

# Circular analysis in systems neuroscience: the dangers of double dipping

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick S F Bellgowan & Chris I Baker

**A neuroscientific experiment typically generates a large amount of data, of which only a small fraction is analyzed in detail and presented in a publication. However, selection among noisy measurements can render circular an otherwise appropriate analysis and invalidate results. Here we argue that systems neuroscience needs to adjust some widespread practices to avoid the circularity that can arise from selection. In particular, 'double dipping', the use of the same dataset for selection and selective analysis, will give distorted descriptive statistics and invalid statistical inference whenever the results statistics are not inherently independent of the selection criteria under the null hypothesis. To demonstrate the problem, we apply widely used analyses to noise data known to not contain the experimental effects in question. Spurious effects can appear in the context of both univariate activation analysis and multivariate pattern-information analysis. We suggest a policy for avoiding circularity.**

"Show me the data," we say. But we don't mean it. Instead of the numbers generated by measurement, which can be billions for a single experiment, we wish to see results. This frequent confusion illustrates an important point. We think of the results as reflecting the data so closely that we can disregard the distinction. However, interposed between data and results is analysis, and analysis is often complex and always based on assumptions (Fig. 1a).

Ideally, the results reflect some aspect of the data without any distortion being caused by the assumptions or hypotheses (Fig. 1a). Consider the hypothesis that neuronal responses in a particular region reflect the difference between two experimental stimuli. We might measure the neuronal responses, average across repetitions and present the results in a bar graph with one bar for the response to each stimulus. The set of stimuli (or, more generally, the experimental conditions) is decided on the basis of assumptions and hypotheses, thus determining which bars are shown. But the results themselves (that is, the heights of the two bars) are supposed to reflect the data without any effect of assumptions or hypotheses.

Untangling how data and assumptions influence neuroscientific analyses sometimes reveals that assumptions predetermine results to some extent<sup>1-5</sup>. When the data are altogether lost in the process, the

analysis is completely circular (Fig. 1a). More frequently, in practice, the results do reflect the data, but are distorted to varying degrees by the assumptions (Fig. 1a). Such distortions can arise when the data are first analyzed to select a subset and then the subset is reanalyzed to obtain the results. In this context, assumptions and hypotheses determine the selection criterion and selection can, in turn, distort the results.

In neuroimaging, an example of selection is the definition of a region of interest (ROI) by means of a statistical mapping that highlights voxels that are more strongly active during one condition than another. In single-cell recording, an example of selection is the restriction of in-depth analysis to neurons with certain response properties. In electro- and magnetoencephalography, an example of selection is the restriction to a subset of sensors or sources that show expected responses. In gene microarray studies, an example of selection is inferential analysis performed for a statistically selected subset of genes<sup>6</sup>. In behavioral studies, an example of selection is the division of a group of subjects into subgroups on the basis of task performance. Weighting and sorting of data can be construed as variants of selection, and we will use the latter term in a general sense to refer to all three (Fig. 1b).

Selection can entail two distinct forms of bias: selective reporting of accurate results and distortion of estimates and invalidation of statistical tests. Both forms deserve a wider debate, but we focus on the latter here.

If selection were determined only by true effects in the data, then there would be no distortion of the results of the selective analysis. However, data are always a composite of true effects and noise. Selection is therefore affected by noise. In neuroimaging, for example, the voxels included at the fringe of an ROI tend to reflect the noise to some extent, even if the ROI highlights a truly active brain region (as in Example 2, see below). When the selection process is based on the design matrix, it creates spurious dependencies between the noise in the selected data and the experimental design, thus violating the assumption of random sampling. This can bias selective analysis.

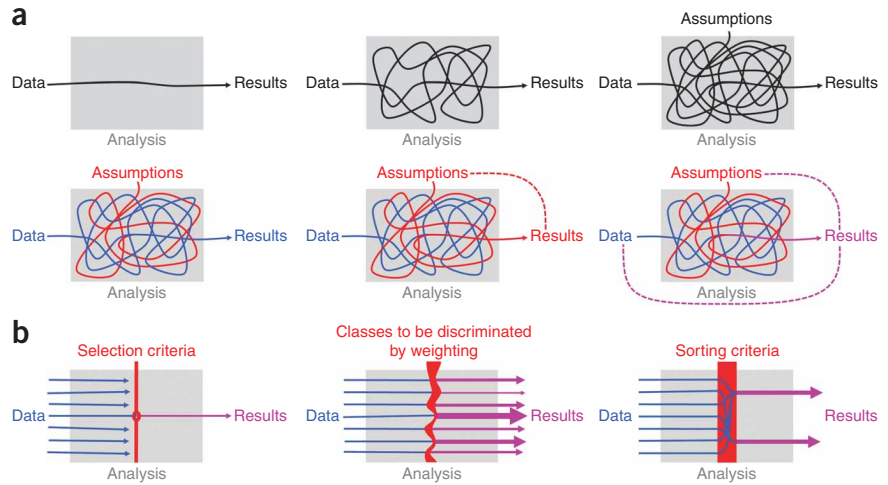
Selective analysis is a powerful tool and is perfectly justified whenever the results are statistically independent of the selection criterion under the null hypothesis. However, double dipping (the use of the same data for selection and selective analysis) will result in distorted descriptive statistics and invalid statistical inference whenever the test statistics are not inherently independent of the selection criteria under the null hypothesis. Non-independent selective analysis is incorrect and should not be acceptable in neuroscientific publications.

