Ground-truth data cannot do it alone

Verifying automated analysis methods via ground-truth data remains an essential step of algorithm development. But as datasets increase in size and complexity, this classical test is often insufficient. Integrated editing tools can help.

Researchers using satellite imaging to remotely observe features on the Earth enjoy the luxury of a simple solution for verifying the interpretation of their data with the truth on the ground or 'ground truth'. They or a surrogate can go observe it firsthand.

In contrast, researchers using an algorithm to analyze data on complex biological phenomena rarely have the luxury of a straightforward ground truth. So what can one do? For algorithm developers, the first solution is to create synthetic biological data and run that through the algorithm, similar to how developers of image-analysis and signal-processing algorithms have historically tested their methods. This works well for structured data such as cell shape, metagenome sequences or particle trajectories. The next solution, appropriate for developers and users, is the use of curated or 'gold-standard' datasets that allow testing on more complex data. Finally, a reasonable approximation of the actual ground truth of an experiment can often be obtained for verification by incorporating data on the same system from alternative protocols (Peng *et al. Nat. Methods* **8**, 493–500; 2011).

Biologists often make do without the use of a ground truth in the classical sense and instead rely on appropriate controls and validation experiments. This is sufficient for studies that do not require automation, but as studies increase in size and complexity, more data analysis is dependent on computational methods. It is essential that the performance of these automated methods be evaluated using synthetic data, gold-standard datasets and ground-truth data.

But biological complexity and variability create many different 'truths' and challenge the reliability of ground-truth testing. An automated analysis method that performs superbly on one or more datasets in one laboratory may fail in another using different systems or protocols. Developers of analysis software can and should approximate this by testing their algorithms on a variety of ground-truth datasets.

In addition to better defining and using ground-truth datasets, a good way to deal with biological and experimental variability when using automated analysis methods is the integration of efficient proofreading and editing tools. Although this will be less ideal for some data types than others, in the case of visual data, the human eye remains unmatched for evaluating and classifying images and is therefore highly suitable for proofreading this kind of data.

Biologists have such faith in their own eyes that they will often trust them over a fully automated solution even when presented with data clearly showing that the automated solution performs better. Proofreading and editing tools integrated into automated systems provide the assurance of being able to evaluate the output of the system and correct mistakes and thus lower the adoption barrier to automated methods. But because such editing negates one of the chief benefits of automated systems by allowing the investigator to alter the results and removing the ability of others to faithfully replicate the results, some form of Good Laboratory Practice guidelines should also be implemented.

Objective data are crucial for evaluating the effects of drugs and chemicals and government regulatory agencies created Good Laboratory Practice guidelines to ensure the integrity of data produced in labs under their jurisdiction. These include guidelines for handling data from automated instruments, which state that if somebody makes a correction to an automated result there needs to be a reason for it, and both the edit and its justification must be robustly documented. This could be a database entry that cannot be erased but that can be rolled back if an audit requires it. As an example, the ability to efficiently proofread automated image analysis tasks and log edits has been implemented in the Farsight image analysis toolkit (http://www.farsight-toolkit.org/) (Luisi *et al. Neuroinformatics* **9**, 305–315; 2010).

Documented editing is not only useful for the user but provides valuable feedback to the developer who can use it to determine where the algorithm is breaking down and correct it. When dealing with high-throughput data, it is possible to include statistical analysis in the proofreading that allows a level of certainty to be calculated. Editing and the use of statistics also make it possible to sort the errors into specific types, allowing easy global changes and simplifying algorithm redesign and refinement. Ultimately, it may be possible to automate the proof-editing process itself (Peng *et al. Neuroinformatics* **9**, 103–105; 2010).

Development of automated analysis tools necessary to meet the changing research landscape depends more than ever on active communication and collaboration between developers of automated analysis tools and their users. This must include not only a discussion of the role and use of ground-truth data but methods for users to evaluate and improve the performance of automated methods and obtain statistical measures of confidence in the results when applied to their own data.