Large-scale automated synthesis of human functional neuroimaging data

Tal Yarkoni[1], Russell A Poldrack[2–4], Thomas E Nichols[5,6], David C Van Essen[7] & Tor D Wager[1]

The rapid growth of the literature on neuroimaging in humans has led to major advances in our understanding of human brain function but has also made it increasingly difficult to aggregate and synthesize neuroimaging findings. Here we describe and validate an automated brain-mapping framework that uses text-mining, meta-analysis and machine-learning techniques to generate a large database of mappings between neural and cognitive states. We show that our approach can be used to automatically conduct large-scale, high-quality neuroimaging meta-analyses, address long-standing inferential problems in the neuroimaging literature and support accurate 'decoding' of broad cognitive states from brain activity in both entire studies and individual human subjects. Collectively, our results have validated a powerful and generative framework for synthesizing human neuroimaging data on an unprecedented scale.

The development of noninvasive neuroimaging techniques such as functional magnetic resonance imaging (fMRI) has spurred rapid growth of literature on human brain imaging in recent years. In 2010 alone, more than 1,000 fMRI articles had been published[1]. This proliferation has led to substantial advances in our understanding of the human brain and cognitive function; however, it has also introduced important challenges. In place of too little data, researchers are now besieged with too much. Because individual neuroimaging studies are often underpowered and have relatively high false positive rates[2–4], multiple studies are required to achieve consensus regarding even broad relationships between brain and cognitive function. It is therefore necessary to develop new techniques for the large-scale aggregation and synthesis of human neuroimaging data[4–6].

Here we describe and validate a new framework for brain mapping, NeuroSynth, that takes an instrumental step toward automated large-scale synthesis of the neuroimaging literature. NeuroSynth combines text-mining, meta-analysis and machine-learning techniques to generate probabilistic mappings between cognitive and neural states that can be used for a broad range of neuroimaging applications. Whereas previous approaches have relied heavily on researchers' manual efforts (for example, refs. 7,8), which limits the scope and efficiency of resulting analyses[1], our framework is fully automated and allows rapid and scalable synthesis of the neuroimaging literature. We show that this framework can be used to generate large-scale meta-analyses for hundreds of broad psychological concepts; support quantitative inferences about the consistency and specificity with which different cognitive processes elicit regional changes in brain activity; and decode and classify broad cognitive states in new data solely on the basis of observed brain activity.

RESULTS

Overview

Our methodological approach includes several steps (**Fig. 1a**). First, we used text-mining techniques to identify neuroimaging studies that used specific terms of interest (for example, 'pain', 'emotion', 'working memory' and so on) at a high frequency (>1 in 1,000 words) in the article text. Second, we automatically extracted activation coordinates from all tables reported in these studies. This approach produced a large database of term-to-coordinate mappings; here we report results based on 100,953 activation foci drawn from 3,489 neuroimaging studies published in 17 journals (Online Methods). Third, we conducted automated meta-analyses of hundreds of psychological concepts, producing an extensive set of whole-brain images that quantified relationships between brain activity and cognition (**Fig. 1b**). Finally, we used a machine-learning technique (naive Bayes classification) to estimate the likelihood that new activation maps were associated with specific psychological terms, which allowed relatively open-ended decoding of psychological constructs from patterns of brain activity (**Fig. 1c**).

Automated coordinate extraction

Our approach differs from previous work in its heavy reliance on automatically extracted information, raising several potential concerns about data quality. For example, the software might incorrectly classify noncoordinate information in a table as an activation focus (a false positive); different articles report foci in different stereotactic spaces, resulting in potential discrepancies between anatomical locations represented by the same set of coordinates; and the software did not discriminate activations from deactivations.

[1]Department of Psychology and Neuroscience, University of Colorado at Boulder, Boulder, Colorado, USA. [2]Imaging Research Center, University of Texas at Austin, Austin, Texas, USA. [3]Department of Psychology, University of Texas at Austin, Austin, Texas, USA. [4]Department of Neurobiology, University of Texas at Austin, Austin, Texas, USA. [5]Department of Statistics, University of Warwick, Coventry, UK. [6]Warwick Manufacturing Group, University of Warwick, Coventry, UK. [7]Department of Anatomy and Neurobiology, Washington University School of Medicine, St. Louis, Missouri, USA. Correspondence should be addressed to T.Y. (tal.yarkoni@colorado.edu).

RECEIVED 24 JANUARY; ACCEPTED 24 MAY; PUBLISHED ONLINE 26 JUNE 2011; DOI:10.1038/NMETH.1635

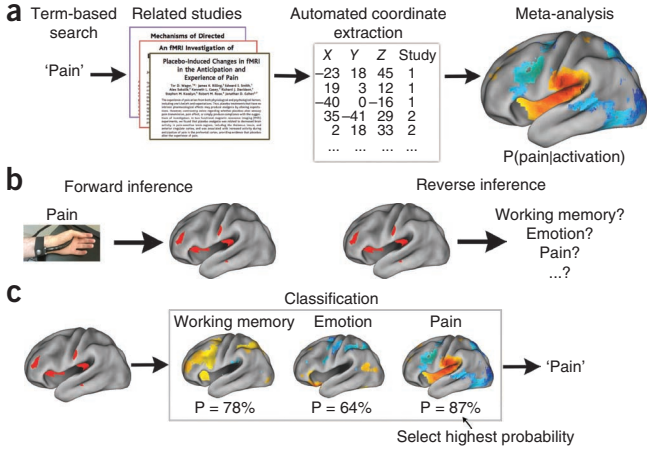


Figure 1 | Schematic overview of NeuroSynth framework and applications. (**a**) Outline of the NeuroSynth approach. The full text of a large corpus of articles is retrieved and terms of scientific interest are stored in a database. Articles are retrieved from the database on the basis of a user-entered search string (for example, 'pain') and peak coordinates from the associated articles are extracted from tables. A meta-analysis of the peak coordinates is automatically performed, producing a whole-brain map of the posterior probability of the term given activation at each voxel (P(pain|activation)). (**b**) Outlines of forward and reverse inference in brain imaging. Given a known psychological manipulation, one can quantify the corresponding changes in brain activity and generate a forward inference, but given an observed pattern of activity, drawing a reverse inference about associated cognitive states is more difficult because multiple cognitive states could have similar neural signatures. (**c**) Given meta-analytic posterior probability maps for multiple terms (for example, working memory, emotion and pain), one can classify a new activation map by identifying the class with the highest probability, P, given the new data (in this example, pain).

To assess the effect of these issues on data quality, we conducted supporting analyses (**Supplementary Note**). First, we compared automatically extracted coordinates with a reference set of manually entered foci in the Surface Management System Database (SumsDB)[7,9], and found high rates of sensitivity (84%) and specificity (97%). Second, we quantified the proportion of activation increases versus decreases reported in the neuroimaging literature. We found that decreases constituted a small proportion of results and had minimal effect on our results. Third, we developed a preliminary algorithm (based on ref. 10) to automatically detect and correct for between-study differences in stereotactic space (**Supplementary Fig. 1**). Although automated extraction missed a minority of valid coordinates, and work remains to be done to increase the specificity of the extracted information, most coordinates were extracted accurately and several factors of a priori concern had relatively small influences on the results.

Large-scale automated meta-analysis

We used the database of automatically extracted activation coordinates to conduct a comprehensive set of automated meta-analyses for several hundred terms of interest. For each term, we identified all studies that used the term at high frequency anywhere in the article text[11] and submitted all associated activation foci to a meta-analysis. This approach generated whole-brain maps that showed the strength of association between each term and every location in the brain, enabling us to make multiple kinds of quantitative inference (for example, if the term 'language' had been used in a study, how likely was the study to report activation in Broca's area? If activation had been observed in the amygdala, what was the probability that the study frequently used the term 'fear'?).