Although the dangers of double dipping in the pool of data are well understood in statistics and computer science, the practice is common in systems neuroscience and, in particular, in neuroimaging and electrophysiology. To assess how widespread non-independent selective analyses are in the literature, we examined all of the functional

Laboratory of Brain and Cognition, US National Institute of Mental Health, Bethesda, Maryland, USA. Correspondence should be addressed to N.K. (nikokriegeskorte@gmail.com) or C.I.B. (bakerchris@mail.nih.gov).

Published online 26 April 2009; doi:10.1038/nn.2303

**Figure 1** Intuitive diagrams for understanding circular analysis. **(a)** The top row serves to remind us that our results reflect our data indirectly, through the lens of an often complicated analysis, whose assumptions are not always fully explicit. The bottom row illustrates how the assumptions (and hypotheses) can interact with the data to shape the results. Ideally (bottom left), the results reflect some aspect of the data (blue) without distortion (although the assumptions will determine what aspect of the data is reflected in the results). Sometimes (bottom center), however, a close inspection of the analysis reveals that the data get lost in the process and the assumptions (red) predetermine the results. In this case, the analysis is completely circular (red dotted line). More frequently in practice (bottom right), the assumptions tinge the results (magenta). The results are then distorted by circularity but still reflect the data to some degree (magenta dotted lines). **(b)** Three diagrams that illustrate the three most common causes of circularity: selection (left), weighting (center) and sorting (right). Selection, weighting and sorting criteria reflect assumptions and hypotheses (red). Each of the three can tinge the results, distorting the estimates presented and invalidating statistical tests, if the results statistics are not independent of the criteria for selection, weighting or sorting.



© 2009 Nature America, Inc. All rights reserved.

npg

magnetic resonance imaging (fMRI) studies published in five prestigious journals (*Nature*, *Science*, *Nature Neuroscience*, *Neuron* and *Journal of Neuroscience*) in 2008. Of these 134 fMRI papers, 42% (57 papers) contained at least one non-independent selective analysis (not considering supplementary materials). Another 14% (20 papers) may contain non-independent selective analyses, but the methodological information that was given was insufficient for us to reach a conclusion.

Are all these studies incorrect in their main claims? We do not think so. First, we counted any study containing at least one non-independent selective analysis. For a given paper, the overall claim may not depend on the distorted result. Second, we have no way of assessing the severity of the distortions. They might be small in many cases. If circularity consistently caused only slight distortions, one could argue that it is a statistical quibble. However, the distortions can be very large (Example 1, below) or smaller, but significant (Example 2), and they can affect the qualitative results of significance tests. To decide which neuroscientific claims hold, the community needs to carefully consider each particular case, guided by both neuroscientific and statistical expertise. Reanalyses and replications may also be required.

The problem arises so frequently because the desired selection criterion is often identical with or related to the desired results statistics for the selective analysis. In neuroimaging, for example, we may hypothesize that there is a region responding more strongly to stimulus A than to B, select voxels showing this effect to define an ROI and then selectively analyze that ROI to test our hypothesis. One way to ensure statistical independence of the results under the null hypothesis is to use an independent dataset for the final analysis of the selected channels (for example, neurons or voxels).

