There was a time when people believed the earth was flat. Or that humans would never walk on the moon. Or that computers had no place in the law. But then the non-believers proved them wrong. The earth is round, men have walked on the moon, and it is hard to imagine practicing law without a computer.

What about technology-assisted review? Are there myths surrounding TAR that will fall by the wayside as we better understand the process? Will we look back and smile at what people believed about TAR way back then? Turns out, that is already happening. Here are five myths that early TAR adopters believed true but that modern TAR systems prove wrong.

1. You only get one bite at the apple.

One early myth about TAR was that you would run it just once and that was the end of it. This myth grew out of the first TAR processes (TAR 1.0), which required an initial seed set of documents selected at random from the total population. A subject matter expert (usually one senior lawyer) tagged each seed document as relevant or irrelevant. The expert’s tags were then used to “train” the system. Eventually, after reviewing a few thousand documents, the expert could stop. The system would get no better; it was as well trained about your documents as it could be.

With the training complete, a review administrator applied the TAR algorithm to the rest of the document population. The system ranked the unviewed documents in relevance order. Depending on the effectiveness of the ranking, the administrator set a “cutoff” point to govern the review. Documents ranked higher than the cutoff were reviewed and tagged. Documents below the cutoff were discarded (after confirmatory sampling).

Under this approach, the TAR process was static and run once at the beginning. As reviewers progressed through the documents, there was no easy way to feed their findings back into the system to improve the ranking even further. The myth was that “one bite at the apple” was all you could get.

TAR 2.0 systems let you keep biting away, thanks to their capacity for continuous learning. Now, reviewers are given the next most likely relevant documents for consideration. As they tag the documents (either relevant or not), that information is fed back to the system. As it is, the system gets smarter and smarter about your document population.

2. Subject matter experts are required for TAR training.

Another myth of TAR 1.0 was that only a subject matter expert can do the training. Although the expert didn’t have to be a lawyer, it did have to be someone senior in the field who would know how the documents should be classified. Underlying this myth was the fear that, without an expert, inconsistency in training would degrade the algorithm’s effectiveness. That would lead to more documents falling above the cutoff and thus require more expensive human review. Recent evidence suggests this is wrong. First, these senior experts are not always consistent in their tagging. People are fallible. Document review can be mind numbing. On one day, you tag them one way; on another, the opposite.

Second, review teams, while not perfect, turn out to do a pretty good job of tagging documents for training. This is particularly true because most TAR 2.0 systems take this natural variation into account. They can also present outliers to an expert for correction as part of a quality control process. Using reviewers to train the system makes the review cheaper (experts typically bill at higher
Rates. It also means review can start right away, without the delay of waiting for the busy expert to focus on the review and complete the initial training. Most senior attorneys I know feel they have better things to do than TAR training in any event.

3. You must train on randomly selected documents.

Many TAR proponents believe that you need to train the system at least initially using documents selected randomly from the review population. If you select training documents by other means (key word searching, for example), you may bias the training, they argue. Their fear is that you will unwittingly place undue emphasis on documents you think are relevant while ignoring others that might be equally relevant. “You don’t know what you don’t know,” they say. TAR systems following this approach present the subject matter expert with randomly selected documents for training. This may be tolerable when there are a reasonable number of relevant documents in the population, often called richness. But it can drive you crazy when the population is not rich. You have to click through hundreds if not thousands of documents before you find relevant ones for training.

Modern TAR systems prove this to be a myth. They allow and encourage you to submit as many documents as you can find for training, regardless of how you find them. You supplement this training with documents you don’t know about. They can be selected through some form of diversity sampling (specifically, to find documents you know the least about), systematic sampling (sampling every Nth document from top to bottom) or even simple random sampling as a supplement but not the main course. The more relevant documents you can find for training, the better the results. Clicking through thousands of random documents is boring and not required for a good TAR result.

4. You can’t start TAR training until you have all your documents.

One of the bugaboos of TAR 1.0 was the requirement that you collect all documents before beginning training. Early systems required this because they trained against a control set rather than against all of the documents. These systems lacked the horsepower to rank all of the documents for each training round. In order for the control set to be valid, it had to be selected randomly from all of the documents being referenced. If you received additional documents the next week, this created a problem. The addition of new documents in the population meant the control set was no longer valid. It was no longer representative of the larger set.

In the real world of litigation, where collections were ongoing, this meant that training had to be redone each time new collections arrived. For review administrators, this represented an impossible burden. They did not have the luxury of waiting until all the documents were collected or of conducting new rounds of training each time new documents were found. TAR 2.0 systems have made this a myth. With the capacity to handle “big data,” they rank all of the documents each time and don’t use a control set to determine the effectiveness of each ranking.

As a result, new documents can be added continually as they are collected. The new documents may require a few added rounds of training but the process no longer has to start from scratch. They are simply added to the mix and ranked along with the others.

5. TAR doesn’t work for non-English documents.

Many early TAR users believed that the process worked only on English documents. They assumed that TAR systems “understood” the words and concepts in documents. That being the case, there was no way it could understand other languages. This, too, was a myth. TAR is a mathematical process that ranks documents based on word frequency. It has no idea what the words mean. If the documents are prepared properly, TAR can be just as effective with any language as it can with English. For some languages—such as Chinese, Japanese and Korean—this requires that the text is first broken into individual word segments, a process also called tokenizing. Many TAR 1.0 systems did not have tokenizing engines. Many TAR 2.0 systems are able to tokenize as long as your trainers understand the documents and can tag them properly, TAR should be just as effective with non-English documents as with English ones.

Myths help us understand our world.

Myths evolved to help us make sense of things that were beyond our comprehension. We created myths about the sun being drawn by chariots or the moon being made of green cheese. Myths helped us get started in understanding our solar system. As we learn more, myths get replaced by facts, which help us to better navigate our world. As we learn more about TAR and the cost-saving benefits it can provide, many of the initial myths about how it worked have fallen away too.

Turns out, the moon is not made of green cheese, nor is the sun drawn by chariots. And TAR is far more versatile and adaptable than early adopters believed.

John Tredennick is the president and CEO of Catalyst Repository Systems, based in Denver.

Reprinted with permission from the January 2014 edition of Law Technology News. © 2014 ALM Media Properties, LLC. All rights reserved. Further duplication without permission is prohibited. For information, contact 877-257-3382 or reprints@alm.com. #010-01-14-01