Supplementary Discussion

A policy for noncircular analysis

We describe one possible policy that ensures correct inference and undistorted descriptive statistics. The description of this policy is summarized by the flow diagram of Fig. 4. (For additional informal notes that might be helpful to consider for researchers, authors, and reviewers dealing with this issue, see below: *Q&A, Preventing circularity*.)

Is a selective analysis needed? The first question to be decided is whether a selective analysis is needed. We can avoid selection altogether by performing an inferential statistical mapping of the whole measurement volume.[1] This powerful approach allows us to analyze and report results for all locations equally, while accounting for the multiple tests performed across locations.[a] We can, thus, avoid both the bias of selective reporting of accurate results and the inaccuracies that can arise from selection. If we are interested in regions exhibiting multiple effects, mapping can be performed for conjunctions of contrasts[2,3] or other more complex test statistics[4]. This approach is data-driven with respect to the spatial dimension, but hypothesis-driven with respect to the effects to be investigated and has considerable advantages. Despite the beauty and completeness of a nonselective mapping analysis, selective in-depth analysis of regions defined by mapping can yield additional insights.

Are all results statistics independent of the selection criteria? In case a selective analysis is to be performed, the next question is whether the results statistics are independent of the selection criterion under the null hypothesis. For example, if we define an ROI purely on the basis of brain anatomy, say the amygdala, then we can argue that the selection of functional data by this criterion cannot possibly bias the results statistics. The argument that the results statistics are independent is less straightforward if we use the same functional data set for selection and selective analysis. It is sometimes argued that a test contrast vector $\mathbf{c}_{test}$ orthogonal to the selection contrast vector $\mathbf{c}_{selection}$ will not yield biased results. Unfortunately this is not true in general, as can easily be shown analytically or by simulation (Fig. S3).

Contrast-vector orthogonality does not ensure independent statistics. For example, consider the selection contrast A+B ($\mathbf{c}_{selection} = [1\ 1]^{\mathbf{T}}$) and the test contrast A-B ($\mathbf{c}_{test} = [1\ {-1}]^{\mathbf{T}}$). These are orthogonal contrast vectors in that their inner product is zero: $\mathbf{c}_{selection}{}^{\mathbf{T}} \cdot \mathbf{c}_{test} = 0$. However, whether selecting with $\mathbf{c}_{selection}$ biases testing with $\mathbf{c}_{test}$ also depends on the design matrix $\mathbf{X}$.

[a] Statistical mapping can also be performed for restricted search volumes. This increases sensitivity, because there are fewer tests whose familywise-error or false-discovery rate needs to be controlled. The search volume for such restricted mapping analyses needs to be defined by criteria independent of the test statistic – just like any ROI to be selectively analyzed.

Nature Neuroscience: doi:10.1038/nn.2303

For the two condition example, selecting by the contrast A+B can bias testing the contrast A-B if the design is not balanced with respect to A and B. For an intuitive understanding of this, consider the case when there is more data for condition A than for condition B. Condition B will be less stably estimated. As a result selecting by contrast A+B will favor voxels having high positive noise in condition B. High positive noise in condition A will not be favored equally since the noise is more strongly reduced by averaging for condition A. As a result, the contrast A-B will be negatively biased.

For an ordinary-least-squares analysis, the effect of the design matrix can be taken into account by using the criterion $\mathbf{c}_{selection}^{\mathbf{T}} \cdot (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \cdot \mathbf{c}_{test} = 0$. However, if the errors are temporally dependent (as is the case for fMRI data), small biases can still arise. The temporal dependence can be characterized by a time-point mixing matrix $\mathbf{S}$. Taking temporal dependence into account yields the criterion $\mathbf{c}_{selection}^{\mathbf{T}} \cdot (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \cdot \mathbf{X}^{\mathbf{T}} \cdot \mathbf{S} \cdot \mathbf{S}^{\mathbf{T}} \cdot \mathbf{X} \cdot (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \cdot \mathbf{c}_{test} = 0$. Temporal dependence is not a concern in second-level between-subject analyses.

If it can be demonstrated that all results statistics are independent of the selection process under the null hypothesis, then all data should be used for selection and selective analysis. This maximizes the power for the selective analysis and obviates the complication of dividing the data.

Can the data be divided into independent sets? If any of the results statistics are not inherently independent of the selection process under the null hypothesis (a frequent case in practice), then independence of the results can be achieved by dividing the data into independent subsets. Data set 1 is then to be used for selection and data set 2 for the selective analysis. In the context of fMRI, a good way to divide the data is to number the experimental runs chronologically as measured and combine all odd runs to form data set 1 and all even runs to form data set 2. This approach prevents the temporal dependencies within runs from translating into dependencies between the two data sets.

In the case of a pattern-classifier analysis, set 1 can be used for both selection and classifier training (e.g. the determination of linear weights). Set-2 statistics then provide undistorted statistics and valid results.

Although the amount of data available for testing will be reduced, this approach allows us to use arbitrary selection criteria related to (or even identical with) the test statistics. Note also, that the selection process need not involve statistical inference. In pattern-classification analysis, for example, classifier training typically does not involve statistical inference. Inference can be performed on data set 2. Dividing the data into independent sets allows us to explore data set 1 and fit arbitrarily complex models, as long as we do not repeat the cycle of exploration and confirmation using any piece of data set 2. This flexibility, granted by splitting the data, is very powerful.

Crossvalidation allows us, in a sense, to have our cake and eat it too: by splitting off small independent subsets of the data repeatedly, we can take advantage of the benefits of a split-data analysis, and still use almost all data for fitting and all data testing. The price of this is added computational demands. The implementation of a correct crossvalidation scheme requires some care as each split-off subset needs to be independent of the remainder of the data. Importantly, the selection process must be performed independently for each crossvalidation fold. In neuroimaging, this means that the statistical mapping is repeated and the ROI redefined on each fold.

Nature Neuroscience: doi:10.1038/nn.2303

What if all this fails? What if there are nonindependent results statistics, but the data is not to be divided? Is there a correct way to use a single data set? One method would be to model the effect of selection on the results statistics (Fig. 4, bottom left box). Another approach would be to acknowledge the circular aspects of the analysis (Fig. 4, bottom right box). The methods for the former have yet to be developed and the latter should be viewed as a last resort. The study will be questionable with respect to all results statistics not demonstrated to be independent of the selection process.

Modeling the effect of selection on the results statistics under the null hypothesis (for hypothesis testing) or under an alternative hypothesis (for estimating the degree of distortion) may not be tractable analytically, but could be achieved by simulation. For fMRI data, for example, we would need to simulate the variability of the ROI definition given the noise (leaving the effect that defines the ROI intact) and estimate the distribution of the results statistics on this basis, thus taking the selection bias into account. We are not aware of an example of this approach in the literature. Appropriate methods would have to be developed.

Tolerating selection biases and acknowledging them both visually in the figure and verbally in the text of the paper should be viewed as a last resort. This shortcut might be chosen if the results in question are not central to the conclusions of the study and the cost of a proper analysis (which might require additional experiments) outweighs its benefit. We suggest using visual “circularity indicators” to mark all biased effects (Fig. S4 demonstrates this for a set of bar graphs). The purpose of circularity indicators is to prevent readers from drawing conclusions on the basis of the biased aspects of the results. Again, the caveat that even contrasts with orthogonal weight vectors can be biased needs to be considered. Therefore explicit demonstrations of independence as discussed above (under *Are all results statistics independent of the selection criteria?*) will be required in this context as well: for all effects unmarked by circularity indicators.

Policy summary. To summarize the core of our policy, we first consider a nonselective analysis (e.g. brain mapping with correction for multiple comparisons). If selective analysis is needed, we next assess whether the results statistics are independent of the selection criterion under the null hypothesis. If this has been demonstrated, then all data should be used for selective analysis. Otherwise, an independent data set for the selective analysis can serve to ensure independence of the results under the null hypothesis and prevent circularity.

Alternative perspectives for understanding circularity

(1) The cycle of exploration and confirmation

The cycle of exploration and confirmation in science provides a useful perspective on the problem of circular analysis. Hypotheses generated by exploring the data require confirmation by means of independent data, because a relationship observed in a data set will be consistent with that data set, whether or not it reflects a true relationship or just the noise.

A prespecified hypothesis will be related to the noise in the data only by chance. This renders statistical testing straightforward. Hypotheses generated by exploring the data are therefore generally thought to

Nature Neuroscience: doi:10.1038/nn.2303

require confirmation by means of independent data. Selection (e.g. of an ROI), weighting (e.g. in linear classification), and sorting (e.g. of neurons according to their tuning) can all be viewed as exploratory analyses that fit a model to the data so as to generate a specific hypothesis (e.g. “This particular region will respond more strongly to stimulus A than B.”), which is to be confirmed by a subsequent test.

A statistical significance test assesses the probability of observing the hypothetical relationship (as strongly or more strongly than it has been observed) under the null hypothesis that it does not truly exist. Using the same data set to generate and test a hypothesis is circular unless the multiplicity of hypotheses considered in the exploration process is taken into account in modeling the null-hypothesis scenario. In other words, a test using the same data would need to address the question: If the data contained only noise and we searched for an effect the way we did, with what probability would we find an effect as strong as (or stronger than) the one we observed?

Statistical brain mapping is a case in point: Hypotheses are tested in many brain locations, but this exploratory process (i.e. the multiple testing) is accounted for in statistical inference. This allows us to perform exploration and confirmation in one go using a single data set. For complex exploratory analyses, including classifier training and ROI definition, this can be a difficult feat and using independent data for confirmation is often preferred. We will first consider the case of a classifier training, then the case of ROI definition (or, more generally, channel selection).