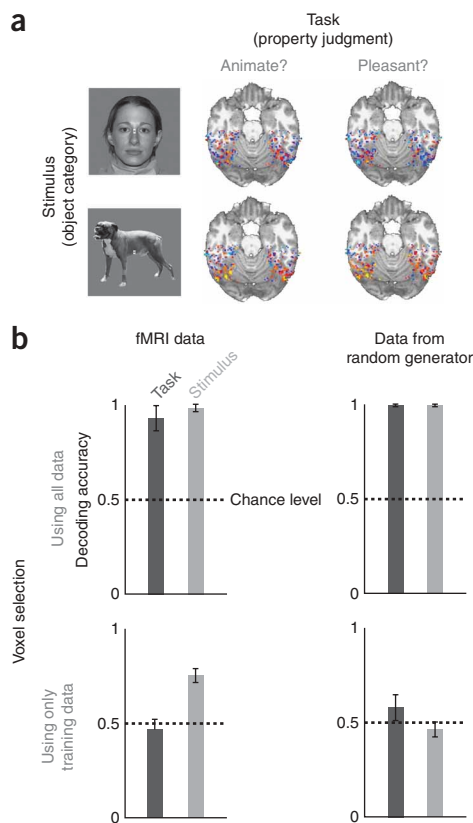
Another way to ensure independence is to use inherently independent statistics for selection and selective analysis. For example, we may select channels with a large average response to stimuli A and B (contrast  $A + B$ ) and test for a difference between the conditions (contrast  $A - B$ ). The contrast vectors ( $[1, 1]^T$  and  $[1, -1]^T$ ) are orthogonal. Unfortunately, contrast-vector orthogonality, by itself, is not sufficient to ensure independence (see **Supplementary Discussion** online). In practice, the same data are frequently used for selection and selective analysis, even when the selection

criteria are not inherently independent of the results statistics. In this case, the results are questionable.

Distortions arising from selection tend to make results look more consistent with the selection criteria, which often reflect the hypothesis being tested. Circularity is therefore the error that beautifies results, rendering them more attractive to authors, reviewers and editors, and thus more competitive for publication. These implicit incentives may create a preference for circular practices so long as the community condones them.

Analyzing multiple channels and reporting results for a statistically selected subset is essential in electrophysiology and neuroimaging. Neuroimaging is faced with even more parallel sites than electrophysiology, typically on the order of 100,000 voxels in the measured volume. However, selection is also an issue in electrophysiology and will gain importance as multi-electrode arrays become more widely used. To its great credit, neuroimaging has developed rigorous methods for statistical mapping from its beginning<sup>7–11</sup>. Note that mapping the whole measurement volume avoids selection altogether; we can analyze and report results for all locations equally, while accounting for the multiple tests performed across locations<sup>12</sup>. The sense of discovery associated with brain mapping derives from this data-driven approach, which avoids both the bias of selective reporting of accurate results and the circularity that can invalidate non-independent selective analyses. Despite the beauty and completeness of a nonselective mapping analysis, selective in-depth analysis of ROIs can yield additional insights<sup>13</sup>.

Here, we demonstrate the problem using two examples from neuroimaging (**Figs. 2 and 3**). In each example, a widely accepted practice is applied to random data known to not contain the experimental effect in question. This exercise reveals the distortion and spurious significance that can arise in circular analysis. We view the problem from three perspectives: as ‘selection bias’, as ‘exploration and confirmation using the same data’ and as ‘overfitting’ (these perspectives are elaborated on in **Supplementary Figures 1–4** and the **Supplementary Discussion** online, which also contain further analyses and simulations and a comprehensive set of questions and answers about circular analysis). Finally, we suggest a policy for noncircular analysis of brain-activity data (**Fig. 4** and **Supplementary Discussion**).



**Figure 2** Example 1: data selection can bias pattern-information analysis. **(a)** To assess the extent to which human inferior-temporal activity patterns reflect bottom-up sensory signals and top-down task constraints, we measured activity patterns with fMRI while subjects viewed object images of different categories and judged whether the object shown was ‘animate’ (task 1) or whether it was ‘pleasant’ (task 2)<sup>2</sup>. **(b)** We selected all inferior-temporal voxels for which any two-sided *t* test contrasting two conditions was significant at  $P < 0.001$  (uncorrected for multiple tests). We then cleanly divided the data into independent sets, using odd runs for training and even runs for testing. We used a linear classifier to determine whether the activity pattern would allow us to decode the stimulus category (light gray bars) and the judgment task (dark gray bars). Results (top left) suggested that both stimulus and task can be decoded with high accuracy, significantly above chance. However, application of the same analysis to Gaussian random data (top right) also suggested high decoding accuracies significantly above chance. This shows that spurious effects can appear when data from the test set are used in the initial data-selection process. Such spurious effects can be avoided by performing selection using data that are independent of the test data (bottom row). Error bars indicate  $\pm 1$  across-subject s.e.m.

are dimensions. Counter to the intuitive simplicity and rigidity of a planar decision boundary, fitting a hyperplane in a 100-dimensional space to separate 100 data points is like separating two points on a plane by a line; separation is always perfect, even if the points are drawn from identical distributions (**Supplementary Discussion**). Separability thus provides no evidence for separate distributions.