To validate this automated approach, which rests on the assumption that simple word counts are a reasonable proxy for the substantive content of articles, we conducted several supporting analyses (**Supplementary Note**). First, we found that NeuroSynth accurately recaptured conventional boundaries between distinct anatomical regions by comparing lexically defined regions of interest to anatomically defined regions of interest (**Supplementary Fig. 2**). Second, we used NeuroSynth to replicate previous findings of visual category-specific activation

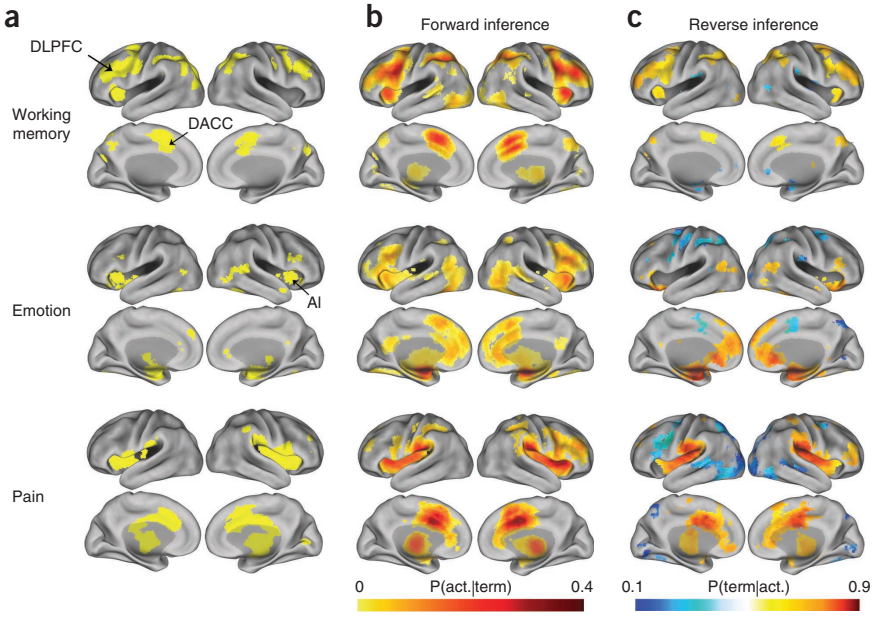


Figure 2 | Comparison of previous meta-analysis results with forward and reverse inference maps produced automatically using the NeuroSynth framework. (**a**) Meta-analytic maps produced manually in previous studies[14–16]. (**b**) Automatically generated forward inference maps showing the probability of activation given the presence of the term (P(act.|term)). (**c**) Automatically generated reverse inference maps showing the probability of the term given observed activation (P(term|act.)). Meta-analyses were carried out for working memory (top), emotion (middle) and physical pain (bottom) and mapped to the PALS-B12 atlas[30]. Regions in **b** were consistently associated with the term and regions in **c** were selectively associated with the term. To account for base differences in term frequencies, reverse inference maps assumed uniform priors (equal 50% probabilities of 'term' and 'no term'). Activation in orange or red regions implies a high probability that a term is present, and activation in blue regions implies a high probability that a term is not present. Values for all images are shown only for regions that survived a test of association between term and activation, with a whole-brain correction for multiple comparisons (false discovery rate was 0.05). DLPFC, dorsolateral prefrontal cortex; DACC, dorsal anterior cingulate cortex; AI, anterior insula.

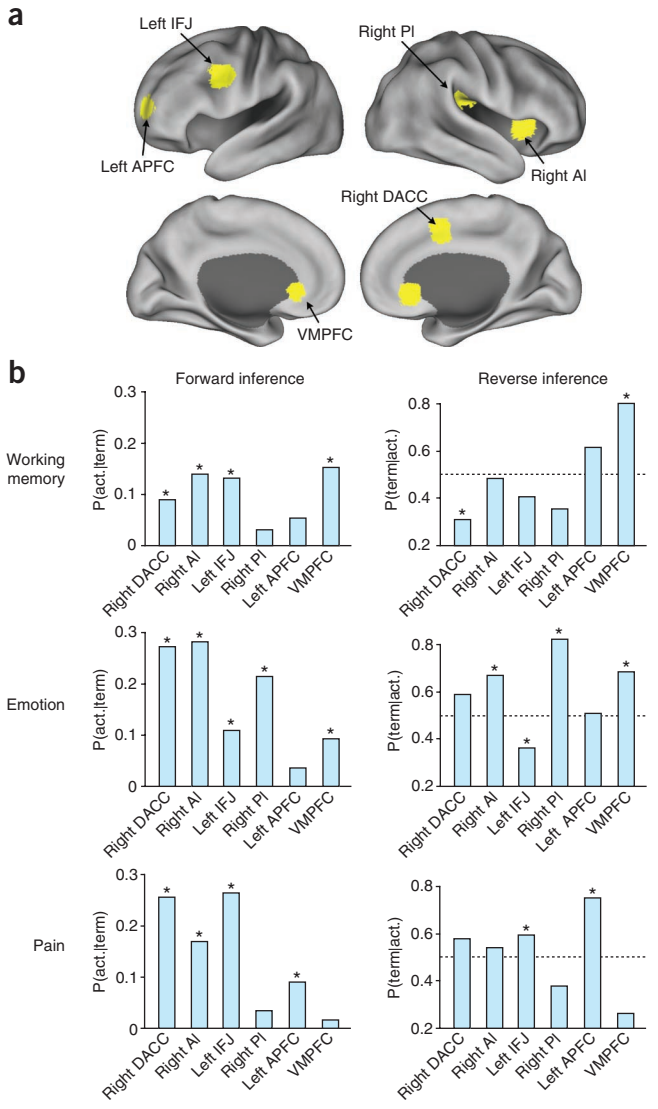


Figure 3 | Comparison of forward and reverse inference in regions of interest. (**a**) Labeled regions of interest shown on lateral and medial brain surfaces. (**b**) Comparison of forward inference (probability of activation given term P(act.|term)) and reverse inference (probability of term given activation P(term|act.)) for the domains of working memory, emotion and pain as marked. * denotes results at a false discovery rate threshold of 0.05; (whole-brain false discovery rate, (q) = 0.05). DACC, dorsal anterior cingulate cortex (stereotactic coordinates in Montreal Neurological Institute space: +2, +8, +50); AI, anterior insula (+36, +16, +2); IFJ, inferior frontal junction (−50, +8, +36); PI, posterior insula (+42, −24, +24); APFC, anterior prefrontal cortex (−28, +56, +8); VMPFC, ventromedial prefrontal cortex (0, +32, −4). Dashed lines indicate even odds of a term being used (P(term|act.) = 0.5).

in regions such as the fusiform face area[12] and visual word form area[13] (**Supplementary Fig. 3**). Third, we found that more conservative meta-analyses in which the lexical search space had been restricted to article titles yielded similar, but less sensitive, meta-analysis results (**Supplementary Fig. 4**).

Finally, we compared our results with those produced by previous manual approaches. Comparison of automated meta-analyses of three broad psychological terms ('working memory', 'emotion' and 'pain') with previously published meta- or mega-analytic maps[14–16] revealed marked qualitative (**Fig. 2**) and quantitative convergence (**Supplementary Fig. 5**) between approaches. To directly test the convergence of automated and manual approaches when applied to similar data, we manually validated 265 automatically extracted pain studies and performed a standard multilevel kernel density analysis[15] to compare experimental pain stimulation with baseline (66 valid studies). There was a notable overlap between automated and manual results (correlation across voxels, 0.84; **Supplementary Fig. 6**). These results showed that, at least for broad domains, an automated meta-analysis approach generated results that were comparable in sensitivity and scope to those produced with more effort in previous studies.

Quantitative reverse inference

The relatively comprehensive nature of the NeuroSynth database enabled us to address a long-standing inferential problem in the neuroimaging literature, namely how to quantitatively identify cognitive states from patterns of observed brain activity. This problem of 'reverse inference'[17] arises because most neuroimaging studies are designed to identify neural changes that result from known psychological manipulations and not to determine what cognitive state(s) a given pattern of activity implies[17] (**Fig. 1b**). For instance, fear consistently activates the human amygdala, but this does not imply that people in whom the amygdala is activated must be experiencing fear because other affective and nonaffective states have also been reported to activate the amygdala[4,18]. True reverse inference requires knowledge of which brain regions and networks are selectively, and not just consistently, associated with particular cognitive states[15,17].

Because the NeuroSynth database contains a broad set of term-to-activation mappings, our framework is well suited for drawing quantitative inferences about mind-brain relationships in both the forward and reverse directions. We could quantify both the probability that there would be activation in specific brain regions given the presence of a particular term (P(activation|term) or 'forward inference'), and the probability that a term would occur in an article given the presence of activation in a particular brain region (P(term|activation) or reverse inference). Comparison of these two analyses allowed us to assess the validity of many common inferences about the relationship between neural and cognitive states.