Classifier analysis as exploration and confirmation

We can think of fitting a linear classifier in order to discriminate two experimental conditions as generating a hypothesis by exploring the data (Fig. S1, left). The hypothesis is typically subject-specific. For example, it could amount to a statement such as: “If this subject’s neuronal responses in these locations are weighted with these particular weights and summed, the resulting number will reflect the perceived stimulus.” The weights will be overfitted to some degree. The training-data classification accuracy is therefore a positively biased estimate of the actual classification accuracy. In other words, we do not know to what extent the hypothesis we generated by exploring the data reflects the noise. We therefore need independent data to confirm the hypothesis (Fig. S1, right).[b] Note that the same-data bias can be extreme: We could have perfect classification on the training set even under the null hypothesis of identical response patterns to the stimuli. (When and why this occurs is explained in detail in the *Q&A* below, question *What is overfitting?*) The classifier will perform at chance level on the independent test data if the null hypothesis is true.

ROI-average activation analysis as exploration and confirmation

An ROI can similarly be viewed as part of a specific hypothesis defined by exploration and requiring independent confirmation (Fig. S1, left). Again, the hypothesis is typically subject-specific. For example, it could amount to the statement: “This particular set of voxels in this subject’s brain responds more

[b] One might ask if we cannot test for a multivariate effect using a single data set in this context. The answer is: yes. For example, a multivariate analysis of covariance would take the many different possible dimensions of the effect into account. However, the test would only be valid if the noise were multivariate normal, which might not be the case. Using a Fisher linear discriminant, we would rely on the multivariate normal assumption for sensitivity (the classifier would work best for normal noise), but a test on independent data would still be valid if the multivariate normal assumption were violated.

Nature Neuroscience: doi:10.1038/nn.2303

strongly to stimulus A than B." Note that this is distinct from the hypothesis tested by the statistical mapping. The hypothesis tested by the mapping is: "There is a blob of activation somewhere." The mapping, ideally, confirms this hypothesis and generates a more specific one: the ROI hypothesis, which – unlike the mapping hypothesis – specifies a particular set of voxels.

Defining an ROI is equivalent to assigning a weight of either 0 (for outside the ROI) or 1 (for inside) to each voxel. As we have seen in regional-activation analysis of Example 2 in the paper, these binary weights will be overfitted to some degree. The ROI activation in the data set used for mapping is therefore a positively biased estimate of the actual ROI activation for stimulus A, compared to B. In other words, we do not know exactly to what extent the ROI hypothesis we generated by exploring the data reflects the noise. We therefore need independent data to confirm the hypothesis (Fig. S1, right).

The hypothesis of greater ROI activation for stimulus A than B is less interesting to us after the mapping result than that of pattern discriminability by linear classification: The ROI activation appears to be already confirmed by the mapping analysis. However, the mapping is confirmatory only with respect to the hypothesis that there is a blob of activation; it is exploratory with respect to the exact ROI. The ROI definition therefore is a new hypothesis generated by the mapping. Although we expect the ROI activation to be significantly positive, the data used for mapping do not give us an accurate estimate of its magnitude.

Moreover, if we are to test a distinct hypothesis on the basis of the ROI, then the ROI definition becomes a component of the tested hypothesis. The test statistic then must not be biased by the ROI definition. In Example 2 (Fig. 3), the mapping contrast was A-D, so any contrast involving either condition A or condition D would be biased. Such biased contrasts include A, A-B, A-C, and A+B. Biased results can occur even for orthogonal test contrast vectors (e.g. A+D in this case). This is explained above in the section *A policy for noncircular analysis* (for more details, see *Q&A*, below). One safe way to ensure independence of the test statistic under the null hypothesis is to use independent data.

(2) Overfitting of model parameters

We have seen that independent training and test data are required in pattern-classifier analysis. In the previous section we viewed the fitting of a model to the data (e.g. classifier training) as an exploratory analysis requiring independent confirmation. In this section, we explore the concept of "overfitting" of model parameters, which provides another perspective on the problem of circular analysis.

The accuracy of a classifier (i.e. its percentage of correct classifications) on the training data is an inflated estimate of its accuracy on independent test data (i.e. its generalization accuracy). The cause of this phenomenon is "overfitting": The fitted model will reflect not only the true effects in the data, but also the noise to some degree. To make this point and its range of consequences intuitive, we will consider training a nearest-neighbor classifier, training a linear classifier, computing a simple mean, and defining an ROI or selection mask. All of these suffer from different degrees of overfitting.

Nature Neuroscience: doi:10.1038/nn.2303

Training a nearest-neighbor classifier – so as to test its performance

A nearest-neighbor classifier is perhaps the simplest possible classifier: Each new response pattern is classified as the stimulus associated with the most similar response pattern in the data. It is easy to see that such a classifier will perfectly decode the data it is defined by: Each response pattern in the data is unique and, trivially, will be most similar to itself among the set of patterns. It will, thus, receive the correct stimulus label, even if the brain region the patterns are taken from contains no information about the stimulus at all or the data are produced by a random generator. We might erroneously conclude that the region distinguishes the stimuli.

This is an extreme case of overfitting. A model is said to be overfitted to the data when its parameters reflect the noise. A more complex model (i.e. one with more parameters) will be more susceptible to overfitting. Because of overfitting, the quality of a pattern classifier is always assessed by its accuracy on an independent test data set.

Training a linear classifier– so as to test its performance

What if we chose a linear classifier (also called a linear discriminant) to assess whether the stimulus can be decoded from the response pattern? A linear classifier assigns a weight w_i to each response channel i (e.g. each voxel or neuron) and computes a weighted sum across the channels, reducing each response pattern j to a single score $s_j = \Sigma_i w_i \cdot response_{i,j}$. Each pattern is then classified by determining whether its score s_j exceeds a particular threshold. The weights w_i are chosen with the goal to maximize the classification accuracy.

To make this intuitive, imagine computing each voxel's t value for the contrast between stimuli A and B. We could simply use these t values as weights. Voxels with a larger response to A than B will get a positive weight; voxels with a smaller response to A than B will get a negative weight. Voxels with a greater absolute t value will have greater influence on the decision than voxels responding about equally to both stimuli. Commonly used linear classifiers employ more sophisticated methods for optimizing the weights, but this naive method might work quite well for neuroimaging data.

Classification by thresholding of a weighted sum of the inputs is equivalent to placing a linear decision boundary (i.e. a hyperplane) in response-pattern space. Response-pattern space is the space spanned by the single-voxel activity levels. A response pattern corresponds to a point in this space. All patterns on one side of the decision hyperplane are classified as stimulus A and those on the other side as stimulus B.

In neuroimaging, the number of patterns used to "train" the classifier (called the "training data") is not typically greater than the number of voxels in the region. For example, we may have 100 voxels in the region and 100 response-pattern estimates (50 per stimulus condition). In that case, there exists a hyperplane that perfectly separates the 100 response patterns, even if the brain region the patterns are taken from contains no information at all about the stimulus or the data are produced by a random generator. As for the nearest-neighbor classifier, we might erroneously conclude that the region distinguishes the stimuli if we used the training data to assess decoding accuracy.

For intuition, imagine red and blue dots (response patterns) in a space (response space). Consider one red and one blue dot on a single dimension: they can always be separated by a threshold (unless they fall on the same point, a case whose probability is infinitesimally small under noise). Next, imagine one

Nature Neuroscience: doi:10.1038/nn.2303

red and two blue dots on a plane (two dimensions) and consider the fact that they can always be separated by a line so as to divide them according to their color – splitting off the red one. Next, imagine four points in three dimensions: they can always be separated by a plane. This generalizes: up to n points in n-1 dimensions can always be separated by a hyperplane so as to divide them according to their color – no matter which dots are colored red and which blue. This holds under one condition: the points need to be "in general position", which means that no two of them are located on the same point in space, no three on a line, no four on a plane, and so on.[5,6] Linear separability, then, provides no evidence at all, that the red and blue dots correspond to distinct distributions.

The reason why a hyperplane can fit the data so well as to perfectly separate the 100 response patterns according to stimulus condition in the above example is that a hyperplane in 100 dimensions has 101 parameters (the 100 weights define the orientation, the threshold shifts the hyperplane to the optimal position). The large number of parameters gives the model a lot of flexibility: Fitting a plane in 100 dimensions is much like drawing a convoluted line in a plane so as to exactly separate two sets of points.

Like the nearest-neighbor classifier described above, this is an extreme example of overfitting: Even if the response patterns contain no information about the stimulus, classification will be perfect on the training set. The fitted hyperplane reflects only noise in that case, so it will not perform above chance level on independent data. If there were more data points or less response channels, overfitting would be ameliorated. However, there would still be some degree of overfitting. This is why independent test data are needed to estimate classifier accuracy and determine if the region contains information about the stimulus.

Computing a mean – so as to estimate variance

Note that overfitting also affects models with very few parameters, albeit to a much lesser degree. Consider the opposite extreme: In order to estimate some effect from a set of noisy measurements, we might compute the mean of the measured values. Although the mean may be the best possible estimate, there will be some deviation between the true mean (i.e. the population mean) and the mean of the measurements (the sample mean). This can be interpreted as overfitting: The estimate reflects not only the true quantity, but also the noise. If we now estimate the noise variance (i.e. the average squared noise displacement) on the basis of the mean, we will underestimate the noise variance. This is because the mean is the reference value that minimizes the sum of squared deviations of the measurements. Estimating the variance as the average squared deviation from the mean can thus be viewed as circular. One solution would be to split the data and use independent sets to (a) estimate the mean and (b) estimate the deviations from that mean. In this particular case there is a better solution: We can correct for the bias by estimating the noise variance as the sum of squared deviations from the mean divided by n-1 (instead of n), where n is the number of measurements.

Defining an ROI or selection mask – so as to perform a selective analysis

Many analyses in systems neuroscience rely on data selection based on brain-activity analyses. Selecting data can be viewed as fitting a model of binary weights. The model will necessarily be overfitted to some degree. As in the other cases, one way to avoid circularity is to use independent data for further analyses based on the fit.