Using the same data to train and test a linear classifier can lead us to believe that there is information about the stimulus in regions where actually there is none. In this context, double dipping entails extreme distortions and is widely understood to be unacceptable. We are not aware of examples of this error in the systems neuroscience literature. However, the error here is fundamentally the same as that of non-independent selective analysis. Linear classification is based on a weighted sum of the responses. Weighting can be construed as a continuous variant of selection. Conversely, we can think of selection as binary weighting, a special case.

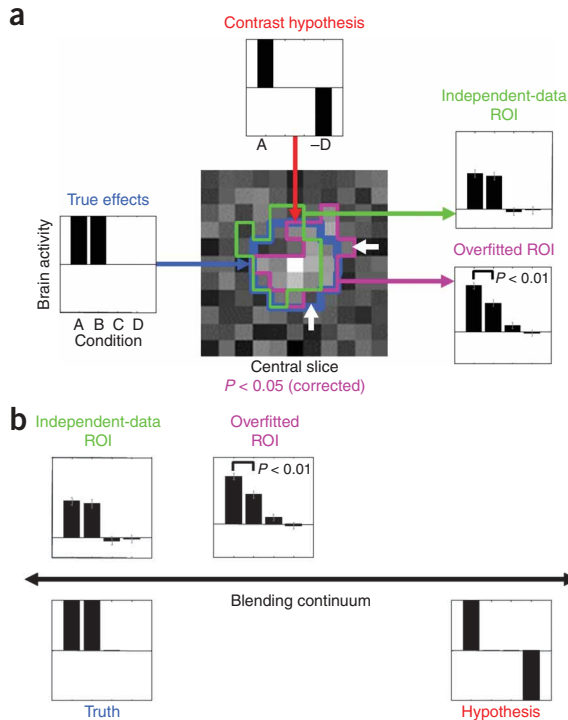
Can selection produce similar distortions as continuous weighting in the context of pattern-information analysis? To test this possibility, we performed a classifier analysis on human inferior-temporal response patterns measured with fMRI while subjects viewed object images<sup>2</sup>. The experiment had two independent variables: object category and task (**Fig. 2a**). In task 1, subjects judged whether the object presented was animate or inanimate. In task 2, they judged whether the object was pleasant or unpleasant. The experiment can reveal the extent to which inferior-temporal activity patterns reflect stimulus category and task.

We first analyzed all experimental runs together to define an ROI. We included all inferior-temporal voxels for which any two-sided *t* test for a pair-wise condition contrast was significant at  $P < 0.001$  (uncorrected). We then cleanly divided the data into independent training and test sets by designating all odd runs as being training data and all even runs as being test data. For the training and the test set separately, we computed the average activity pattern for each condition (combination of task and stimulus category). For each pair of conditions, we decoded a given test pattern by assigning the condition label of the training pattern that was more similar to the test pattern<sup>14</sup>. This nearest-neighbor method is a linear classifier because the condition-average patterns are used. Pattern similarity was measured by the Pearson correlation across voxels. For each subject, decoding accuracy was computed for each pair-wise task comparison in each stimulus category and for each pair-wise stimulus-category comparison in each task (chance level was 50%). Task decoding accuracies were averaged, first within subjects and then across subjects.

### Example 1: pattern-information analysis

In pattern-information analysis<sup>14–18</sup>, the objective is to determine whether the pattern of response in a brain region contains stimulus information. Considering pattern-information analysis is relevant not only because this approach is gaining importance in systems neuroscience, but also because it provides a powerful general perspective on circular analysis<sup>19,20</sup>.

One popular approach to pattern-information analysis is to attempt to decode the stimulus from the response pattern with a pattern classifier<sup>21–23</sup>. If we can predict the stimuli from the response patterns significantly above chance level, then the patterns must contain information about the stimuli. The most common method is linear classification, where a linear decision boundary (that is, a hyperplane) is placed in response-pattern space to discriminate the stimuli. After training the classifier to discriminate example patterns, we can determine its accuracy (percentage of correct classifications). However, if we used the training data to assess the accuracy, we would overestimate the accuracy and conclude that there was stimulus information even if there was none. The reason for this is a phenomenon known as overfitting: a model will capture the noise to some extent as its parameters are fitted to the data. A more flexible model (that is, one with many parameters) will tend to be more susceptible to overfitting. However, even the fitting of a one-parameter model (for example, a mean) is affected by noise to some extent. When thinking about fitting a linear decision boundary, we tend to imagine a line separating two clouds of points in a plane. When there are many points (much data) and few dimensions (for example, two dimensions: a plane), overfitting may be negligible. However, response-pattern space has as many dimensions as there are response channels (for example, neurons or voxels), and a linear decision boundary has as many parameters as there