For illustration, we focused on the sample domains of working memory, emotion and pain, which are of substantial basic and clinical interest and have been extensively studied using fMRI (for additional examples, see **Supplementary Fig. 7**). These domains are excellent candidates for quantitative reverse inference, as they are thought to have confusable neural correlates, with common activation of regions such as the dorsal anterior cingulate cortex (DACC)[19] and anterior insula.

Our results showed differences between the forward and reverse inference maps in all three domains (**Fig. 2**). For working memory, the forward inference map revealed the most consistent associations in the dorsolateral prefrontal cortex, anterior insula and dorsal medial frontal cortex, replicating previous findings[15,20]. However, the reverse inference map instead implicated the anterior prefrontal cortex and posterior parietal cortex as the regions that were most selectively activated by working memory tasks.

We observed a similar pattern for pain and emotion. In both domains, frontal regions that have been broadly implicated in

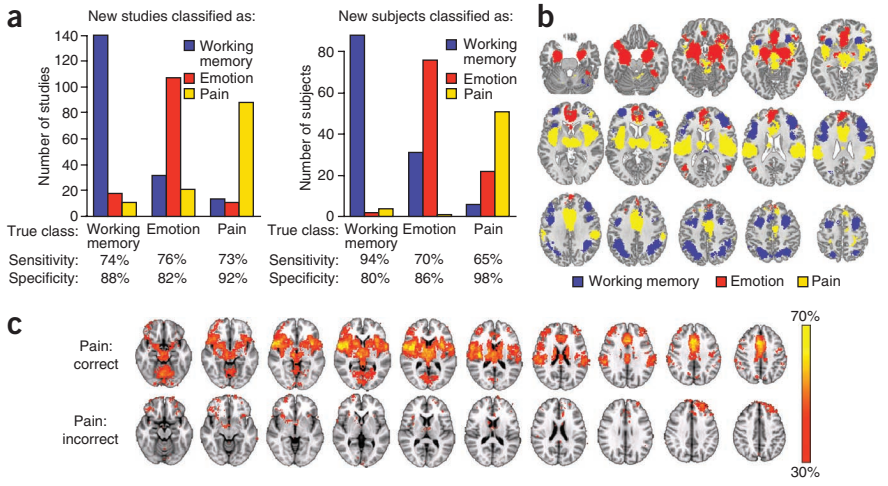


Figure 4 | Three-way classification of working memory, emotion and pain. (**a**) Naive Bayes classifier performance when cross-validated on studies in the database (left) or applied to individual subjects from studies not in the database (right). (**b**) Whole-brain maximum posterior probability map; each voxel is colored by the term with the highest associated probability. (**c**) Whole-brain maps showing the proportion of individual subjects in the three pain studies ($n = 79$ subjects total) who showed activation at each voxel ($P < 0.05$, uncorrected), averaged separately for subjects who were classified correctly ($n = 51$ subjects; top) or incorrectly ($n = 28$ subjects; bottom). Regions are color-coded according to the proportion of subjects in the sample who showed activation at each voxel.

goal-directed cognition[21–23] showed consistent activation in the forward analysis but were relatively nonselective in the reverse analysis (**Fig. 2**). For emotion, the reverse inference map revealed much more selective activation in the amygdala and ventromedial prefrontal cortex (**Fig. 3**). For pain, the regions of maximal pain-related activation in the insula and DACC shifted from anterior foci in the forward analysis to posterior ones in the reverse analysis (**Fig. 3**). This is consistent with studies of nonhuman primates that have implicated the dorsal posterior insula as a primary integration center for nociceptive afferents[24] and with studies of humans in which anterior aspects of the so-called 'pain matrix' responded nonselectively to multiple modalities[25].

Several frontal regions that showed consistent activation for emotion and pain in the forward analysis were associated with a decreased likelihood that a study involved emotion or pain in the reverse inference analysis (**Fig. 3**). This seeming paradox reflected the fact that even though lateral and medial frontal regions had been consistently activated in studies of emotion and pain, they had been activated even more frequently in studies that did not involve emotion or pain (**Supplementary Fig. 8**). Thus, the fact that these regions showed involvement in pain and emotion probably reflected their much more general role in cognition (for example, sustained attention or goal-directed processing[22,23]) rather than processes specific to pain or emotion.

These results showed that without the ability to distinguish consistency from selectivity, neuroimaging data can produce misleading inferences. For instance, neglecting the high base rate of DACC activity might lead researchers in the areas of cognitive control, pain and emotion to conclude that the DACC has a key role in each domain. Instead, because the DACC is activated consistently in all of these states, its activation may not be diagnostic of any one of them and conversely, might even predict their absence. The NeuroSynth framework can potentially address this problem by enabling researchers to conduct quantitative reverse inference on a large scale.

Open-ended classification of cognitive states

An emerging frontier in human neuroimaging is brain 'decoding': inferring a person's cognitive state from their observed brain activity. The problem of decoding is essentially a generalization of the univariate reverse inference problem addressed above: instead of predicting the likelihood of a particular cognitive state given activation at a single voxel, one can generate a corresponding prediction based on an entire pattern of brain activity. The NeuroSynth framework is well positioned for such an approach: whereas previous decoding approaches have focused on discriminating between narrow sets of cognitive states and have required extensive training on raw fMRI datasets (for example, refs. 26–28), the breadth of cognitive concepts represented in the NeuroSynth database affords relatively open-ended decoding, with little or no training on new datasets.

To assess the ability of our approach to decode and classify cognitive states, we trained a naive Bayes classifier[29] that could discriminate between flexible sets of cognitive states given new images as input (**Fig. 1c**). First, we tested the classifier's ability to classify studies in the NeuroSynth database that had been associated with different terms. In a tenfold cross-validated analysis, the classifier discriminated between studies of working memory, emotion and pain with high sensitivity and specificity (**Fig. 4a**), showing that each of these domains had a relatively distinct neural signature (**Fig. 4b**).

To assess the classifier's ability to decode cognitive states in individual human subjects, we applied the classifier to 281 single-subject activation maps derived from contrasts between: n-back working memory performance and rest (94 maps); negative and neutral emotional photographs (108 maps); and intense and mild thermal pain (79 maps). The classifier performed substantially above chance, identifying the originating study type with sensitivities of 94%, 70% and 65%, respectively (chance = 33%), and specificities of 80%, 86% and 98% (**Fig. 4a**). Moreover, there were systematic differences in activation patterns for correctly and incorrectly classified subjects. For example, incorrectly classified subjects in physical pain tasks (**Fig. 4c**) systematically activated the lateral orbitofrontal cortex and dorsomedial prefrontal cortex but not secondary somatosensory cortex or the posterior insula, suggesting that the discomfort owing to noxious heat in these subjects may have been qualitatively different (for example, emotionally generated versus physically generated pain). Thus, these findings demonstrate the viability of decoding cognitive states in new subjects without training and suggest new hypotheses for exploration.

Next, to generalize beyond working memory, emotion and pain, we selected 25 broad psychological terms that occurred at high frequency in the database (**Fig. 5**). We estimated classification accuracy in tenfold cross-validated two-alternative and multiclass analyses. The classifier performed substantially above