Nature Neuroscience: doi:10.1038/nn.2303

Questions and answers about circular analysis

Below we have assembled a list of questions and answers about circular analysis. The questions are grouped in six thematic blocks. The answers are written somewhat redundantly for didactical purposes and in order to allow selective reading. Some of the questions go beyond what is explained in the paper, others reiterate points made in the paper or place them in a different context.

Table of contents

Nature Neuroscience: doi:10.1038/nn.2303

Power (i.e. sensitivity) and test validity (i.e. specificity)

1. Will we have enough power when we split the data?

Power considerations are a legitimate criterion for deciding among correct analyses. If the results are demonstrably independent of the selection criteria under the null hypothesis, then the data do not need to be divided and using all data will afford more power. Otherwise the data do need to be divided. Circular analysis is not an option because it is incorrect. If the most powerful correct analysis lacks power, then more data is needed or the null hypothesis should be accepted.

If we use part of the data for selection (or, more generally, for hypothesis generation or training), then this data set should be thought of as "*used up*": it represents the price of generating a specific hypothesis (which requires independent data to be tested). If we had not *used up* part of the data to generate the specific hypothesis to be tested, we could not test this hypothesis at all. We would have to test a more general hypothesis that does not require prior data fitting. Such a more general hypothesis will typically require more data to be tested with the same power. Dividing the data, thus, arguably affords an *increase* in power by allowing us to test a more specific hypothesis to address the same conceptual question.

Instead of dividing the data into two independent sets and using one set only for selection and the other only for testing, we can use crossvalidation. This complicates the analysis, but can afford greater power. Crossvalidation allows us to use most of the data for selection and all of the data for testing, while maintaining independence. For an explanation of crossvalidation, see Questions 21-23.

Nature Neuroscience: doi:10.1038/nn.2303

2. Is it not legitimate to trade off specificity for sensitivity?

Yes, it is legitimate to trade off specificity for sensitivity (i.e. power) – as long as the analysis is not circular. A circular analysis not only sacrifices specificity, but it does so in an uncontrolled manner. A better way to trade off specificity for sensitivity in frequentist hypothesis testing (i.e. testing for significant deviations from a null hypothesis) would be to perform a noncircular analysis and set the p threshold to a higher value than 0.05, such as 0.1 or 0.2. Specificity will suffer (as in circular analysis), but at least we know how much.

Nonindependent selective analysis is never acceptable

3. Is a nonindependent selective analysis acceptable if data were selected by rigorous statistical inference corrected for multiple tests?

No. Example 2 shows that there can still be a bias. Rigorous statistical inference will control the familywise-error rate or the false-discovery rate. It does not prevent overfitting of the selection mask (e.g. the ROI). Statistics related to the selection criterion, therefore cannot be estimated without bias or validly tested without using independent data.

While valid statistical inference for selection does not justify nonindependent selective analysis, using selection criteria without valid statistical inference tends to produce even larger biases in nonindependent selective analysis. For example, if we map an fMRI volume and use an inadequate criterion such as $p<0.001$ (uncorrected), we may well find spurious regions, even in pure noise data. And, even in pure noise data, these regions will exhibit the effect they are selected for. Inadequate statistical inference compounds the problem of circular selective analysis and leads to biases that are typically even greater than those of Example 2, where a valid inferential mapping highlighted a truly activated region.

4. Can nonindependent selective analysis be used for descriptive rather than inferential purposes?

No. Nonindependent selective analysis distorts descriptive statistics. It is because of this distortion that statistical inference is also invalid. The distinction between descriptive and inferential does not help in this context.

Common descriptive analyses include bar graphs of brain activation levels, scatterplots of brain activity versus task- or subject-related covariates, as well as the computation of effect estimates such as the r value. Descriptive analyses distorted by selection are questionable, because they don't accurately "describe" the brain region under study. To make matters worse, nonindependent selective analyses tend to be distorted so as to suggest evidence for the hypothesis when there is none or so as to exaggerate the evidence for the hypothesis. Moreover, distortions can take unexpected forms that could

Nature Neuroscience: doi:10.1038/nn.2303

only be predicted by a detailed study of the selection effects for the scenario at hand. The magnitude of the distortions is also unknown.

Visual inference based on plots of distorted statistics is misleading for the same reason that *statistical inference* based on these statistics is invalid. Both can suggest conclusions that are not supported by the evidence.

5. Isn't a nonindependent analysis of statistically selected data acceptable unless it is interpreted as independent validation?

It certainly will not provide independent validation. Beyond that, however, it will exaggerate the effect the data are selected for, thus also not serving the purpose of accurately characterizing either the size of that effect or the profile of effects across conditions, if some or all effects plotted are affected by the distortion.

6. Can a nonindependent selective analysis not reveal important additional information?

Yes, it can. However, the additional information will be obscured by the inevitable distortions whose form and magnitude is unknown.

It's like taking a photo of a scene through a distortion lens with unknown properties in order to estimate the sizes of different objects. While the picture will contain novel information, its unknown distortion makes it impossible to draw compelling conclusions about either the sizes of the objects (descriptive statistics) or size differences between them (statistical inference).

One can construct cases, where it is revealing – as part of a quick-and-dirty data exploration – to look at distorted analyses. But this requires a keen awareness of the presence, likely form, and inherent unpredictability of the distortions. It should also be viewed as exploratory, not confirmatory. We feel that it is bad scientific style to use such analyses in papers when undistorted analyses could be provided.

(a) In particular, can a scatterplot for nonindependently selected data not reveal outliers? Yes, it might reveal outliers (i.e. data points whose noise component is not well accounted for by the noise model used in the analysis). However, it might alternatively obscure outliers: Nonindependent selection might (1) select data with outliers that happen to conform to the selection criterion and (2) obscure that they are outliers by simultaneously favoring data whose non-outlier noise is more consistent with the selection criterion. In effect, this would move other data points toward the regression line, thus giving the impression that the outliers are just extreme points and that the apparent correlation is not accounted for by outliers alone.

(b) In particular, are time courses for statistically selected voxels not helpful to look at? Yes, they can be helpful to look at. However, they will tend to reflect the effects they were selected for more strongly than they should, reproducing the selection contrast as well as the temporal shape of the

Nature Neuroscience: doi:10.1038/nn.2303

hemodynamic response model used. Other aspects of same-data time courses may also be distorted – in unexpected ways.

Consider the case of inspecting event-related fMRI time courses for conditions A and B after defining the ROI by the contrast A-B. If the design matrix used for selection contains hemodynamic response predictors for A and B, thus assuming a shape of the hemodynamic response, then the ROI voxels will also be selected so as to conform to that shape.

7. Can aspects of the data independent of the selection criterion not be revealed by same-data analysis?

Yes. If it can be demonstrated that all aspects of the results are independent of the selection criterion, then all data should be used for selective analysis (see section *A policy for noncircular analysis* above and Fig. 4).

The problem is that it is not easy to predict exactly how selection will affect different aspects of the analysis. Interpreting results that are distorted in complex ways to unknown degrees is questionable, even if those results do contain information not predetermined by the selection process. The novel information is often rendered useless because it is buried among distorted effects. The burden of proof is on the researcher to demonstrate what aspects of the results are strictly independent of the selection criterion.

8. Aren't descriptive visualizations helpful to illustrate the claims of a paper?

Yes, descriptive visualizations (such as scatterplots) can provide illustrations that are helpful in communicating the claims of a paper, thus "telling the story". This constitutes an important part of scientific communication.

For the purpose of illustrating a hypothesis, it is entirely legitimate to include plots designed by hand. If a plot claims to present empirical evidence, however, the evidence should not be distorted.

The ideal scientific visualization simultaneously provides (1) an undistorted view of the data biased by minimal assumptions and (2) an intuitive illustration of the claim. This is only possible when the data support the claim and are sufficiently clean for the visualization to clearly communicate it.

In systems neuroscience, reality is typically more complex than our hypotheses. As a result, the ideal scientific visualization described above is often out of reach. Something has got to give. So we are faced with a choice. Legitimate options include:

- a visualization that is undistorted and biased by minimal assumptions, but suggests a more complex picture than our hypothesis in its strong form

Nature Neuroscience: doi:10.1038/nn.2303

- a visualization that is undistorted, but utilizes strong assumptions to reduce complexity (perhaps to a single dimension on which a positive value would support our hypothesis)

Nonindependent selection will tend to “clean up”, appearing to give us both a view of the data and a clear illustration of our hypothesis. However, it is not a legitimate option because it misrepresents the data. Including plots that are based on data but distorted in favor of the hypothesis is akin to morphing between a hand-drawn plot illustrating the hypothesis and a data-based plot showing actual results. While each of these two is useful in its pure form, their amalgamation is misleading.

It is important to be aware that there is a tradeoff between the clarity of our story telling and the accuracy of our presentation of the empirical evidence. Present incentives may favor the former at the expense of the latter. We feel that this is unhealthy for our field.

9. Is selective same-data analysis valid if an orthogonal contrast is used for selection?

In multifactor designs, it is common practice to define ROIs using orthogonal contrasts (sometimes referred to as localizing contrasts). This is a valid and useful way to increase the power of statistical inference, which precludes the bias addressed by this paper.

However, it is important to appreciate the precise meanings of the statistical concepts of “contrast” and “contrast orthogonality”. The “contrast” is not the contrast weight vector, but the linear combination of the data, i.e. the effect estimate itself (a single number). “Contrast orthogonality” means that two contrasts are statistically independent under the null hypothesis. (If the null hypothesis were true and one repeated the experiments many times, the two contrast estimates would be uncorrelated across repetitions.) Under this definition, selection with an orthogonal contrast cannot bias the test results.