**Figure 3** Example 2: ROI definition can bias activation analysis. A simulated fMRI block-design experiment demonstrates that non-independent ROI definition can distort effects and produce spuriously significant results, even when the ROI is defined by rigorous mapping procedures (accounting for multiple tests) and highlights a truly activated region. Error bars indicate  $\pm 1$  s.e.m. (a) The layout of this panel matches the intuitive diagrams of **Figure 1a**; the data in **Figure 1a** correspond to the true effects (left), the assumptions to the contrast hypothesis (top) and the results to ROI-average activation analyses (right). A 100-voxel region (blue contour in central slice map) was simulated to be active during conditions A and B but not during conditions C and D (left). The  $t$  map for contrast A – D is shown for the central slice through the region (center). When thresholded at  $P < 0.05$  (corrected for multiple tests by a cluster threshold criterion), a cluster appears (magenta contour), which highlights the true activated region (blue contour). The ROI is somewhat affected by the noise in the data (difference between blue and magenta contours). The noise pushes some truly activated voxels below the threshold and lifts some nonactivated voxels above the threshold (white arrows). This can be interpreted as overfitting. The bar graph for the overfitted ROI (bottom right, same data as used for mapping) reflects the activation of the region during conditions A and B, as well as the absence of activation during conditions C and D. However, in comparison to the true effects (left), it is substantially distorted by the selection contrast A – D (top). In particular, the contrast A – B (simulated to be zero) shows spurious significance ( $P < 0.01$ ). When we use independent data to define the ROI (green contour), no such distortion is observed (top right). (b) The simulation illustrates how data selection blends truth (left) and hypothesis (right) by distorting results (top) so as to better conform to the selection criterion.

**Example 2: regional activation analysis**

A widespread approach to neuroimaging analysis is to perform a statistical mapping, followed by a selective activation analysis of one or more ROIs. The ROIs are typically defined by the mapping; and their analysis is often based on the same data. In many cases, the conclusion that the ROI analysis serves to support is directly or indirectly related to the mapping contrast. Is this a valid approach?

Let us assume that the ROI is defined by a valid statistical mapping analysis with adequate correction for multiple tests. If the statistical mapping were not performed correctly, one could argue that whatever problem arises thereafter is not caused by non-independent selection, but instead by the inadequate statistical mapping. We further assume that the mapping analysis successfully localizes a truly active region. The alternative case that the mapping falsely highlights a region will be rare; it will have a probability of 0.05 or less under the null hypothesis, as the mapping is assumed to be correct. If the mapping did not highlight any region, then there would be no ROI to selectively analyze.