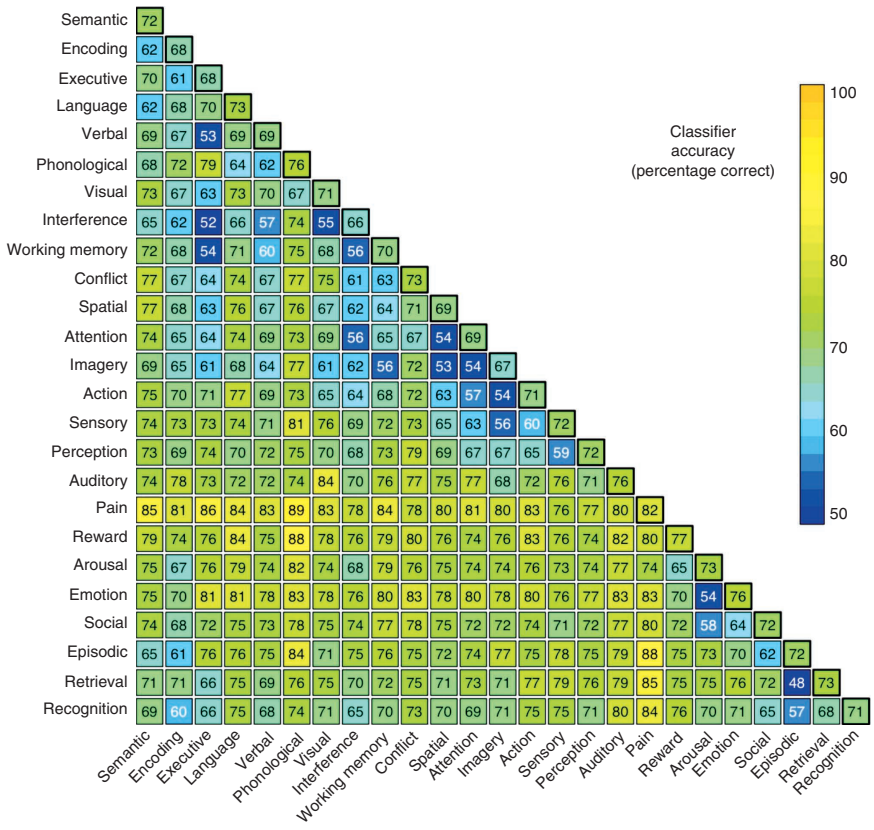


Figure 5 | Accuracy of the naive Bayes classifier when discriminating between all possible pairwise combinations of 25 key terms. Each cell represents a cross-validated binary classification between the intersecting row and column terms. Off-diagonal values reflect accuracy averaged across the two terms. Diagonal values reflect the mean classification accuracy for each term. Terms were ordered using the first two factors of a principal components analysis. All accuracy rates above 58% and 64% are statistically significant at $P < 0.05$ and $P < 0.001$, respectively.

chance in both two-alternative classification (mean pairwise accuracy of 72%; **Fig. 4**) and relatively open-ended multiclass classification on up to ten simultaneous terms (**Supplementary Fig. 9**). The results provided insights into the similarity structure of neural representation for different processes. For instance, pain was highly discriminable from other psychological concepts (all pairwise accuracies > 74%), which suggests that pain perception might be a distinctive state that is grouped neither with other sensory modalities nor with other affective concepts such as arousal and emotion. Conversely, conceptually related terms such as 'executive' and 'working memory' could not be distinguished at a rate different from chance, reflecting their closely overlapping usage in the literature.

DISCUSSION

Using the NeuroSynth framework, first we conducted large-scale automated neuroimaging meta-analyses of broad psychological concepts that are lexically well represented in literature. A key benefit of NeuroSynth is the ability to quantitatively distinguish forward inference from reverse inference, which should allow researchers to assess the specificity of mappings between neural and cognitive function, a long-standing goal of cognitive neuroscience research. Although considerable work remains to be done to improve the specificity and accuracy of the tools developed here, we expect quantitative reverse inference to be increasingly important in future meta-analytic studies.

Second, we decoded broad psychological states in a relatively open-ended way in individual subjects; this was, to our knowledge, the first application of a domain-general classifier that can distinguish a broad range of cognitive states based solely on prior literature. The ability to decode brain activity without previous training data or knowledge of the 'ground truth' for an individual is particularly promising. Our results raise the prospect that legitimate 'mind reading' of more nuanced cognitive and affective states might eventually become feasible with additional technical advances. However, the present NeuroSynth implementation provides no basis for such inferences, as it distinguishes only between relatively broad psychological categories.

Third, we designed our platform to support immediate use in a broad range of neuroimaging applications. To name just a few potential applications, researchers could use these tools and results to define region-of-interest masks or Bayesian priors in hypothesis-driven analyses; to conduct quantitative comparisons between meta-analysis maps of different terms of interest; to use the automatically extracted coordinate database as a starting point for more refined manual meta-analyses; to draw more rigorous reverse inferences when interpreting results by referring to empirically established mappings between specific regions and cognitive functions; and to extract the terms that are most frequently associated with an active region or distributed pattern of activity, thereby contextualizing new research findings on the basis of published data.

Of course, the NeuroSynth framework is not a panacea for the many challenges that face cognitive neuroscientists, and several limitations remain to be addressed. We focus on two in particular here. First, the present reliance on a purely lexical coding approach, albeit effective, is suboptimal in that it relies on traditional psychological terms that do not carve the underlying neural substrates at their natural joints, do not capitalize on redundancy across terms (for example, 'pain', 'nociception' and 'noxious' overlap closely but are modeled separately) and do not allow closely related constructs to be easily distinguished (for example, physical versus emotional pain). Future efforts could overcome these limitations by using controlled vocabularies or ontologies for query expansion, developing extensions for conducting multiterm analyses and extracting topic-based representations of article text (**Supplementary Note**).

Second, although our automated tools accurately extract coordinates from articles, they cannot extract information about fine-grained cognitive states (for example, different negative emotions). Thus, the NeuroSynth framework is currently useful primarily for large-scale analyses involving broad domains and

should be viewed as a complement to, and not as a substitute for, manual meta-analysis approaches. We are currently working to develop improved algorithms for automatic coding of experimental contrasts, which should substantially improve the specificity of the resulting analyses. In parallel, we envision a 'crowd-sourced' collaborative model in which multiple groups participate in the validation of automatically extracted data, thereby combining the best elements of both automated and manual approaches. Such efforts should further increase the specificity and predictive accuracy of the decoding model, and we hope that they will lead to the development of many other applications that we have not anticipated here.

To encourage application and development of a synthesis-oriented approach, we have publicly released most of the tools and data used in the present study through a web interface (http://neurosynth.org/). We hope that cognitive neuroscientists will use, and contribute to, this new resource, with the goal of developing new techniques for interpreting and synthesizing the wealth of data generated by modern neuroimaging techniques.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank T. Braver (Washington University), J. Gray (Yale University) and K. Ochsner (Columbia University) for data; E. Reid for help with validation analyses; members of the Wager lab for manually coding the pain database; members of the Neuroimaging and Data Access Group (http://nidag.org/), and particularly R. Mar, for suggestions; and R. Bilder, R. Raizada and J. Andrews-Hanna for comments on a draft of this paper. This work was supported by awards from US National Institute of Nursing Research (F32NR012081 to T.Y.), National Institute of Mental Health (R01MH082795 to R.A.P. and R01MH076136 to T.D.W.), US National Institutes of Health (R01MH60974 to D.C.V.E.) and National Institute on Drug Abuse (R01DA027794 and 1RC1DA028608 to T.D.W.).

