Correspondence

ArXiv screens spot fake papers

Unlike the computer-generated nonsense papers in some peer-reviewed subscription services (see *Nature* http://doi.org/r3n; 2014), the 500 or so preprints received daily by the automated repository arXiv are not pre-screened by humans. But sometimes automated assessment can be better than human diligence at enforcing standards.

The automated screens for outliers in arXiv include analysis of the probability distributions of words and their combinations, ensuring that they fall into patterns that are consistent with existing subject classes. This serves as a check of the subject categorizations provided by submitters, and helps to detect non-research content.

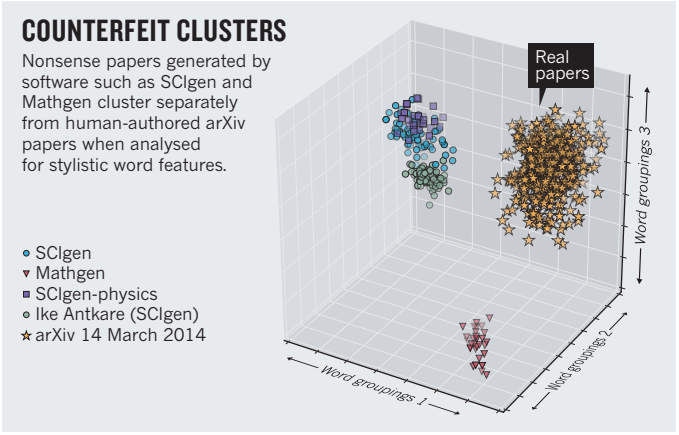
Fake papers generated by SCIgen software, for example, have a 'native dialect' that can be picked up by simple stylometric analysis (see J. N. G. Binongo *Chance* **16,** 9–17; 2003). The most frequent words used in English text (stop words such as 'the', 'of', 'and') encode stylistic features that are independent of content. On average, these words follow a power-law distribution that is evident in even relatively small amounts of text; significant deviations signal outliers.

The effect can be seen in principal-component analysis plots (see 'Counterfeit clusters'). Computer-generated articles form tight clusters that are well separated from human-authored articles.

Paul Ginsparg *Cornell University, Ithaca, New York, USA.*
ginsparg@cornell.edu

COUNTERFEIT CLUSTERS

Nonsense papers generated by software such as SCIgen and Mathgen cluster separately from human-authored arXiv papers when analysed for stylistic word features.



Free up systems for funding and advice

As president of the New Zealand Association of Scientists, I endorse Peter Gluckman's principles for effective science advice to government (*Nature* **507,** 163–165; 2014). As he remarks, however, science advisers may encounter a conflict of interest if they are involved in administering public research funding.

Gluckman is the New Zealand Prime Minister's chief science adviser and chaired the panel that last year selected the National Science Challenges. He has been instrumental in publicizing and defending the new funding mechanism for meeting these goals (see go.nature.com/cmgkx1), which the government has signalled are likely to set the default funding strategy for New Zealand science in the next decade and beyond (see, for example, go.nature.com/srrtym).

The community of scientists is concerned about the perceived conflict of interest and loss of trust inherent in combining these roles. They are worried that the challenges will shut out excellent science that does not fit with the goals. Another issue is the perception among Maori researchers that the processes for identifying the national challenges have so far marginalized Maori participation.

It is to be hoped that Gluckman's ten principles will help in future to separate science advisory and funding systems, and that the promised National Statement of Science Investment will address the wider (and no less important) research agenda.

Nicola Gaston *New Zealand Association of Scientists, Wellington, New Zealand.*
president@scientists.org.nz

Journals must boost data sharing

The journal ecosystem is a powerful filter of scientific literature, promoting the best work into the best journals. Why not use a similar mechanism to encourage more comprehensive data sharing?

Several journals have introduced policies mandating that data be shared on a public archive at publication (see, for example, go.nature.com/b7u4ed). However, these policies have met with limited success, perhaps because of authors' fears of losing control, being scooped in subsequent papers or having errors exposed. Moreover, compliance with data-sharing policies is typically checked only after the paper has been accepted.

To spur excellence in data sharing, journals must recognize that better sharing leads to stronger papers, and judge submissions accordingly. Articles associated with feeble sharing efforts should either improve or be rejected.

A focus on publishing verifiable research will boost journal reputation. It also signals to the community of authors that withholding data will restrict them to publication in less-prestigious journals.

Timothy H. Vines *University of British Columbia, Vancouver, Canada.*
vines@zoology.ubc.ca

Projects powered by free computing grid

Herman Tse describes the scientific output of IBM's World Community Grid as "lacklustre" (*Nature* **507,** 431; 2014). This is not the case: the 22 projects we have supported so far have generated more than 35 peer-reviewed papers in prominent journals. Our donated computing power has resulted in several important practical scientific advances.

For example, Japan's Chiba Cancer Center used our free computing power to screen three million drug candidates for treating neuroblastoma, a common childhood cancer. This yielded seven promising compounds that have no apparent side effects (Y. Nakamura *et al. Cancer Med.* **3,** 25–35; 2014).

Last June, Harvard University's Clean Energy Project announced some 35,000 organic materials that could double the efficiency of carbon-based solar cells, after using our grid to scan more than two million candidate materials (see J. Hachmann *et al. Energy Environ. Sci.* **7,** 698–704; 2014, and go.nature.com/cxt181).

Neither should Tse underestimate papers that focus "solely on the technical aspect of distributed computing". Such computing accelerates research and underpins scientific advances. Take the 2013 Nobel Prize in Chemistry: it was awarded to three scientists who developed the kind of computer-modelling techniques on which the work of World Community Grid researchers is based. As the Nobel committee noted: "Today the computer is just as important a tool for chemists as the test tube."

Juan Hindo *World Community Grid, Chicago, Illinois, USA.*
juan.hindo@us.ibm.com

CONTRIBUTIONS

For Correspondence author guidelines, see **go.nature.com/cmchno**.