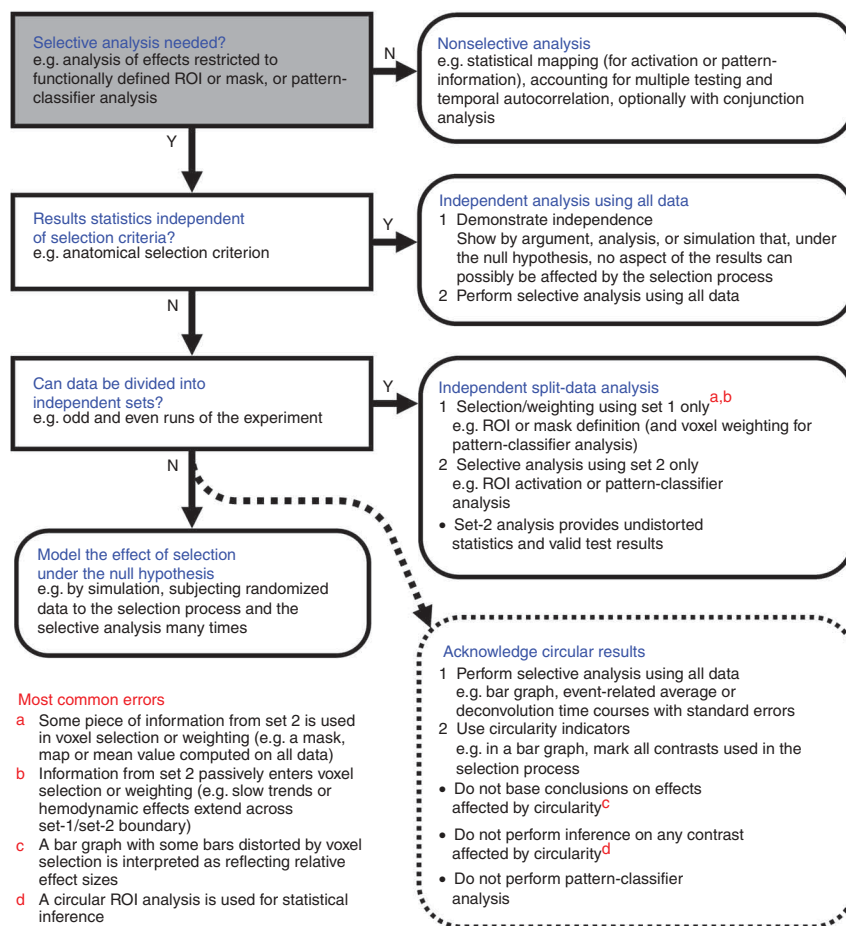
To assess whether an ROI analysis can be distorted by selection under these assumptions, we simulated a neuroimaging dataset of  $30 \times 30 \times 20$  voxels and 200 time points. The simulated experiment was a block design with four conditions (A, B, C and D). We placed a 100-voxel activation ( $5 \times 5 \times 4$  voxels) at the center of the volume. The region was simulated to be active during conditions A and B, but not during C and D (**Fig. 3a**). The resulting spatiotemporal dataset was added to independent spatiotemporal Gaussian noise and spatially smoothed by convolution with a 3-voxel-wide cubic kernel. The data were analyzed by means of a general linear model using the same design matrix as was used to simulate the effects, with one predictor per condition. We mapped the dataset by voxel-wise univariate linear modeling using the contrast A – D (**Fig. 3a**). We thresholded the resulting  $t$  map using a primary threshold corresponding to  $P < 0.0001$  (uncorrected). We then assessed the size of each contiguous cluster exceeding this primary threshold and highlighted all clusters whose size exceeded a cluster-size threshold that controlled the family-wise error rate at  $P < 0.05$ , thus correcting for multiple tests. The cluster-size threshold was determined by simulating the map-maximum cluster-size distribution under the

Stimulus-category decoding accuracies were averaged in the same way. Similar methods are widespread in the literature.

This analysis suggested that both stimulus category and judgment task can be decoded with accuracies above 90% and significantly better than chance (**Fig. 2b**). Therefore, we would conclude that both the task and the stimulus category are strongly reflected in inferior-temporal response patterns. However, when we applied the same analysis to data generated with a Gaussian random generator, we obtained equivalent results (**Fig. 2b**). The random data are known to not contain any information about either task or stimulus category, so any correct analysis should indicate decoding accuracies whose deviations from 50% are in the margin of error and are significant in only 5% of the cases. This demonstrates that selection of ROI voxels using all data can strongly bias estimates of decoding accuracy and yield spuriously significant test results.

The cause of the distortion is the selection of voxels whose time series, by chance, show some consistency between training and test set in the way that they are related to the experimental conditions. For the selected voxel set, training and test datasets are therefore no longer independent. When we corrected the error of non-independent voxel selection, decoding accuracies dropped to chance level for the Gaussian random data (**Fig. 2b**). For the actual experimental data, task decoding accuracy dropped to chance level, whereas stimulus-category decoding accuracy dropped to about 75%, but remained significant (**Fig. 2b**). The latter result replicates a previous study<sup>14</sup>.

Beyond neuroimaging, pattern-information analyses are increasingly used in invasive and scalp electrophysiology. Circularity will cause similar distortions when cells or sensors are preselected by non-independent criteria. We conclude that selection of response channels can strongly inflate estimates of decoding accuracy and misleadingly suggest substantial amounts of information in a brain region, where there is actually none. We can avoid such spurious results by performing selection using data that is independent of the test data.



**Figure 4** A policy for noncircular analysis. This flow diagram suggests a procedure for choosing an appropriate analysis that avoids the pitfalls of circularity. Considering the most common errors (bottom left, red-letter references) can help to recognize circularity in assessing a given analysis. We first consider performing a nonselective analysis only. If selective analysis is needed and we can demonstrate that the results are independent of the selection criterion under the null hypothesis, then all data are used for selective analysis. If we cannot demonstrate this, then a split-data analysis can serve to ensure independence (for details, see **Supplementary Discussion**).

for the null hypothesis that the proportion of results that were significant at  $P < 0.05$  is 0.05). We conclude that non-independent selection can distort the results of selective analyses, even when rigorous statistical tests are used during selection.