AUTHOR CONTRIBUTIONS

T.Y. conceived the project and carried out most of the software implementation, data analysis and writing. R.A.P. provided data and performed analyses. T.E.N. provided statistical advice, reviewed all statistical procedures and contributed to the implementation of the naive Bayes classifier. D.C.V.E. provided data, contributed to automated data extraction and coordinated data validation. T.D.W. conceived the classification analyses, wrote part of the software, provided data and suggested and performed analyses. All authors contributed to writing and editing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Derrfuss, J. & Mar, R.A. Lost in localization: the need for a universal coordinate database. *Neuroimage* **48**, 1–7 (2009).
2. Yarkoni, T. Big correlations in little studies: inflated fMRI correlations reflect low statistical power—commentary on Vul *et al*. (2009). *Perspect. Psychol. Sci.* **4**, 294–298 (2009).
3. Wager, T.D., Lindquist, M. & Kaplan, L. Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* **2**, 150–158 (2007).
4. Yarkoni, T., Poldrack, R.A., Van Essen, D.C. & Wager, T.D. Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends Cogn. Sci.* **14**, 489–496 (2010).
5. Van Horn, J.D., Grafton, S.T., Rockmore, D. & Gazzaniga, M.S. Sharing neuroimaging studies of human cognition. *Nat. Neurosci.* **7**, 473–481 (2004).
6. Fox, P.T., Parsons, L.M. & Lancaster, J.L. Beyond the single study: function/location metanalysis in cognitive neuroimaging. *Curr. Opin. Neurobiol.* **8**, 178–187 (1998).
7. Van Essen, D.C. Lost in localization—but found with foci?! *Neuroimage* **48**, 14–17 (2009).
8. Laird, A.R. *et al.* ALE meta-analysis workflows via the BrainMap database: progress towards a probabilistic functional brain atlas. *Front Neuroinformatics* **3**, 23 (2009).
9. Dickson, J., Drury, H.A. & Van Essen, D.C. "The surface management system" (SuMS) database: a surface-based database to aid cortical surface reconstruction, visualization and analysis. *Phil. Trans. R. Soc. Lond. B* **356**, 1277–1292 (2001).
10. Lancaster, J.L. *et al.* Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Hum. Brain Mapp.* **28**, 1194–1205 (2007).
11. Nielsen, F.A., Hansen, L.K. & Balslev, D. Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics* **2**, 369–380 (2004).
12. Kanwisher, N., McDermott, J. & Chun, M.M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
13. McCandliss, B.D., Cohen, L. & Dehaene, S. The visual word form area: expertise for reading in the fusiform gyrus. *Trends Cogn. Sci.* **7**, 293–299 (2003).
14. Atlas, L.Y., Bolger, N., Lindquist, M.A. & Wager, T.D. Brain mediators of predictive cue effects on perceived pain. *J. Neurosci.* **30**, 12964–12977 (2010).
15. Wager, T.D., Lindquist, M.A., Nichols, T.E., Kober, H. & Van Snellenberg, J.X. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage* **45**, 210–221 (2009).
16. Kober, H. *et al.* Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage* **42**, 998–1031 (2008).
17. Poldrack, R.A. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10**, 59–63 (2006).
18. Zald, D.H. The human amygdala and the emotional evaluation of sensory stimuli. *Brain Res. Brain Res. Rev.* **41**, 88–123 (2003).
19. Shackman, A.J. *et al.* The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nat. Rev. Neurosci.* **12**, 154–167 (2011).
20. Owen, A.M., McMillan, K.M., Laird, A.R. & Bullmore, E. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* **25**, 46–59 (2005).
21. Dosenbach, N.U. *et al.* A core system for the implementation of task sets. *Neuron* **50**, 799–812 (2006).
22. Duncan, J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci.* **14**, 172–179 (2010).
23. Yarkoni, T., Barch, D.M., Gray, J.R., Conturo, T.E. & Braver, T.S. BOLD correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI analysis. *PLoS ONE* **4**, e4257 (2009).
24. Craig, A.D. How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* **3**, 655–666 (2002).
25. Legrain, V., Iannetti, G.D., Plaghki, L. & Mouraux, A. The pain matrix reloaded: a salience detection system for the body. *Prog. Neurobiol.* **93**, 111–124 (2011).
26. 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).
27. Mitchell, T.M. *et al.* Predicting human brain activity associated with the meanings of nouns. *Science* **320**, 1191–1195 (2008).
28. Poldrack, R.A., Halchenko, Y.O. & Hanson, S.J. Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychol. Sci.* **20**, 1364–1372 (2009).
29. Lewis, D. Naive (Bayes) at forty: The independence assumption in information retrieval. *Mach. Learn.* **ECML-98**, 4–15 (1998).
30. Lang, P.J., Bradley, M.M. & Cuthbert, B.N. *International Affective Picture System (IAPS): Instruction Manual and Affective Ratings* (Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, USA, 1999).

ONLINE METHODS

Automated coordinate extraction. To automatically extract stereotactic coordinate information from published neuroimaging articles, we developed a software library written in the Ruby programming language. We released the NeuroSynth automated coordinate extraction (ACE) tools under an open-source license and encourage other researchers to contribute to the codebase (http://github.com/neurosynth/). Because the code is freely available for inspection and use, we provide only a functional, non-technical overview of the tools here.

In brief, ACE consists of a parsing engine that extracts coordinate information from published articles by making educated guesses about the contents of the columns reported in tables in neuroimaging articles (at present, ACE does not attempt to extract coordinates that are reported in the main text of an article or in supplementary materials). For each full-text HTML article provided as input, ACE scans all tables for rows that contain coordinate-like data. Rows or tables that do not contain values that correspond to a flexible template used to detect coordinates are ignored. Moreover, all extracted coordinates are subjected to basic validation to ensure that they reflect plausible locations in stereotactic space (for example, all coordinates with absolute values > 100 in any plane are discarded).

Although neuroimaging coordinates are reported in a variety of stereotactic spaces in the neuroimaging literature[31,32], for technical reasons, the results we reported in the main text ignore such differences and collapse across different spaces. Moreover, the parser did not distinguish activations from deactivations, and aggregates across all reported contrasts in each article, that is, it makes no attempt to code different tables within an article, or different contrasts within a table, separately. As extensive validation analyses showed (**Supplementary Note**), these factors appear to exert only a modest influence on results, and in some cases can be automatically accounted for; however, for present purposes we simply note that the net effect of these limitations should be to reduce fidelity rather than to introduce systematic bias.

As well as extracting coordinates, ACE parses the body of each article and generates a list of all words that appear at least once anywhere in the text, along with a corresponding frequency count for each word. All data are then stored in a relational (MySQL) database that maintains associations between words, articles and activation foci, allowing flexible and powerful structured retrieval of information.

Database. The foci used to generate the results of the present study were extracted from 17 source journals: *Biological Psychiatry, Brain, Brain and Cognition, Brain and Language, Brain Research, Cerebral Cortex, Cognitive Brain Research, Cortex, European Journal of Neuroscience, Human Brain Mapping, Journal of Neurophysiology, Journal of Neuroscience, NeuroImage, NeuroLetters, Neuron, Neuropsychologia* and *Pain*. We deliberately focused on journals that contained a high incidence of functional neuroimaging studies; thus, some important general neuroscience or science journals (for example, *Science, Nature* and *Nature Neuroscience*) were not included. The range of years represented varied by journal, with the earliest studies dating to 2000 and the latest to early 2010. The database contains 3,489 articles and 100,953 foci, and is, to our knowledge, the largest extant database of neuroimaging foci, though it still captures only a minority of the published literature[1].

The database has considerable potential for additional growth. Because neuroimaging studies appear in dozens of journals (besides those analyzed here), which typically require the use of publisher-specific or journal-specific filters to correctly obtain coordinates from tables, the database will continue to grow as new filters are added. An additional limitation is that many journals have not yet made full-text HTML versions of older articles available online; as such efforts proceed, the database will grow correspondingly. Finally, because authors report coordinate information in a variety of formats, false negatives can occur (that is, ACE might not extract real coordinate information). However, these limitations do not bias the present analyses in any particular way, and suggest that if anything, one can expect the sensitivity and specificity of the reported results to improve as the database grows and additional quality assurance procedures are implemented.

Statistical inference and effect size maps. In keeping with previous meta-analyses that used multilevel kernel density analysis (MKDA[15]), we represented reported activations from each study by constructed a binary image mask, with a value of 1 (reported) assigned to each voxel in the brain if it was within 10 mm of a focus reported in that article and 0 (not reported) if it was not within 10 mm of a reported focus[15]. These maps used 2 mm × 2 mm × 2 mm voxels, with $n_V = 231{,}202$ voxels in the brain mask. We denote the activation map for study i as $A_i = (A_{ij})$, a length-n_V binary vector.

A frequency cut-off of 0.001 was used to eliminate studies that only used a term incidentally (that is, to be considered about pain, a study had to use the term 'pain' at a rate of at least one in every 1,000 words). Subsequent testing revealed that the results were largely insensitive to the exact cut-off used (including no cut-off at all), except that very high cut-offs (for example, 0.01 or higher) tended to leave too few studies for reliable estimation of most terms. (As the database grows, even very conservative thresholds that leave no ambiguity at all about the topic of a study should become viable.) The total number of terms collected was $n_T = 10{,}000$ (although the majority of these were nonpsychological in meaning—for example, 'activation' and 'normal'). We write the term indicator for study i as $T_i = (T_{ik})$, a length-n_T binary vector marking each term 'present' (frequency above the cut-off) or 'absent' (frequency below the cut off).

For each term of interest in the database (for example, 'pain', 'amygdala') we generated whole-brain meta-analysis maps that showed the strength of statistical association between the term and reported activation at each voxel. For each voxel j and term k, every study can be cross-classified by activation (present or absent) and term (present or absent), producing a 2 × 2 contingency table of counts.

Statistical inference maps were generated using a χ^2 test of independence, with a significant result implying the presence of a dependency between term and activation (that is, a change in activation status would make the occurrence of the term more or less likely). This approach departs from the MKDA approach used in previous meta-analyses[3,15,16] in its reliance on a parametric statistical test in place of permutation-based family wise error rate (FWE) correction; however,

permutation-based testing was not computationally feasible given the scale of the present meta-analyses (multiple maps for each of several thousand terms).