If it can be demonstrated that all results statistics are independent of the selection process under the null hypothesis, then all data should be used for selection and selective analysis. This maximizes the power for the selective analysis and obviates the complication of dividing the data.

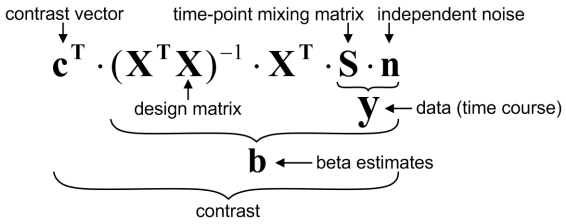
10. Do orthogonal contrast vectors ensure contrast orthogonality?

No. Contrast-vector orthogonality does not imply contrast orthogonality. Even for orthogonal contrast vectors, unbalanced design matrices and dependent errors can lead to non-orthogonal contrasts, which will introduce selection bias. This can easily be shown analytically or by simulation (Fig. S3).

Let us assume the null hypothesis holds. Let us further assume that the data $\mathbf{y}$ (time by 1, a single time course) originate from independent equal-variance Gaussian noise values $\mathbf{n}$ (time by 1), which are mixed in the data according to the time-point mixing matrix $\mathbf{S}$ (time by time, this can account e.g. for noise autocorrelation, as is present in fMRI data): $\mathbf{y} = \mathbf{S}\cdot\mathbf{n}$. We typically perform an ordinary least-squares fit of a design matrix $\mathbf{X}$, obtaining beta estimates $\mathbf{b} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\cdot\mathbf{X}^\mathbf{T}\cdot\mathbf{y}$. We then compute a contrast, i.e. an effect

Nature Neuroscience: doi:10.1038/nn.2303

estimate, for contrast vector $\mathbf{c}$ (predictor by 1) as $\mathbf{c}^{\mathbf{T}}\cdot\mathbf{b}$. To see how this contrast $\mathbf{c}^{\mathbf{T}}\cdot\mathbf{b}$ relates to the independent noise input $\mathbf{n}$, we can put these steps together:



Note that we are considering the case when the null hypothesis is true. This is why the input here consists only in the independent noise $\mathbf{n}$. The independent noise goes through a sequence of stages of linear mixing, which can be summarized as a multiplication by a single row vector ($\mathbf{c}^{\mathbf{T}}\cdot(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\cdot\mathbf{X}^{\mathbf{T}}\cdot\mathbf{S}$), which contains a weight for each time point. In other words, the contrast estimate is a linear combination of the noise $\mathbf{n}$, using the weights in this row vector.

For independent selection and test statistics (i.e. contrast orthogonality) $\mathbf{c}_{selection}{}^{\mathbf{T}}\cdot(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\cdot\mathbf{X}^{\mathbf{T}}\cdot\mathbf{S}$ needs to be orthogonal to $\mathbf{c}_{test}{}^{\mathbf{T}}\cdot(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\cdot\mathbf{X}^{\mathbf{T}}\cdot\mathbf{S}$. This is equivalent to:

$$\mathbf{c}_{selection}{}^{\mathbf{T}}\cdot(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\cdot\mathbf{X}^{\mathbf{T}}\cdot\mathbf{S}\cdot\mathbf{S}^{\mathbf{T}}\cdot\mathbf{X}\cdot(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\cdot\mathbf{c}_{test} = 0.$$

Intuitively, the noise vector $\mathbf{n}$ is a point in the space spanned by the axes corresponding to the time points. If we repeat the experiment many times (under the null hypothesis), we sample many such points from an isotropic multivariate Gaussian distribution in this space. A contrast is a linear combination of the time points and corresponds to a projection of a point $\mathbf{n}$ onto a single dimension. Moreover if two contrasts are orthogonal (the contrasts, not the contrast weight vectors), then they correspond to projections onto orthogonal dimensions. These projections will be uncorrelated across repetitions of the experiment, because the underlying distribution is isotropic.

We see that contrast-vector orthogonality, i.e. $\mathbf{c}_{selection}{}^{\mathbf{T}}\cdot\mathbf{c}_{test} = 0$, does not ensure independent statistics. To determine whether selection and test contrasts are independent, we also need to consider the design matrix $\mathbf{X}$ and the time-point mixing matrix $\mathbf{S}$.

11. How can the design matrix make orthogonal contrast vectors yield dependent estimates?

For an intuitive understanding of the effect of the design matrix, consider the selection contrast A+B ($\mathbf{c}_{selection} = [1\ 1]^{\mathbf{T}}$) and the test contrast A-B ($\mathbf{c}_{test} = [1\ {-1}]^{\mathbf{T}}$). These are orthogonal contrast vectors, i.e. their inner product is zero: $\mathbf{c}_{selection}{}^{\mathbf{T}}\cdot\mathbf{c}_{test} = 0$. However, selecting by the contrast A+B can bias testing the contrast A-B if the design is not balanced with respect to A and B. Consider the case when there is more data for condition A than for condition B. Condition B will be less stably estimated. As a result selecting by contrast A+B will favor channels (e.g. voxels) having high positive noise in condition B. High positive noise in condition A will not be favored equally since the noise is more strongly reduced by averaging for condition A. As a result, the contrast A-B will be negatively biased.

Nature Neuroscience: doi:10.1038/nn.2303

For an ordinary-least-squares analysis, the effect of the design matrix can be taken into account by using the simplified criterion $\mathbf{c}_{selection}^{T} \cdot (\mathbf{X}^{T}\mathbf{X})^{-1} \cdot \mathbf{c}_{test} = 0$. This criterion does not account for the temporal dependency of the noise.

In practice, many selective analyses will not meet this criterion. For example, the experiment may consist in measuring brain activity while objects of three different categories (faces, places, and objects) are visually presented. Let us assume the experiment is analyzed with a linear model comprising three predictors, one for each category. It would seem innocuous to select visually responsive brain regions by the contrast $\mathbf{c}_{selection} = [1\ 1\ 1]^{T}$ and then selectively analyze those regions for the difference between the activity elicited by faces and the activity elicited by other images on average, using the contrast $\mathbf{c}_{test} = [1\ -0.5\ -0.5]^{T}$. The two contrast vectors are orthogonal. However, if the three categories of image had been presented for different amounts of time (e.g. face block, place block, face block, object block – giving faces as much time as the other two combined), then the selection by visual responsiveness would certainly bias the test contrast. Caution is also required, when behavioral measures (such as task errors, subjective judgments, or reaction times) are used either as covariates or to define classes of trials, which are to be modeled using separate predictors. In these cases, orthogonal contrast vectors will usually yield dependent selection and test statistics.

12. How can temporal noise dependency make orthogonal contrast vectors yield dependent estimates?

As explained in the answer to Question 10, temporal noise dependency (which can be characterized by a time-point mixing matrix $\mathbf{S}$) is one of the factors that can render two contrast estimates dependent (for repetitions of the experiment under the null hypothesis). The full orthogonality criterion $\mathbf{c}_{selection}^{T} \cdot (\mathbf{X}^{T}\mathbf{X})^{-1} \cdot \mathbf{X}^{T} \cdot \mathbf{S} \cdot \mathbf{S}^{T} \cdot \mathbf{X} \cdot (\mathbf{X}^{T}\mathbf{X})^{-1} \cdot \mathbf{c}_{test} = 0$ (introduced above, in the answer to Question 10) takes both the design matrix and temporal noise dependency into account for ordinary least-squares analyses.

Importantly, temporal dependency is not a concern in second-level between-subject analyses. Moreover, the biases caused by temporal dependency might be small for fMRI analyses, especially when block designs are used – but it is advisable to confirm this expectation for each particular data set, design matrix, and contrast pair used for selection and test.

13. Can an omnibus F test safely be used to select channels for a subsequent selective analysis?

No. An omnibus F test determines whether the model as a whole explains significant variance in the data. It is, thus, sensitive to all effects modeled by the design matrix. Selection according to the omnibus F value will select channels whose data best conform to the model.

Consider the case of a single-predictor model (e.g. stimulation versus baseline). The omnibus F statistic will select channels that exhibit either positive or negative effects. Although there is no preference for either effect direction, channels with effect estimates close to zero will not be selected. If we were to test selected channels, t values under the null hypothesis would not follow a t distribution. In the extreme

Nature Neuroscience: doi:10.1038/nn.2303

case of selection of a single channel among many candidates, the null distribution will be bimodal instead: The selected channel will tend to contain data showing either a large positive or a large negative effect, even if the null hypothesis is true. A t test applied to the selected channel will therefore no longer provide valid inference: The false-positives rate would be inflated for one-sided t tests in either direction and for two-sided t tests as well.

Consider the case of two conditions A and B in an fMRI experiment. We could model the responses by means of two hemodynamic response predictors, one for each condition. Selection by the omnibus F statistic will bias subsequent selective analysis in complex ways. One might expect equal and canceling biases for the contrasts A-B and B-A. However, this depends on the design matrix (in particular, it does not hold when there are different numbers of repetitions for the two conditions). In any case, as in the previous example of the single-predictor model, the distribution of effect statistics under the null hypothesis will be affected by selection, thus invalidating statistical tests.

In general, channels whose noise component happens to better conform to the model will be favored when selection is performed using an omnibus F statistic. The selection can therefore cause biases in favor of all effects modeled. The selected channels will tend to look more like linear combinations of the model's predictors under the null hypothesis than unselected noise data would. For example, if the model assumes a temporal shape for response, the selected time courses will tend to exhibit that shape even when the data are pure noise.

14. Can correlations between regional activation and subject covariates (such as personality traits) be affected by circularity?

Yes. For example, we may perform a group-statistical mapping that localizes a region whose activation in some task is correlated with a personality trait across subjects. An ROI analysis will be affected by circularity, if it is related to the personality trait used for defining the ROI. In particular, plotting the ROI activation as a function of the personality trait would be misleading, the correlation would be inflated, and a test of it invalid.