Independence of the selective analysis could have been ensured either by using independent test data (**Fig. 3a**) or by using selection and test statistics that are inherently independent. For the contrasts that were used (selection contrast,  $A - D$ ; test contrast,  $A - B$ ), the inherent dependence is obvious; voxels with higher signals during condition A are more likely to be selected by chance using contrast  $A - D$ . Thus, test contrast  $A - B$  will be biased. However, selection bias can arise even for orthogonal contrast vectors (**Supplementary Discussion** and **Supplementary Fig. 3**).

Non-independent selection causes bias because the selection is somewhat affected by the noise (**Fig. 3a**), even when the statistical criterion is stringent and the ROI highlights a truly activated region. Our statistical selection method controls the family-wise error rate; it does not ensure that the ROI perfectly captures the shape of the region. The ROI will be overfitted to the data to some extent, just like the weights of a linear classifier.

To temper this conclusion, we note that overfitting will typically be less severe in fitting an ROI than in fitting a linear classifier with continuous weights. The restriction to binary weights and the constraint of selecting a contiguous set of voxels effectively regularize an ROI fit. In contrast, discontinuous selection (as in Example 1) and data sorting can be extremely susceptible to overfitting (for two simple simulations on sorting effects, see **Supplementary Fig. 2**).

In practice, statistical mapping for ROI definition is not always performed with rigorous correction for multiple tests as assumed here. Many studies rely on a threshold of  $P < 0.001$  (uncorrected). The selective analysis of the same data is then sometimes interpreted as though it confirmed the effect selected for. Although it does not confirm the effect, the selective analysis effectively serves to help us forget about the multiple-testing problem during selection. The inadequacy of the inference during selection will compound the circularity of the selective analysis, and strong biases and large false-positives rates are to be expected.

Although the example here concerns the selection of voxels in a neuroimaging experiment, the same caution should be applied in analyzing other types of data. In single-cell recording, for example, it is common to select neurons according to some criterion (for

null hypothesis by running the above simulation 1,000 times for the same contrast without any effect placed in the data.

The ROI defined by the mapping analysis (**Fig. 3a**) correctly highlighted the activated region. However, the ROI was somewhat affected by noise in the data. Some voxels at the fringe of the ROI will be included because their noise component makes them look as though they conformed slightly better to the selection criterion; others will be excluded because their noise makes them look as though they did not conform as well to the selection criterion. This can be interpreted as overfitting of the ROI.

We averaged all time courses in the ROI (same data as used for mapping) and fitted the linear model. The resulting bar graph (**Fig. 3a**) reflects the activation of the region during conditions A and B as well as the absence of activation during conditions C and D. However, it is substantially distorted by the non-independent selection; recall that the mapping was based on the contrast  $A - D$  (**Fig. 3a**). Although the region was equally activated during conditions A and B, it appears to be more activated during condition A than during B, and this effect is significant ( $P < 0.01$  in the particular example run shown). When we used independent data to define the ROI, no such distortion was observed (**Fig. 3a**).

To assess the proportion of cases in which the contrast  $A - B$  would yield a spuriously significant result caused by non-independent voxel selection, we repeated the simulation 100 times. The one-sided  $t$  test for the ROI contrast  $A - B$  (whose ground-truth value is zero in the simulation) was significant in 20 of the 100 simulations for  $P < 0.05$  and in 9 of the 100 simulations for  $P < 0.01$ . These false-positives rates are significantly larger than for a correct test ( $P = 0.00005$ ,  $\chi^2$  test

example, visual responsiveness or selectivity) before applying further analyses to the selected subset. If the selection is based on the same dataset as is used for selective analysis, biases will arise for any statistic not inherently independent of the selection criterion. For neurons and voxels, selection should be based on criteria that are independent of any selective analysis. In sum, Example 2 shows that non-independent selective analysis can cause significant biases, even when selection is performed with rigorous statistical inference correcting for multiple tests.

### A policy for noncircular analysis

One possible policy that ensures correct inference and undistorted descriptive statistics is summarized by the flow diagram of **Figure 4**. The core of our policy is as follows: we first consider a nonselective analysis (for example, brain mapping with correction for multiple comparisons). If selective analysis is needed, we then assess whether the results statistics are independent of the selection criterion under the null hypothesis. If this has been explicitly demonstrated, then all data are used for selective analysis. Otherwise, an independent dataset is used for the selective analysis to ensure independence of the results under the null hypothesis and prevent circularity. Each of these steps is explained in detail in the **Supplementary Discussion**.