To stabilize results and to ensure that all cells had sufficient observations for the parametric χ^2 test, we excluded all voxels that were active in fewer than 3% of studies (**Supplementary Fig. 10**). The resulting *P*-value map was false discovery rate (FDR)-corrected for multiple comparisons using a whole-brain FDR threshold of 0.05, identifying voxels where there was significant evidence that term frequency varied with activation frequency. Intuitively, one can think of regions that survive correction as those that show differential activation for studies that include a term versus studies that do not include that term.

We also computed maps of the posterior probability that term k was used in a study given activation at voxel j

$$P(T_k = 1 \mid A_j = 1) = P(A_j = 1 \mid T_k = 1)P(T_k = 1)/P(A_j)$$

(generically referred to as P(Term|Activation) in the text). We use the 'smoothed' estimates for the likelihood of activation given the term

$$p(A_j = 1 \mid T_k = 1) = \left(\sum_i A_{ij}T_{ik} + mp\right) \Big/ \left(\sum_i T_{ik} + m\right)$$

where p(·) reflects an estimated versus true probability, m is a virtual equivalent sample size and p is a prior probability. The parameters m and p are set equal to 2 and 0.5, respectively; this smoothing was equivalent to adding two virtual studies that have term k present, one having an activation one, one without. This regularization prevents rare activations or terms from degrading accuracy[33]. For $P(T_k = 1)$ we impose a uniform prior for all terms, $P(T_k = 1) = P(T_k = 0) = 0.5$. We use this uniform prior because terms differed widely in frequency of usage, leading to very different posterior probabilities for different terms. This is equivalent to making an assumption that the usage and nonusage of the term would be equally likely in the absence of any knowledge about brain activation. Note that this is a conservative approach; using uniform priors will tend to reduce classifier accuracy relative to using empirically estimated priors (that is, allowing base rate differences to play a role in classification) because it increases the accuracy of rare terms at the expense of common ones. Nonetheless, we used uniform priors because they place all terms on a level footing and provide more interpretable results.

The estimate of $P(A_j = 1)$ reflects the regularization and the prior on term frequency:

$$p(A_j = 1) = p(A_j = 1 \mid T_k = 1)P(T_k = 1) + p(A_j = 1 \mid T_k = 0)P(T_k = 0)$$

where

$$p(A_j = 1 \mid T_k = 0) = \left(\sum_i A_{ij}(1 - T_{ik}) + mp\right) \Big/ \left(\sum_i (1 - T_{ik}) + m\right)$$

To ensure that only statistically robust associations were considered, all posterior probability maps were masked with the FDR-corrected *P*-value maps. For visualization purposes, thresholded maps were mapped to the PALS-B12 surface atlas[34] in SPM5 stereotaxic space. Average fiducial mapping values are presented. Datasets associated with **Figure 2** and **Supplementary Figure 3** are available in the SumsDB database (http://sumsdb.wustl.edu/sums/directory.do?id=8285126).

Naive Bayes classifier. We used a naive Bayes classifier[29] to predict the occurrence of specific terms using whole-brain patterns of activation. In classifier terminology, we have n_S instances of feature-label pairs (A_i, T_i). The use of a naive Bayes classifier allows us to neglect the spatial dependence in the activation maps A_i. As simultaneous classification for the presence or absence of all n_T terms is impractical owing to the larger number (2^{n_T}) of possible labels, for this work, we only considered mutually exclusive term labels, ranging from binary classification of two terms (for example, pain versus working memory) to multiclass classification of ten terms (for example, pain, working memory, language, conflict and so on).

For this setting we revised notation slightly from the previous section describing calculation of the posterior probability maps, letting scalar T_i take values 1, …, n_T^* for the subset of n_T^* terms under consideration. For a new study with activation map A, the probability of term t is

$$P(T = t \mid A) = P(A \mid T = t)P(T = t) \,/\, P(A)$$

$P(A)$ is computed as above for the studies under consideration, and by independence,

$$P(A \mid T = t) = \prod_j P(A_j \mid T = t),$$

and we use a regularized estimate for voxel j,

$$p(A_j = 1 \mid T = t) = \left(\sum_i A_{ij} I(T_i = t) + mp\right) \Big/ \left(\sum_i I(T_i = t) + m\right),$$

$$p(A_j = 0 \mid T = t) = 1 - P(A_j = 1 \mid T = t)$$

where $I(\cdot)$ is the indicator function (1 if the operand is true, 0 otherwise).

Although the assumption of conditional independence is usually violated in practice, leading to biased posterior probabilities, this generally does not affect classification performance because classification depends on the rank-ordered posterior probabilities of all classes rather than their absolute values. In complex real-world classification settings, naive Bayes classifiers often substantially outperform more sophisticated and computationally expensive techniques[35].

In the context of the present large-scale analyses, the naive Bayes classifier has several advantages over other widely used classification techniques (for example, support vector machines[36,37]). First, it requires substantially less training data than many other techniques because only the cross-classified cell counts are needed. Second, it is computationally efficient, and can scale up to extremely large sets of features (for example, hundreds of thousands of individual voxels) or possible outcomes without difficulty. Third, it produces easily interpretable results: the naive Bayes classifier's assumption of conditional independence ensures that the strength of each feature's contribution to the overall classification simply reflects the posterior probability of class membership conditioned on that feature (that is, $P(T = t|A_i)$).

Cross-validated classification of study-level maps. For cross-validated classification of studies included in the NeuroSynth database (**Figs. 3** and **4**), we used the NBC to identify the most probable term from among a specified set of alternatives (for example, 'pain', 'emotion' and 'working memory') for each map. We used fourfold

doi:10.1038/nmeth.1635

cross-validation to ensure unbiased accuracy estimates. Because the database was known to contain errors, we took several steps to obtain more accurate classification estimates. First, to improve the signal-to-noise ratio, we trained and tested the classifier only on the subset of studies with at least 5,000 'active' voxels (that is, studies that satisfied $\Sigma_j A_{ij} \geq 5{,}000$) occurring when there were more than about four reported foci; $n_S = 2{,}107$ studies satisfied this criterion. This step ensured that studies with few reported activations (a potential marker of problems extracting coordinates) did not influence the classifier. Second, we only considered voxels that were activated in at least 3% of studies ($\Sigma_i A_{ij}/n_S \geq 0.03$), which ensured that noisy features would not exert undue influence on classification. Third, any studies in which more than one target term was used were excluded to ensure that there was always a correct answer (for example, if a study used both 'pain' and 'emotion' at a frequency of 0.001 or greater, it was excluded from classification). No further feature selection was used (all remaining voxels were included as features).

We calculated classifier accuracy by averaging across classes (terms) rather than studies (for example, if the classifier correctly classified 100% of 300 working memory studies, but 0% of 100 pain studies, we would report a value of 50%, reflecting the mean of 0% and 100%, rather than the study-wise mean of 75%, which allows for inflation due to differing numbers of studies). Using this accuracy metric, called balanced loss in the machine learning literature, eliminated the possibility of the classifier capitalizing on differences in term base rates and ensured that chance accuracy was always 50%. Note that this is the appropriate comparison for a naive Bayes classifier when uniform prior probabilities are stipulated because the classifier should not be able to capitalize on base rate differences even if they exist (as it has no knowledge of base rates beyond the specified prior).

For binary classification ($n_T^* = 2$), we selected 25 terms that occurred at high frequency in our database (>1 in 1,000 words in at least 100 studies; **Fig. 5** and **Supplementary Fig. 3**) and ran the classifier on all possible pairs. For each pair, the set of studies used included all those with exactly one term present ($n_s = 23$–794; mean = 178.2, median = 141). Statistical significance for each pairwise classification was assessed using Monte Carlo simulation. Across all comparisons, the widest 95% confidence interval was 0.4–0.6, and the majority of observed classification accuracies (283/300) were statistically significant ($P < 0.05$; 257/300 were significant at $P < 0.001$).

For multiclass analyses involving $n_T^* > 2$ terms (**Supplementary Fig. 5**), an exhaustive analysis of all possible combinations was not viable owing to combinatorial explosion and the increased processing time required. We therefore selected 100 random subsets of n_T^* terms from the larger set of 25, repeating the process for values of n_T^* between 3 and 10. All procedures were otherwise identical to those used for binary classification.

