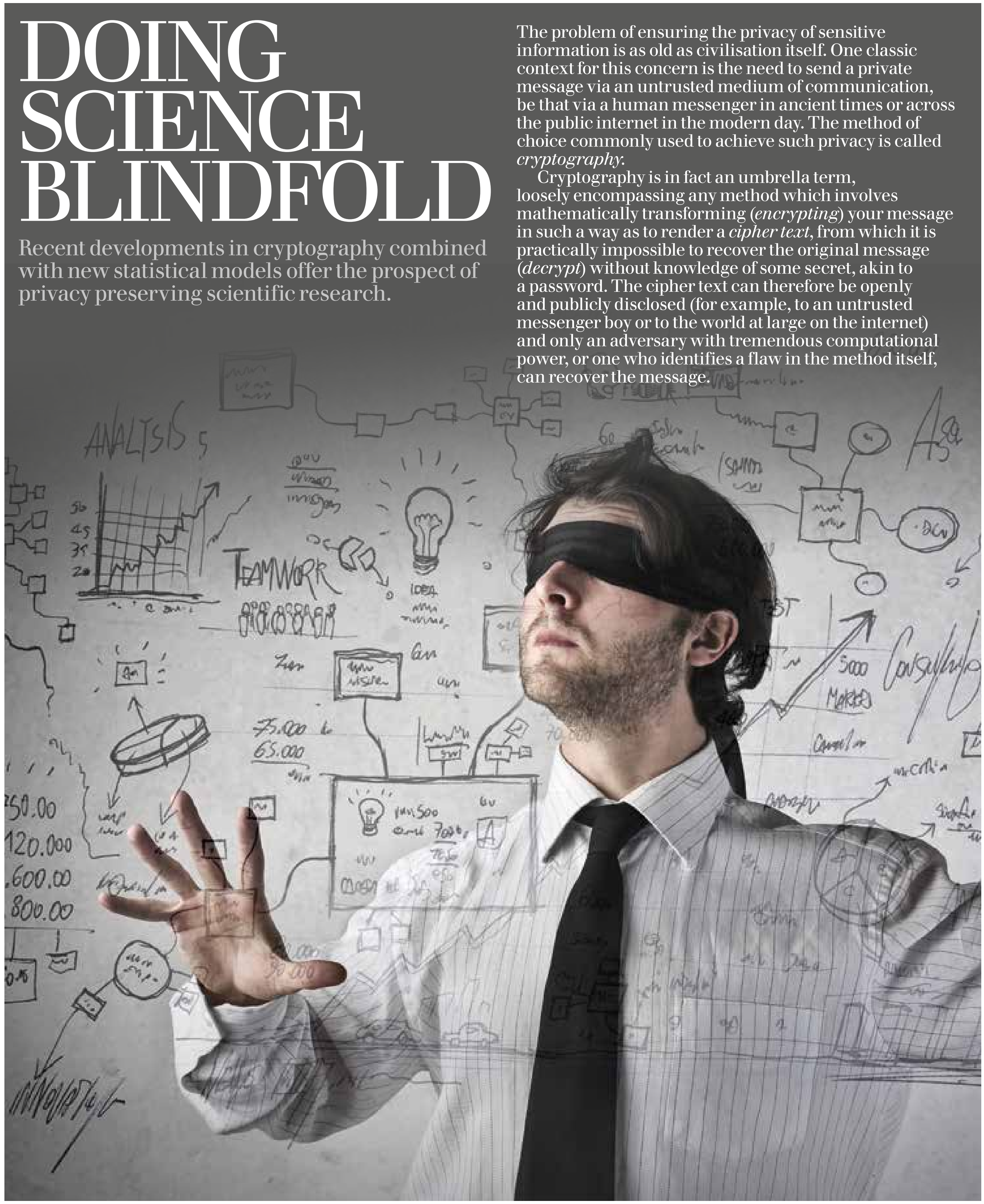
Dr Louis Aslett
Junior Research Fellow, Corpus Christi College, Oxford
Postdoctoral Researcher, Department of Statistics, University of Oxford



©OLLY/FOTOLIA

DOING SCIENCE BLINDFOLD

Recent developments in cryptography combined with new statistical models offer the prospect of privacy preserving scientific research.

The problem of ensuring the privacy of sensitive information is as old as civilisation itself. One classic context for this concern is the need to send a private message via an untrusted medium of communication, be that via a human messenger in ancient times or across the public internet in the modern day. The method of choice commonly used to achieve such privacy is called *cryptography*.