Forms of circular analysis and severity of biases

15. What are the different forms of circularity and how prevalent are they in the systems neuroscience literature?

Some forms of circularity that occur in systems neuroscience are as listed below. We have ordered them according to our sense of how prevalent they are in systems neuroscience (from rare to frequent). Note that this order is not based on any quantitative analysis of the literature, but on our subjective impression. We suspect that the most prevalent circular practices (at the bottom of the list) are those associated with less extreme distortions.

- same data used for training and testing a classifier (extreme distortion, rare error)

Nature Neuroscience: doi:10.1038/nn.2303

- non-independent data used for training and testing
- all data used to define the ROI for a classifier analysis with independent training and test sets
- set averages analyzed on the same data used for sorting voxels (or neurons) into the sets
- example neurons selectively analyzed after statistical selection using the same data
- ROI-average activation regressed onto some factor that is related to the ROI-definition contrast
- descriptive or inferential analysis of ROI-average activation not independent of ROI definition (smaller distortions, very widespread)

16. What determines the severity of the distortion resulting from circular analysis?

The magnitude of the distortions incurred by circular analysis will depend on the complexity of the model. A more complex model (i.e. one with more parameters) will tend to be more susceptible to overfitting, producing more strongly distorted results.

The model here is often a weight vector that determines the relative influence of the response channels. The definition of an ROI is a special case, where the weights are binary. Model complexity, then, is dependent on the number of channels selected from. Systems neuroscience often deals with many channels of brain activity data of which a small subset is selectively analyzed. This is one reason why our analyses can be quite susceptible to selection bias. Greater distortions are to be expected for nonindependent selective analyses based on more channels (e.g. high-resolution fMRI).

The effective complexity of the model (and with it the magnitude of the distortions) will be reduced by constraints that regularize the selection (such as spatial contiguity of ROI voxels or spatial smoothing of the data). For a given data set and selection contrast, a contiguous ROI (Example 2) will therefore be less severely overfitted than a discontiguous ROI (scattered set of voxels, Example 1). The contiguity constraint in effect regularizes the model fit, thus reducing overfitting. Similarly, if data are strongly smoothed, the precise shape of an ROI may have less of an influence on the result. This might reduce the effects of circular analysis, but will not eliminate them.

17. How strong are the biases caused by circularity really? Are they perhaps negligible in many analyses?

The magnitude of the distortions depends on many factors. It can be small and it can be large. To justify circular analysis, the magnitude of the distortions would need to be demonstrated to be negligible for the particular case at hand.

Nature Neuroscience: doi:10.1038/nn.2303

Dividing the data into independent sets

18. What is meant by “independence” in this context?

Independence in this context means *statistical independence between selection statistics and results statistics under the null hypothesis*. If selection and results statistics are not inherently independent, we can render them independent by using independent data to compute each. Even if the same statistic (e.g. the same contrast) is used for selection and selective analysis, selection and results statistics will be independent under the null hypothesis if *the noise is statistically independent between the two data sets*.

So when we say that a hypothesis needs to be tested with “independent data”, this is analogous to asking for an independent expert opinion: Two experts may give the same advice, but based on independent judgment. Similarly, real effects in the data will replicate, but each data set should have independent noise.

We need to imagine the null hypothesis to be true and then ask: might noise dependencies between the two data sets render the results statistics dependent on the selection statistics? If there are no noise dependencies between the data sets, the answer is no.

For example, if we divide an fMRI run into odd and even volumes, then temporally consecutive fMRI volumes will end up in different data sets. Since fMRI noise is temporally autocorrelated, the two data sets will have dependent noise. The noise dependency will almost certainly render effect statistics dependent, because each experimental event likewise affects several consecutive volumes. Imagine the effect of the noise at a single voxel and time point: A positive noise contribution (making the number measured larger than the true value) will tend to be associated with positive noise in the same voxel at the subsequent time point – which is in the other data set. Importantly, this positive noise effect is likely to affect the same condition in both data sets in the same direction. It will, thus, render any effect statistic based on that condition dependent between the data sets under the null hypothesis.

19. Are different sets of subjects required for truly independent data sets?

No. If the experiment is repeated with the same group of subjects, the repetition provides an independent replication for that group of subjects unless the noise in the data is dependent between the experiments. The same logic holds for a single-subject analysis.

It is important to note that the relevant notion of “dependent data sets” here is distinct from the notion of “dependent samples” used in the context of t tests. When the same subjects are measured repeatedly (e.g. before and after some treatment), a dependent t test is appropriate for comparing the two samples. In a dependent t test, each sample corresponds to a separate condition. In the present context, each data set typically contains all conditions and independence means independence of the effect statistics under the null hypothesis, which holds when the noise is independent between the data sets.

Nature Neuroscience: doi:10.1038/nn.2303

If both data sets are from the same subject, they clearly have something in common: the subject. Beyond that, independent data sets typically have many other things in common: same species, same experiment, same measurement technique, same true effects replicated. However, none of these commonalities render the data sets dependent in the sense that matters here. The data sets can still be completely independent of each other in the sense of containing independent noise.

20. How can I make sure that the data sets to be used for selection and selective analysis have independent noise?

The answer depends on the statistical dependencies of the noise in the data. Dependent data points should be kept in the same set in dividing the data. Brain-activity time series often exhibit temporal autocorrelation restricted to small temporal lags. In that case, independence can be achieved by using temporal blocks with sufficient time margins between them.

In fMRI, a good way to divide the data is to number the scanner runs chronologically as measured and designate all odd runs as data set 1 and all even runs as data set 2. The odd-even scheme minimizes slow temporal confounds such as the subject's level of fatigue. For crossvalidation, similarly, a leave-one-*run*-out scheme is recommended. Since slow drifts often have a similar shape over the course of the fMRI run, it is advisable to use a different random condition sequence for each run.

In electrophysiology, data are sometimes divided on the fine time scale of single trials, e.g. by defining temporal windows within the response to a given trial. This may not yield independent data sets for selection and selective analysis because of the underlying physiology: measurements that are close in time may be dependent and can nevertheless end up in different sets. A better approach is to first divide the data into blocks of consecutive trials, ideally with a temporal margin that prevents dependencies between blocks. The set of blocks can then be divided into subsets for selection and testing. For example, the blocks could be chronologically numbered and divided into an odd and an even set, as suggested for fMRI runs above.

21. What is crossvalidation and how does it relate to data splitting?

Crossvalidation is a form of data splitting. (It thus falls under "independent split-data analysis" in Fig. 4.)

When we split the data into two independent sets, we may designate one set as the selection (or training) set and the other set as the test set. Obviously the opposite assignment of the two sets would be equally justified. Since the two assignments will not yield identical results, we are motivated to perform the analysis for each assignment and combine the results statistically, for greater power. This approach is the simplest form of crossvalidation: a 2-fold crossvalidation.

An n-fold crossvalidation generalizes this idea and allows us to use most of the data for selection (or training) and all of the data for selective analysis, while maintaining independence of the sets. For n-fold crossvalidation, we divide the data into n independent subsets. For each fold i=1..n, we use set i for selective analysis after using all other sets for selection (or training). Finally, the n selective analyses are

Nature Neuroscience: doi:10.1038/nn.2303

statistically combined. An n-fold crossvalidation for n>2 potentially confers greater power than a 2-fold crossvalidation, because the n-fold crossvalidation provides more data for selection (or training) on each fold.

Crossvalidation is a very general and powerful method widely used in statistical learning and pattern classification. However, it is somewhat cumbersome and computationally costly. While it is standard practice in pattern classification, it is not widely used for ROI definition in systems neuroscience. Perhaps it should be.

22. Isn't it cumbersome to repeat the selection process along with classifier training on each fold of crossvalidation?

Yes, it is cumbersome. The selected set will be different on each fold. However, it is necessary to do this when using crossvalidation. Example 1 has shown that selection using all data can entail extreme biases even when independent data sets are used for training and testing thereafter. Similar results would be obtained if crossvalidation were used to test the classifier, but the selection process were not included in the crossvalidation. Selection needs to be performed again on each fold of crossvalidation, only using that fold's training data. Selection is binary weighting and should be viewed as part of the training of a classifier.

One way to simplify things is to use only two data sets and use set 1 for selection and training and set 2 for testing. However, a cross-validation scheme, though cumbersome, can make more efficient use of the data, thus increasing power.

23. Could crossvalidation be used for ROI-average analyses nonindependent of the ROI-definition criterion?

This is a good idea in principle. The benefit would be an increase in power compared to an analysis with the data split into an ROI-definition and an ROI-test set. The cost would be a cumbersome and computationally more intensive analysis with the conceptual complication that the ROI would be slightly differently defined on each fold of the crossvalidation.

We are not aware of an implementation of this approach. Most studies using independent data for ROI analyses related to the ROI-definition criterion use one data set for ROI definition and an independent data set for ROI analysis, without utilizing crossvalidation.

Nature Neuroscience: doi:10.1038/nn.2303

Understanding circularity

24. Is every selective analysis affected by selection bias?

No, only *nonindependent* selective analysis is affected by selection bias. (Independent selective analysis is not affected by selection bias, but does raise the concern of selective reporting of accurate results. Selective reporting is not the focus of this paper, but does deserve a wider debate in systems neuroscience.)

An analysis is “selective” when a subset of the data is first selected from the full data set before performing secondary analyses on the selected data only. For example, in neuroimaging subsets of voxels are selected, in EEG and MEG, sensors and time windows are selected and in invasive electrophysiology, cells or sites are selected. (More generally, an analysis is also selective if the data channels, e.g. voxels, are differentially weighted for further analysis.)

A circular analysis is a selective analysis, in which the selection process biases the results of the selective analysis. Because data always contain noise, the selected subset will never be determined by real effects only. Even in the absence of any real effects, the selected data will show the tendencies they were selected for.