Classification of single-subject data. To classify single-subject data, we used data from several previous studies, including a large study of n-back working memory[38,39], five studies of emotional experience and reappraisal[40–44] and three studies of pain[14,45]. Methodological details for these studies can be found in the corresponding references. For working memory, we used single-subject contrasts that compared n-back working memory blocks to a fixation baseline. For emotion studies, we used single-subject contrast maps that compared negative emotional pictures (from the IAPS set) to neutral pictures. For pain studies, we compared high and low thermal pain conditions.

Because the NBC was trained on binary maps (active vs. inactive) and single-subject *P*-maps varied continuously, all single-subject maps were binarized before classification. We used an arbitrary threshold of $P < 0.05$ to identify 'active' voxels. Because the maps on which the classifier was trained did not distinguish activations from deactivations, only positively activated voxels (n-back > fixation, negative emotion > neutral emotion or high pain > low pain) were considered active. Negatively activated voxels and nonsignificantly activated voxels were all considered inactive. To ensure that all single-subject maps had sufficient features for classification, we imposed a minimum cut-off of 1% of voxels—that is, for maps with fewer than 1% of voxels activated at $P < 0.05$, we used the top 1% of voxels, irrespective of threshold. Once the maps were binarized, all classification procedures were identical to those used for cross-validated classification of study-level maps.

The studies and contrasts included in the single-subject classification analysis were selected using objective criteria rather than on the (circular) basis of optimizing performance. The results we report include all studies that were subjected to the classifier (that is, we did not selectively include only studies that produced better results), despite the fact that there was marked heterogeneity within studies. For instance, one of the pain studies produced substantially better results ($n = 41$; sensitivity = 80%) than the other two (total $n = 34$; sensitivity = 47%), probably reflecting the fact that the former study contained many more trials per subject, resulting in more reliable single-subject estimates. Thus, the accuracy levels we report are arguably conservative, as they do not account for the potentially lower quality of some of our single-subject data.

Similarly, the contrasts we used for the three sets of studies were selected a priori on the basis of their perceived construct validity, and not on the basis of observed classification success. In fact, post-hoc analyses showed that alternative contrasts would have produced better results in some cases. Notably, for the emotion studies, using single-subject maps contrasting passive observation of negative IAPS pictures with active reappraisal of negative pictures would have improved classifier sensitivity somewhat (from 70% to 75%). Nonetheless, we opted to report the less favorable results in order to provide a reasonable estimate of single-subject classifier accuracy under realistic conditions, uncontaminated by selection bias. Future efforts to optimize the classifier for single-subject prediction (for example, by developing ways to avoid binarizing continuous maps, improving the quality of the automatically extracted data through manual verification and so on) would presumably lead to substantially better performance.

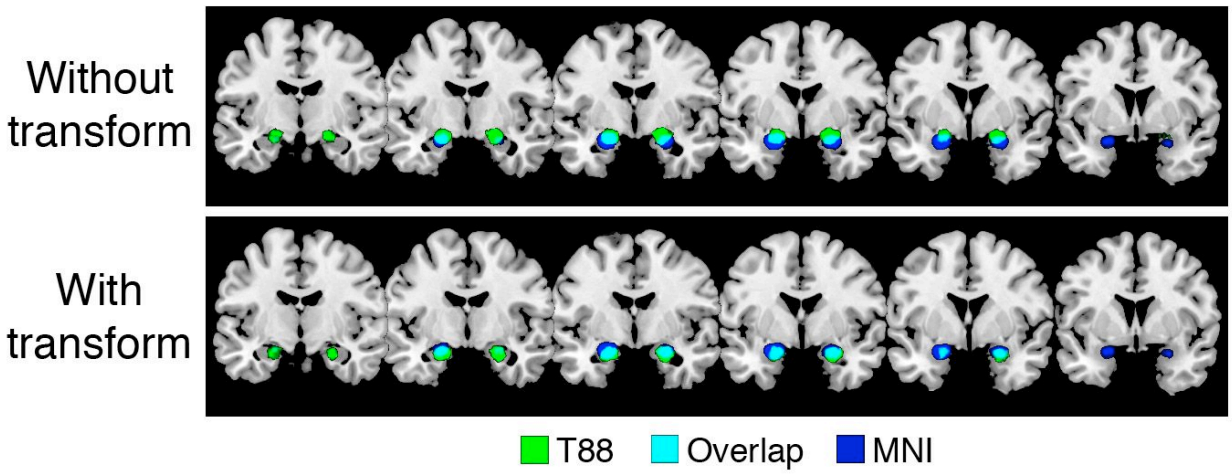
31. Van Essen, D.C. & Dierker, D.L. Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron* **56**, 209–225 (2007).
32. Laird, A.R. *et al.* Comparison of the disparity between Talairach and MNI coordinates in functional neuroimaging data: validation of the Lancaster transform. *Neuroimage* **51**, 677–683 (2010).
33. Nigam, K., McCallum, A.K., Thrun, S. & Mitchell, T. Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* **39**, 103–134 (2000).
34. Van Essen, D.C.A. Population-average, landmark- and surface-based (PALS) atlas of human cerebral cortex. *Neuroimage* **28**, 635–662 (2005).

35. Langley, P., Iba, W. & Thompson, K. An analysis of Bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence* 223–228 (AAAI Press, Menlo Park, California, USA, 1992).
36. Mitchell, T.M. *et al.* Learning to decode cognitive states from brain images. *Mach. Learn.* **57**, 145–175 (2004).
37. 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).
38. DeYoung, C.G., Shamosh, N.A., Green, A.E., Braver, T.S. & Gray, J.R. Intellect as distinct from openness: differences revealed by fMRI of working memory. *J. Pers. Soc. Psychol.* **97**, 883–892 (2009).
39. Shamosh, N.A. *et al.* Individual differences in delay discounting. *Psychol. Sci.* **19**, 904–911 (2008).
40. McRae, K. *et al.* The neural bases of distraction and reappraisal. *J. Cogn. Neurosci.* **22**, 248–262 (2010).
41. Ochsner, K.N. *et al.* For better or for worse: neural systems supporting the cognitive down- and up-regulation of negative emotion. *Neuroimage* **23**, 483–499 (2004).
42. Ochsner, K.N., Bunge, S.A., Gross, J.J. & Gabrieli, J.D. Rethinking feelings: an FMRI study of the cognitive regulation of emotion. *J. Cogn. Neurosci.* **14**, 1215–1229 (2002).
43. Wager, T.D., Davidson, M.L., Hughes, B.L., Lindquist, M.A. & Ochsner, K.N. Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron* **59**, 1037–1050 (2008).
44. McRae, K., Ochsner, K.N., Mauss, I.B., Gabrieli, J.J.D. & Gross, J.J. Gender differences in emotion regulation: an fMRI study of cognitive reappraisal. *Group Process. Intergroup Relat.* **11**, 143–162 (2008).
45. Kross, E., Berman, M.G., Mischel, W., Smith, E.E. & Wager, T.D. Social rejection shares somatosensory representations with physical pain. *Proc. Natl. Acad. Sci. USA* **108**, 6270–6275 (2011).