Cryptography is in fact an umbrella term, loosely encompassing any method which involves mathematically transforming (*encrypting*) your message in such a way as to render a *cipher text*, from which it is practically impossible to recover the original message (*decrypt*) without knowledge of some secret, akin to a password. The cipher text can therefore be openly and publicly disclosed (for example, to an untrusted messenger boy or to the world at large on the internet) and only an adversary with tremendous computational power, or one who identifies a flaw in the method itself, can recover the message.

> An exciting new area of research at the interface between cryptography and statistics which has potential for significant impact in both enabling research and ensuring data privacy.

Although the origins of cryptography lie in this secret communication problem, there are manifold uses, for example simply long term storage of sensitive information: once encrypted there is no need to store it in a safe or bank vault, since the cipher text is itself a mathematical vault. However, this secret storage problem raises an interesting additional question: what if one wishes to manipulate the information stored therein?

To make this concrete with an everyday example, imagine one encrypts the deposits and expenditures on a bank account and after some time the balance is required. The crux of the problem is that all the different approaches to cryptography have shared the common trait of being 'brittle' in the sense that the information within a cipher text cannot be manipulated without first decrypting. Broadly speaking, any attempt to modify, combine or transform cipher texts usually renders nonsense upon decryption – one can only recover exactly what was put into the mathematical vault in the first place.

Returning to the fictional bank account problem, the brittleness of cryptographic methods seem to imply one must first decrypt all of the information on the account and risk disclosure in order to simply ascertain the balance. It would be gratifying if one could instead add up the balance by combining the cipher texts directly and simply decrypt the final result.

In 1979 it was postulated that it may be possible to develop methods of cryptography whereby it was indeed possible to manipulate the information held within a cipher text without compromising privacy. More specifically, the postulate was that arbitrary addition and multiplication of numbers held in the cipher text may be possible so that one ends up with an encrypted version of the answer. In other words, the fictional bank balance could be computed without having to decrypt all the transactions: we simply perform special operations on the the cipher text which correspond to performing addition of the information held within.

Quite literally, if one were to encrypt the numbers 2 and 3, then take these two cipher texts and perform a special 'addition' operation on them, then decryption of the result would render 5. We have computed the sum blindfold because when operating on the cipher texts we had no idea of their true secret values.

It was not until 2009 in a remarkable PhD thesis by Craig Gentry at Stanford University that the 1979 postulate was proved true and a method developed. Any cryptography algorithm which can achieve this (and several now exist since 2009) is called a *fully homomorphic encryption* scheme.

Statistics and homomorphic encryption

Of course, computing bank account balances is a rather mundane and over simplified use for such a remarkable result (and technically can be achieved without *full* homomorphism). However, there are in fact many limitations, some of them rather technical. Applied cryptographers have started publishing applications of these schemes to scientific models where the limitations permit computation, but this restricts the field of use substantially. As a statistician, I am interested in coming from the opposite direction and developing new models which are explicitly designed with the limitations of homomorphic encryption in mind and quantifying the uncertainty in these more approximate methods: statistics is essentially the mathematics of uncertainty.

NICK READ

ABOVE: Dr Louis Aslett

One example area is of importance to both academics and industry. The extensive use of private and personally identifiable information in biomedical applications can make the general public reticent to contribute their data and so impede medical research: a 2009 study found 90% of people were uneasy about so-called 'biobanks' which seek to amass patient data. However, if cryptographically strong guarantees can be provided that their data will not be seen – only the results of models fitted, with the data remaining encrypted throughout – then people may be more willing to engage with researchers and more data will be available. But we need new models which work within the confines of these homomorphic schemes. Our research has already shown that even some very approximate models which are amenable to homomorphic encryption will outperform the traditional models when they have more data available.

Indeed it is not just in academia: industry itself is on the brink on embarking on biomedical applications on a mind boggling scale never before witnessed as a result of the impending wave of 'wearable devices' such as smart watches, which present serious privacy concerns. Companies hope to market the ability to monitor and track vital health signs round the clock, perhaps fitting biomedical models to alert on different health concerns. However, they will almost certainly leverage 'cloud' services, uploading reams of private health diagnostics to corporate servers. Statistical methods developed for homomorphic encryption could allow individual privacy to be preserved, whilst still enabling industry to incorporate such data into statistical analyses.

This represents an exciting new area of research at the interface between cryptography and statistics which has potential for significant impact in both enabling research and ensuring data privacy.