One way to avoid selection bias is to ensure that the data used for selection is independent of the data on which further selective analysis is performed. A replication of the experiment, for example, provides an independent data set. Real effects, but not noise, will replicate. The results, thus, will reflect actual effects at the selected sites, without bias due to the influence of noise on the selection.

A selective analysis performed on the same data as used for selection will be biased unless the statistics of the selective analysis are inherently independent of the statistics used for selection. Whether this is the case is not in general obvious and needs to be explicitly demonstrated. In particular, using orthogonal contrast vectors for selection and test does not ensure independence (see Question 10: *Do orthogonal contrast vectors ensure contrast orthogonality?* and Fig. S3).

In sum, nonindependent selective analyses are circular. Independence of a selective analysis can be ensured by using independent data or by demonstrating inherent independence between the statistics used for selection and selective analysis.

25. Can the distortion caused by selection be quantitatively modeled and corrected for?

In principle, the distortion caused by circularity can be modeled and corrected for (see *What if all this fails?* in section *A policy for noncircular analysis* and Fig. 4). However, this approach is not widely used. Modeling the distortion caused by selection would need to take into account the specifics of each particular case. The inherent nonlinearity of selection makes the process somewhat challenging to model. Appropriate modeling may require simulation and resampling techniques. The methods would need to be developed and validated.

Nature Neuroscience: doi:10.1038/nn.2303

26. Can a selective analysis confirm an effect selected for without valid statistical inference correcting for multiple tests?

Yes, a selective analysis can indeed confirm an effect selected for without valid statistical inference correcting for multiple tests. However, the selective analysis needs to be based on independent data.

As explained under *Example 2: Regional activation analysis*, however, in practice selection is sometimes performed without adequate statistical inference (correcting for multiple tests) and the selective analysis of the same data is then interpreted as though it confirmed the effect selected for. While it does not confirm the effect, the selective analysis effectively serves to help us forget about the multiple testing, which was inadequately accounted for during selection. Independent data would be required to confirm the effect. In fact, the inadequacy of the inference during selection will compound the circularity of the selective analysis and strong biases as well as large false-positives rates are to be expected.

27. How is the multiple-testing problem related to circular analysis?

The multiple-testing problem is closely related to the circularity of nonindependent selective analysis. To see this, consider the case of selecting a single response channel (e.g. a single neuron or neuroimaging voxel) from a large set by testing each with a t test using a threshold corresponding to $p<0.01$. The selective analysis in this case just repeats the analysis used for selection, so the selected channel, by definition, will have $p<0.01$. The analysis of the selected channel is not valid for the same reason that the selection does not provide valid inference: neither of them takes into account the multiple tests performed for the purpose of selection.

For the case of mapping a volume by testing at each location, standard methods for multiple testing can be applied. These include Bonferroni adjustment, Gaussian field theory,[7,8,9] cluster-size thresholding,[10,11,12,13] permutation methods,[14,15] control of the false-discovery rate,[16] and Bayesian techniques.[17] When multiple channels are selected and jointly analyzed, these methods do not apply in a straightforward manner. The effects of selection under the null hypothesis could be modeled by simulation in principle. Otherwise selective analyses of statistics related to the selection criterion are best performed on independent data.

28. What is selective reporting and how is it related to bias of nonindependent selective analysis?

By “selective reporting”, we refer to the reporting of accurate results, which are selected from a larger set of results that could have been reported. This is like taking photos and including only a selection of them in a report. In systems neuroscience, selective reporting occurs when cells or voxels are selected, and then subjected to an independent (noncircular) selective analyses. While the selection may bias the general conclusion, the results themselves are accurate. Nonselective analyses such as statistical mapping with correction for multiple tests, can substantially reduce the bias of selective reporting. Note

Nature Neuroscience: doi:10.1038/nn.2303

however, that the set of contrasts investigated and the conditions included in the experiment still reflect preconceived notions of the subject matter and bias the scientific process. In sum, the bias of selective reporting is an important issue that deserves a debate. However, there may not be a solution to this problem.

The bias of nonindependent selective analysis, which is the topic of this paper, is of a different nature: It is a statistical bias that distorts the magnitudes of effect estimates and invalidates statistical inference. Whereas, the results in selective reporting can be correct results, results from nonindependent selective analyses are not correct results. This is like retouching photos before inclusion in a report. However, that comparison would suggest intent, whereas most circular analyses in systems neuroscience occur inadvertently.

Preventing circularity

29. What can researchers do to prevent circular analyses?

- Don't ask: *How could the selection possibly bias my results?*, ask: *How can I be sure that the selection cannot possibly bias my results?* (Selection effects can be hard to understand, imagine, or predict. So when it's hard to see how, it can still be happening.)
- Consider adopting a zero-tolerance position on circular analyses and choose a policy to prevent circularity (e.g. the one we suggest in this paper).
- Test your complete analysis (including all selection stages) on null data from a random generator or from a brain region where no effect is expected. This can help catch statistical circularities. (Unfortunately, the absence of a bias in such a test does not indicate that the analysis in noncircular. Therefore such a test by itself is not sufficient to justify an analysis.)

30. What can authors do to allow readers to assess whether their results are circular?

- For each selective analysis, communicate clearly (1) how the channels have been selected and (2) whether independent data were used for selection and selective analysis. Consider placing this information in the relevant figure panel itself. At least the figure legend should give this information.
- Explain why each of your selective analyses cannot be affected by selection bias. Ideally, include proof of independence of the statistics used for selection and selective analysis for each same-data selective analysis (see Question 10).
- If an adequate explanation of why results cannot possibly be affected by circularity requires more space than available for the text of the paper, consider elaborating on the issue in the Methods section or Supplementary Information.

31. How can readers and reviewers recognize circular analyses?

- Consider the possibility of results arising purely or partially from a statistical circularity.
- Do not accept a result unless you are sure it cannot have arisen from circular analysis.

Nature Neuroscience: doi:10.1038/nn.2303

- Ensure that authors provide all the information necessary to assess circularity.
- Value statistical correctness over clean and strong appearance of the results. (The latter preference favors circular practices.)
- Complex apparently theory-confirming presences and absences of effects can arise under the null hypothesis if certain effects, but not others, have been selected for.
- Effects can appear to replicate across independent data sets, if all data have been used for selection (as demonstrated in Example 1 in the paper, where patterns appeared to replicate from odd to even runs).
- Distortions caused by circularity can be complex and need not consist simply in an exaggeration of the effects selected for.

32. What caveats on circularity need to be considered in pattern-information analyses?

- Pattern-information analyses are powerful and also very sensitive to nonindependence errors.
- Very large spurious effects can occur if training and test sets are not strictly independent.
- Selection is binary weighting. Like classifier training (which, for linear classifiers, yields a set of weights), voxel selection will be affected by overfitting. Whenever the selection criteria are not proven to be inherently independent of the pattern-information statistics, voxel selection must not be performed on the same data as used in a second step for classifier training and testing.
- Instead, voxel selection must be considered part of the training procedure. The test set must be independent (on each fold, if crossvalidation is used) of the set used for selection and training.
- One option is to use one data set for selection and training and an independent one for testing. This is costly in terms of the data.
- Another option is to use crossvalidation. In this case selection needs to be repeated on each fold of crossvalidation. This is cumbersome, as the selected voxels (or channels) will change on each fold of crossvalidation.
- Crossvalidation may have been correctly used in either voxel selection or pattern-information analysis or both. However, if the same (or overlapping) data have been used for selection and pattern-information analysis, large spurious effects are to be expected.
- Distortions will be greater if selection occurs among more channels or among noisier channels. (High-resolution fMRI is particularly vulnerable here, because there are more and noisier voxels.)
- Distortions will tend to be greater for discontiguously selected voxel sets than for solid, blob-shaped ROIs. (This is because the effective complexity of the ROI model is greater, thus the ROI definition will be more severely affected by overfitting.)
- If the same analysis (including voxel selection and selective pattern-information analysis) is applied to multiple brain regions and fails to show effects in some of them, this is consistent with a noncircular analysis. (However, it does not suffice to establish that the analysis is noncircular.)
- Complex apparently theory-confirming effects (including accurate cross-decoding between independent sets of conditions) can result from selection of response channels (e.g. voxels) if selection is based on data including any of the test data.