### Conclusion

To learn about brain function, systems neuroscience needs to apply complex selective and recurrent analyses to high-dimensional brain-activity data. One challenge that this poses is to avoid circularity. A circular analysis is one whose assumptions distort its results. We have demonstrated that practices that are widespread in neuroimaging are affected by circularity. In particular, data weighting, sorting and selection can distort results and invalidate tests when preceding non-independent further analyses. Similar practices are common in other fields of systems neuroscience including electrophysiology. The distortions may be small in many cases. However, they can be large and can qualitatively affect results. We conclude that some common practices need to be adjusted. In particular, selection criteria should be demonstrated to be independent of further analyses. A simple way to ensure independence is to use independent data for selection and selective analyses. Immanuel Kant<sup>24</sup> observed that Reason, in science, will not be led on by Nature, but rather forces her to answer specific questions. Circular analysis goes one step further, enforcing specific answers as well (or biasing results in their favor), which is one step too far in our opinion.

*Note: Supplementary information is available on the Nature Neuroscience website.*

### ACKNOWLEDGMENTS

We would like to thank P.A. Bandettini, R.W. Cox, J.V. Haxby, D.J. Kravitz, A. Martin, R.A. Poldrack, R.D. Raizada, Z.S. Saad, J.T. Serences and E. Vul for helpful discussions. This work was supported by the Intramural Research Program of the US National Institute of Mental Health.

### AUTHOR CONTRIBUTIONS

N.K., W.K.S., P.S.F.B. and C.I.B. conceived the project and discussed all of the issues. W.K.S. contributed the data, simulation and analysis for Example 1. N.K. designed and performed all other simulations and analyses. P.S.F.B. contributed to data analysis. N.K. wrote the paper and the **Supplementary Discussion** and made all figures. W.K.S. and P.S.F.B. commented on drafts. C.I.B. edited the paper and guided the project.

Published online at <http://www.nature.com/natureneuroscience/>

Reprints and permissions information is available online at <http://www.nature.com/reprintsandpermissions/>

- Baker, C.I., Hutchison, T.L. & Kanwisher, N. Does the fusiform face area contain subregions highly selective for nonfaces? *Nat. Neurosci.* **10**, 3–4 (2007).
- Simmons, W.K. *et al.* Imaging the context-sensitivity of ventral temporal category representations using high-resolution fMRI. *Soc. Neurosci. Abstr.* **262.12** (2006).
- Baker, C.I., Simmons, W.K., Bellgowan, P.S. & Kriegeskorte, N. Circular inference in neuroscience: the dangers of double dipping. *Soc. Neurosci. Abstr.* **104.11** (2007).
- Vul, E. & Kanwisher, N. Begging the question: the non-independence error in fMRI data analysis. in *Foundations and Philosophy for Neuroimaging* (eds. Hanson, S. & Bunzl, M.) (in the press).
- Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly high correlations in fMRI studies of emotion, personality and social cognition. *Perspect. Psychol. Sci.* (in the press).
- Benjamini, Y. Comment: microarrays, empirical bayes and the two-groups model. *Statist. Sci.* **23**, 23–28 (2008).
- Worsley, K.J., Evans, A.C., Marrett, S. & Neelin, P. A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12**, 900–918 (1992).
- Friston, K.J., Jezzard, P. & Turner, R. Analysis of functional MRI time-series. *Hum. Brain Mapp.* **1**, 153–171 (1994).
- Worsley, K.J. & Friston, K.J. Analysis of fMRI time-series revisited—again. *Neuroimage* **2**, 173–181 (1995).
- Nichols, T. & Hayasaka, S. Controlling the family-wise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* **12**, 419–446 (2003).
- Genovese, C.R., Lazar, N.A. & Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15**, 870–878 (2002).
- Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P. & Henson, R.N. A critique of functional localisers. *Neuroimage* **30**, 1077–1087 (2006).
- Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localisers. *Neuroimage* **30**, 1088–1096 (2006).
- Haxby, J.V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
- Cox, D.D. & Savoy, R.L. Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* **19**, 261–270 (2003).
- Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F. & Wang, X. Learning to decode cognitive states from brain images. *Mach. Learn.* **57**, 145–175 (2004).
- Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci. USA* **103**, 3863–3868 (2006).
- Norman, K.A., Polyn, S.M., Detre, G.J. & Haxby, J.V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424–430 (2006).
- Strother, S.C. *et al.* The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage* **15**, 747–771 (2002).
- Strother, S. *et al.* Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics. I. A preliminary group analysis. *Neuroimage* **23**, S196–S207 (2004).
- Mitchell, T.M. *Machine Learning* (McGraw Hill, Portland, Oregon, 1997).
- Duda, R.O., Hart, P.E. & Stork, D.G. *Pattern Classification*. 2nd edn. (Wiley, New York, 2001).
- Bishop, C.M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
- Kant, I. *Kritik der reinen Vernunft*. 2nd edn. <http://www.gutenberg.org/dirs/etext04/Bick210.txt> (1781).