption Machine learning is also used in biomedical research

Dr Allen is working with a group of biomedical researchers at Baylor College of Medicine in Houston to improve the reliability of their results. She is developing the next generation of machine learning and statistical techniques that can not only sift through large amounts of data to make discoveries, but also report how uncertain their results are and their likely reproducibility.

"Collecting these huge data sets is incredibly expensive. And I tell the scientists that I work with that it might take you longer to get published, but in the end your results are going to stand the test of time.

"It will save scientists money and it's also important to advance science by not going down all of these wrong possible directions."

Follow Pallab on Twitter

bbc.com

# Machine learning 'causing science crisis'

*By Pallab Ghosh Science correspondent, BBC News, Washington*

5-6 minutes

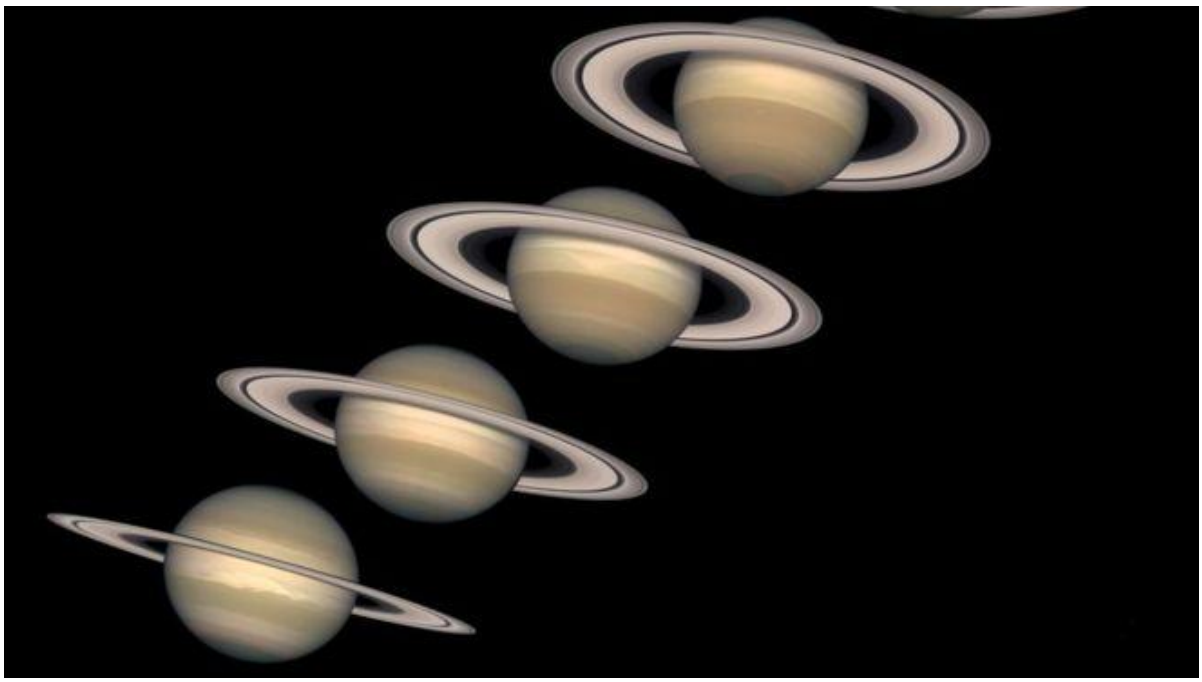

Image copyright Reuters

Image caption Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

Dr Genevera Allen from Rice University in Houston said that the increased use of such systems was contributing to a “crisis in science”.

She warned scientists that if they didn't improve their techniques they would be wasting both time and money. Her research was presented at the American Association for the Advancement of Science in Washington.

A growing amount of scientific research involves using machine learning software to analyse data that has already been collected. This happens across many subject areas ranging from biomedical research to astronomy. The data sets are very large and expensive.

## 'Reproducibility crisis'

But, according to Dr Allen, the answers they come up with are likely to be inaccurate or wrong because the software is identifying patterns that exist only in that data set and not the real world.

Image copyright Rice University
Image caption Dr Allen says flawed machine learning is producing a "crisis in science"

"Often these studies are not found out to be inaccurate until there's another real big dataset that someone applies these techniques to

and says ‘oh my goodness, the results of these two studies don’t overlap‘," she said.

“There is general recognition of a reproducibility crisis in science right now. I would venture to argue that a huge part of that does come from the use of machine learning techniques in science.”

The “reproducibility crisis” in science refers to the alarming number of research results that are not repeated when another group of scientists tries the same experiment. It can mean that the initial results were wrong. One analysis suggested that up to 85% of all biomedical research carried out in the world is wasted effort.

## More from Pallab at the AAAS:

It is a crisis that has been growing for two decades and has come about because experiments are not designed well enough to ensure that the scientists don't fool themselves and see what they want to see in the results.

## Flawed patterns

Machine learning systems and the use of big data sets has accelerated the crisis, according to Dr Allen. That is because machine learning algorithms have been developed specifically to find interesting things in datasets and so when they search through huge amounts of data they will inevitably find a pattern.

"The challenge is can we really trust those findings?" she told BBC News.

"Are those really true discoveries that really represent science? Are they reproducible? If we had an additional dataset would we see the same scientific discovery or principle on the same dataset? And unfortunately the answer is often probably not."

Image copyright Reuters
Image caption Machine learning is also used in biomedical research

Dr Allen is working with a group of biomedical researchers at Baylor College of Medicine in Houston to improve the reliability of their results. She is developing the next generation of machine learning and statistical techniques that can not only sift through large amounts of data to make discoveries, but also report how uncertain their results are and their likely reproducibility.

“Collecting these huge data sets is incredibly expensive. And I tell the scientists that I work with that it might take you longer to get published, but in the end your results are going to stand the test of time.

“It will save scientists money and it’s also important to advance science by not going down all of these wrong possible directions.”

Follow Pallab on Twitter