Nature Neuroscience: doi:10.1038/nn.2303

Supplementary Figures

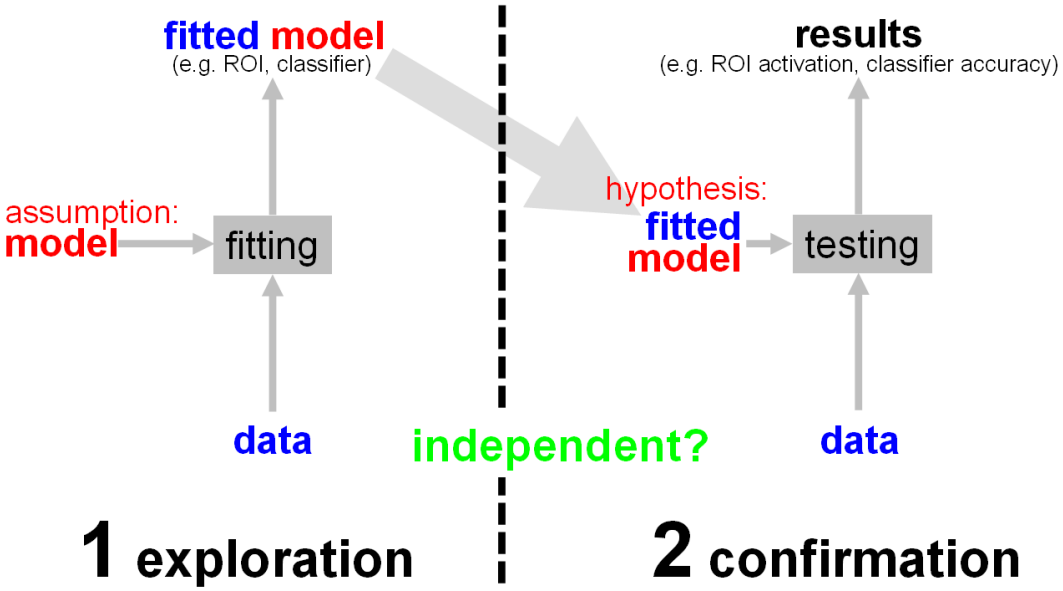


Fig. S1 | Selection, weighting, and sorting can be viewed as exploratory analyses requiring independent confirmation. Selection (e.g. of voxels for ROI definition), weighting (e.g. in linear classification of activity patterns), and sorting (e.g. of neurons according to their tuning) can all be viewed as exploratory analyses (left) that fit a model to the data so as to generate a specific hypothesis (the model with its parameters fitted), which is to be confirmed by a subsequent test (right). The weighting of data in linear classification is conventionally viewed in this context. ROI definition can be construed as a special case of weighting, where the weights are binary. Sorting, similarly, can be viewed as defining multiple sets of binary weights. In each case, the weights will be overfitted to some degree. In other words, the hypothesis generated (the fitted model) will reflect the noise in the data to some degree. We therefore need independent data to confirm the hypothesis (right). (In ROI analysis, for example, the hypothesis generated is the specific set of voxels that defines the ROI. Note that this is related to, but distinct from the hypothesis confirmed by statistical mapping, which states merely that there is a blob of activation. In ROI analysis, we often wish to test additional hypotheses that presuppose the same ROI, e.g. ROI-average contrasts other than the selection contrast. One safe way to ensure that results are not distorted and tests invalidated by circularity is to use independent data.) For details, see section *The cycle of exploration and confirmation*.

Nature Neuroscience: doi:10.1038/nn.2303

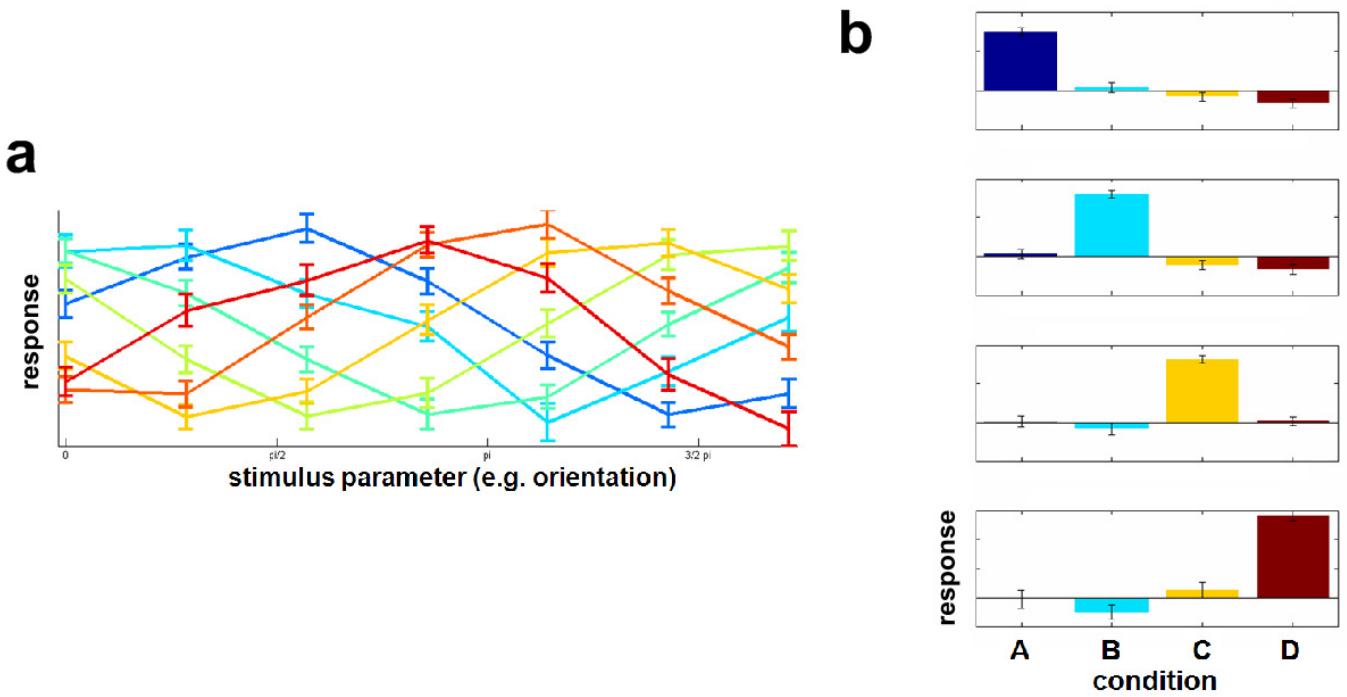


Fig. S2 | Example 3: Sorting of response channels can spuriously suggest response tuning. Two simulations demonstrate the effect of sorting response channels (e.g. single neurons or fMRI voxels) by their response profile across a set of experimental conditions. In each case the set-average responses are shown across experimental conditions with error bars indicating +/- 1 standard error of the mean. The analysis would suggest strong and highly significant tuning in both cases. However, the analysis is based on Gaussian noise containing no real effects. Using independent data to estimate the set-average response profile across conditions would reveal that there are no tuning effects in the data. Note that the distortion is extreme. The error here is more easily understood than the subtler errors in Examples 1 and 2 in the paper. These sorting examples appear in the Supplementary Information because a similar case has been examined before[18] and our goal in this paper is to explain the less obvious cases, which are more critical to rooting out the problem of circularity. (**a**) Set-average tuning curves for response channels sorted according to their tuning (noise data). 500 response channels have been assigned random responses from a Gaussian distribution for each of seven conditions (which could correspond, for example, to stimulus orientation). Each channel's response profile across the conditions was correlated with seven sinusoidal tuning curves, each peaking at a different condition. The response channel was then assigned to a set corresponding to the best fitting tuning curve. The plot shows the average tuning curve for each of the seven sets (colors). (**b**) Set-average response profiles across four conditions for response channels sorted according to their response profile (noise data). 500 response channels have been assigned random responses from a Gaussian distribution for each of four experimental conditions. Each channel was assigned to one of four sets depending on the condition for which it had the maximum response. Each bar graph shows the average response profile for one of the four sets.

Nature Neuroscience: doi:10.1038/nn.2303

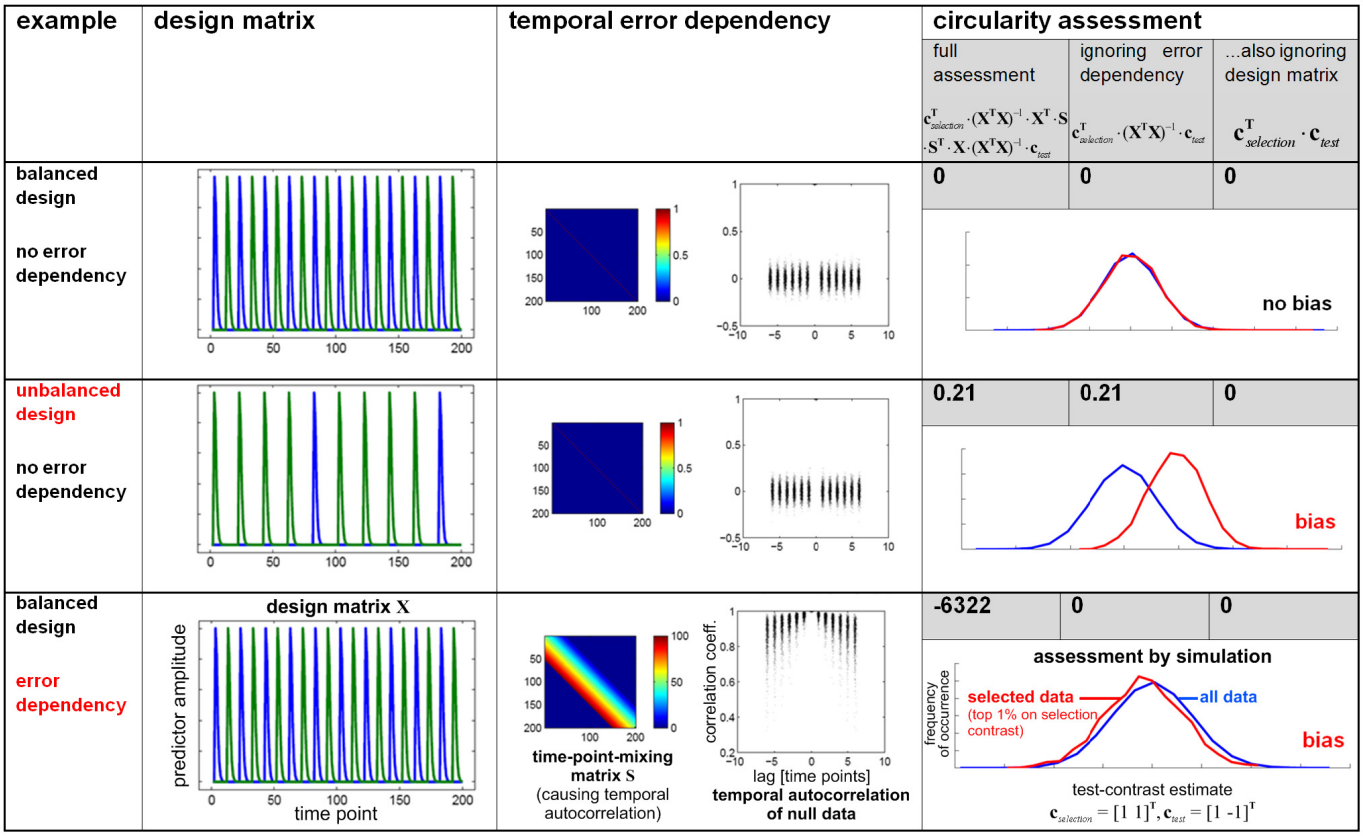


Fig. S3 | Selection can cause bias even when selection and test contrast vectors are orthogonal. Here we simulate the effect of selection among pure noise responses for a two-condition (A, B) experiment with 200 time points. The selection contrast is A+B ($\mathbf{c}_{selection} = [1\ 1]^{\mathbf{T}}$); the test contrast is A-B ($\mathbf{c}_{test} = [1\ -1]^{\mathbf{T}}$). These are orthogonal contrast vectors, i.e. their inner product is zero: $\mathbf{c}_{selection}{}^{\mathbf{T}} \cdot \mathbf{c}_{test} = 0$. We simulate the effect of selection by (1) generating 200,000 Gaussian random time courses of 200 time points each (without spatial or temporal dependencies), (2) optionally introducing temporal dependencies by multiplication of each time course by a time-point-mixing matrix $\mathbf{S}$, (3) analyzing each time course with a design matrix $\mathbf{X}$ (second column from left) to estimate the selection contrast, (4) selecting 1% of the time courses with the highest selection-contrast estimates, and (5) computing the test-contrast estimates for the selected time courses and also for the other time courses. The right column shows the histograms of the test-contrast estimates for all time courses (blue) and selected time courses (red). A deviation between the red and the blue histograms indicates selection bias. The three rows correspond to three scenarios described in the left column. When the experimental design is balanced (i.e. symmetrical with respect to the two conditions A and B) and there are no error dependencies (top row), selecting with contrast A+B does not bias contrast A-B. However, when either the design is not balanced (here we simulated more repetitions for one of the conditions) or there are substantial error dependencies, selecting with contrast A+B does bias contrast A-B. The central column characterizes the temporal error dependencies for each scenario. The gray-shaded cells on the right show the values of three analytical circularity criteria (zero indicates no bias, nonzero indicates bias). These three analytical criteria are motivated and derived in section *A policy for noncircular analysis* and *Questions 10-12* above. The analytical criterion taking design and error dependency into account ("full assessment", leftmost of the three) is consistent with the simulation results in terms of its prediction of bias in all three cases.

Nature Neuroscience: doi:10.1038/nn.2303

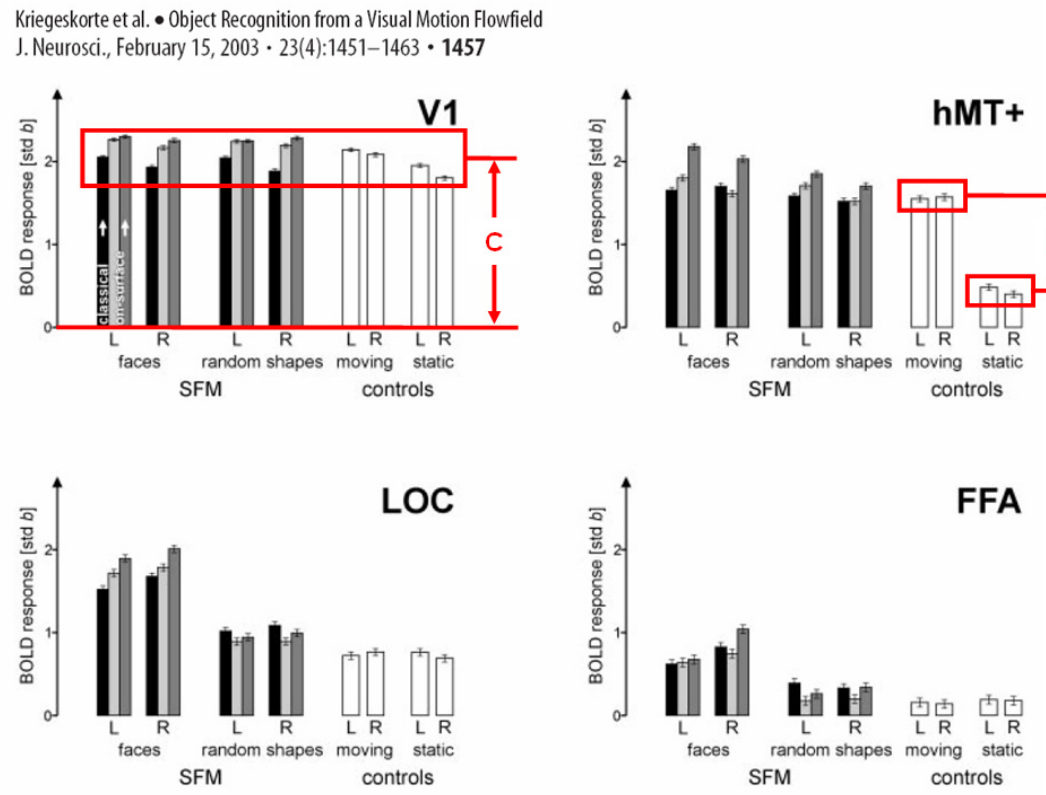


Figure 3. Key-region response profiles (group results). Responses to SFM object and control stimuli in regions of interest as reflected in the linear-regression standardized *b* weights (β estimates) averaged across subjects are shown. Group averaging is based not on Talairach correspondence but on individual localization of the key regions in each subject. V1 has been localized anatomically; hMT+, LOC, and FFA have been localized functionally (see Materials and Methods). For each subject and region, the *b* weight entering into the average has been obtained by multiple-regression analysis of the spatially averaged time course. Error bars indicate the SE of the average *b* weight. *L* and *R* indicate left and right hemisphere responses, respectively. *Black bars* represent responses to classical SFM stimuli; *gray bars* represent responses to on-surface SFM stimuli. *Light-gray bars* represent stationary-implicit-object on-surface SFM responses; *dark-gray bars* represent moving-implicit-object on-surface SFM responses. *White bars* represent responses to control stimuli as labeled (for details, see Statistical analysis in Materials and Methods).

Fig. S4 | Circularity indicators can serve to highlight all aspects of the results that are affected by circularity. The figure above is adapted from a published paper.[19] It serves as an example of a widespread and relatively benign variant of the error of nonindependent selective analysis, where some of the effects shown are biased, but the effects essential to the conclusions are not affected. The figure has been modified only by adding circularity indicators (red) to highlight contrasts that are expected to be affected by a bias because they are related to the selection criterion and the results shown are based on the same data that was used for selection. Note that the indicators very effectively communicate the affected aspects of the bar graph and also inform us about the selection criterion. The ROIs for the lower two panels were defined by independent localizer experiments, as suggested by the absence of circularity indicators. Ideally, circular analysis should be avoided altogether, so circularity indicators are never needed. However, if the essential results are unaffected and the circular aspects are not interpreted or tested, circularity indicators provide a last-resort mechanism for preventing inappropriate interpretations. They should have been added to the figure before publication. (Note that authors sometimes call for caution in interpreting their own findings by stating the circularity of certain aspects of their analysis. One recent example is in the legend of Fig. 4 of ref. 20.)

Nature Neuroscience: doi:10.1038/nn.2303

References

1. Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P., Henson, R.N. (2006). A critique of functional localisers. NeuroImage 30(4), 1077-87.

2. Friston, K., Holmes, A., Price, C., Büchel, C., and Worsley, K. (1999). Multisubject fMRI studies and conjunction analyses. NeuroImage 10, 85-396.

3. Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J. B. (2005). Valid Conjunction Inference with the Minimum Statistic. NeuroImage 25(3), 653-660.

4. Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA 103, 3863-3868.

5. Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford University Press.

6. Cover, T.M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. IEEE Transactions on Electronic Computers, EC-14(3), 326-334.

7. Worsley, K.J., Evans, A.C., Marrett, S., and Neelin P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900-918.

8. Friston, K.J., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. Hum. Brain Mapp. 1, 153-171.

9. Worsley, K.J., and Friston, K.J. (1995). Analysis of fMRI time-series revisited – again. Neuroimage 2, 173-181.

10. Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med. 33(5): 636-47.

11. Poline, J.B., Worsley, K.J., Evans, A.C., Friston, K.J. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. Neuroimage 5(2): 83-96.

12. Cox, R.W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research 29 (3), pp. 162-173.

13. Ward, D.B. (2000). Simultaneous inference for fMRI data. http://afni.nimh.nih.gov/pub/dist/doc/manual/AlphaSim.pdf

14. Nichols, T.E., and Hayasaka, S. (2003). Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. Statistical Methods in Medical Research 12(5), 419-446.

Nature Neuroscience: doi:10.1038/nn.2303

15. Nichols, T.E., and Holmes, A.P. (2002). Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples. Human Brain Mapping 15, 1-25.

16. Genovese, C.R., Lazar, N. A., and Nichols T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15(4), 870-8.

17. Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Smith, S.M. (2005). Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. IEEE Trans Med Imaging 24(1): 1-11.

18. Baker, C.I., Hutchison, T.L., Kanwisher, N. (2007). Does the fusiform face area contain subregions highly selective for nonfaces? Nat Neurosci 10(1), 3-4.

19. Kriegeskorte, N., Sorger, B., Naumer, M., Schwarzbach, J., van den Boogert, E., Hussy, W., Goebel, R. (2003). Human cortical object recognition from a visual motion flowfield. J Neurosci 23(4), 1451-63.

20. Serences, J.T. (2008). Value-Based Modulations in Human Visual Cortex. Neuron 60, 1169-1181.

Nature Neuroscience: doi:10.1038/nn.2303