doi:10.1038/nmeth.1635

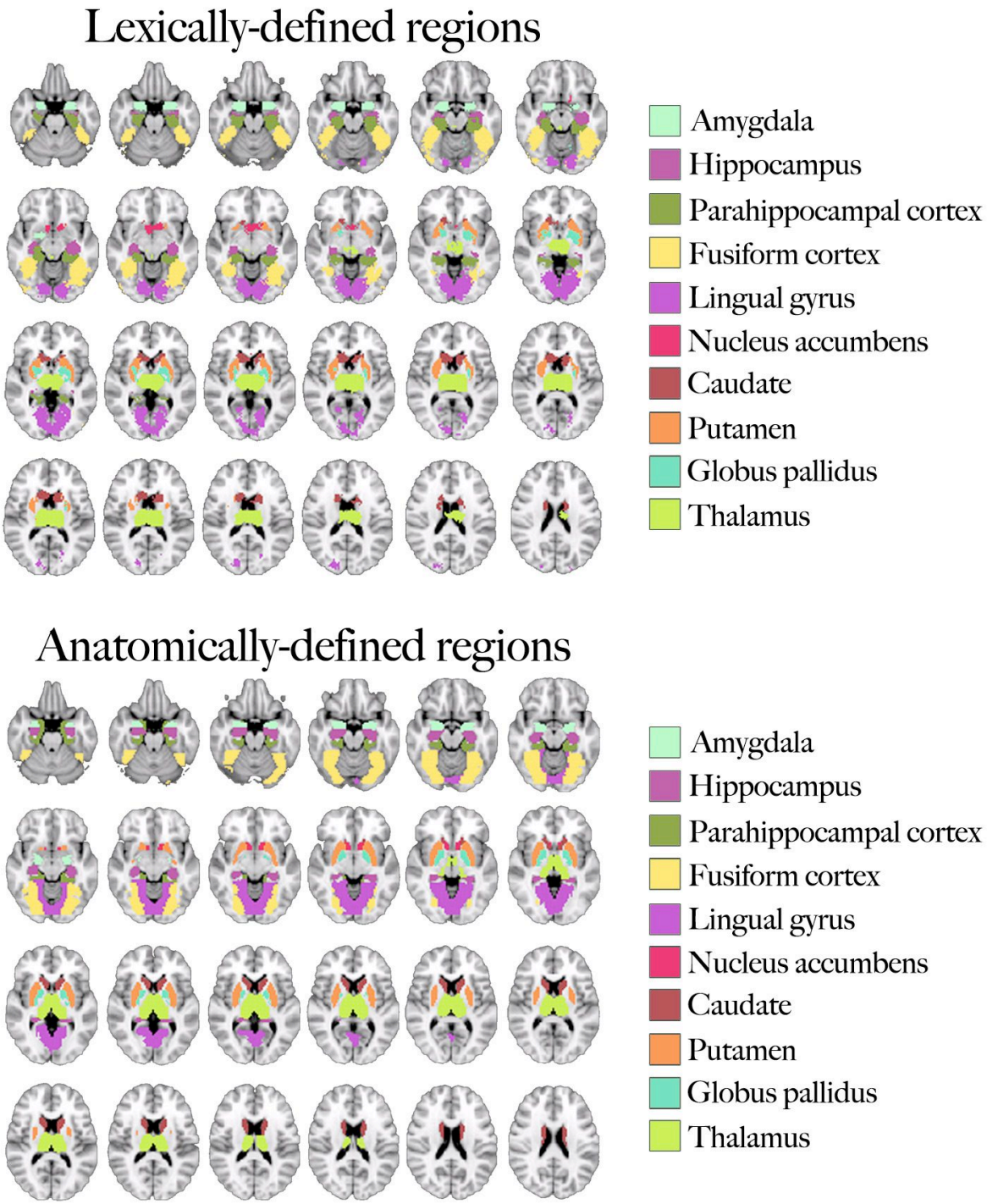
Supplementary Figures

Supplementary Figure 1



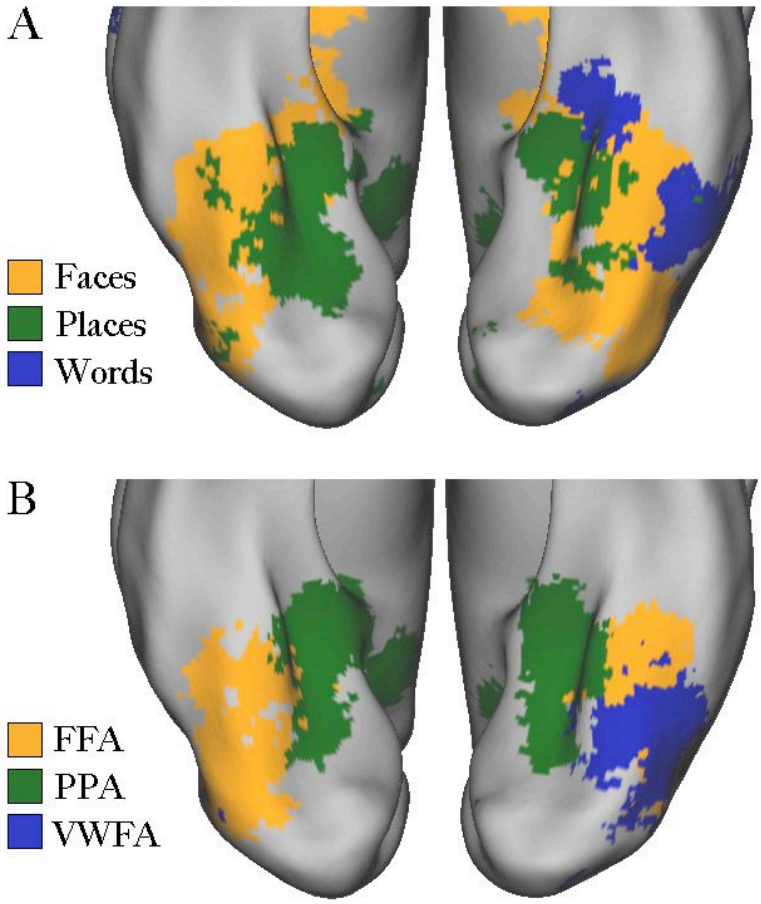
Effects of automated stereotactic space detection and correction on meta-analyses for the term 'amygdala'. Blue: regions maximally active in an automated meta-analysis of all studies that reported coordinates in MNI space. Green: regions maximally active when analyzing all studies reporting coordinates in T88 (i.e., Talairach-Tournoux, 1988[1]). Top row: overlap between MNI and T88 results space prior to application of any transformation. Notice the relatively poor alignment of the T88-based results with both the anatomical underlay and the MNI-based functional results (T88/MNI overlap colored cyan). Bottom row: following automated application of the Lancaster et al transform[2], the T88-based results are substantially better aligned along the dorsal/ventral axis, though differences remain along the rostral/caudal axis. The pearson correlation between MNI and T88 maps across all voxels improved from 0.73 pre-transform to 0.81 post-transform.

Supplementary Figure 2



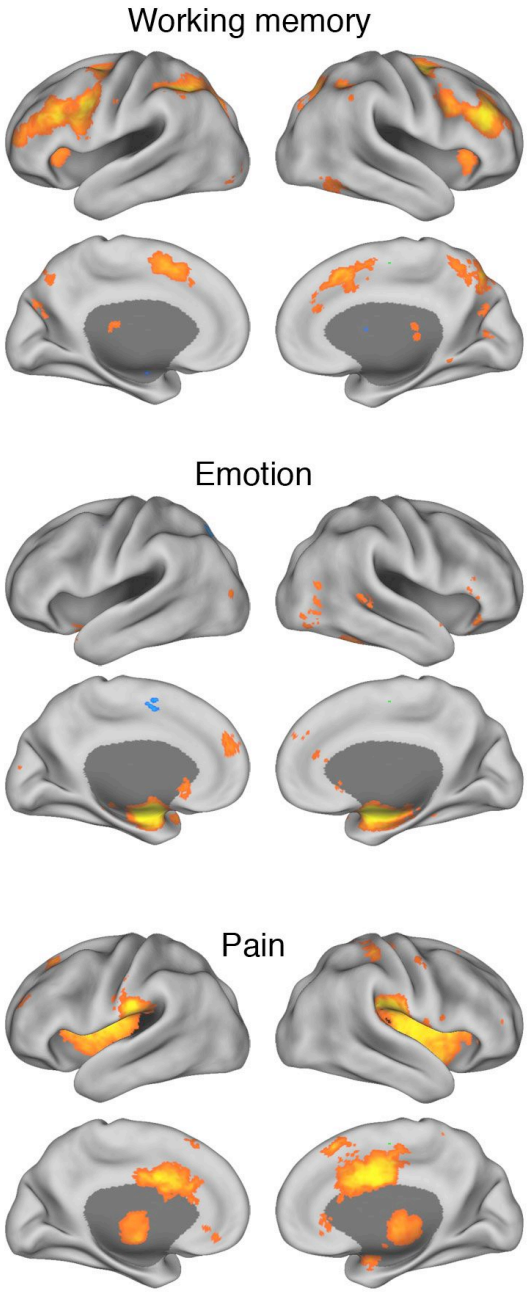
Comparison of selected automatically-defined regions with anatomical regions delineated in the Harvard-Oxford Atlas. Because the posterior probability maps could be thresholded arbitrarily, we restricted each automatically-generated ROI to the same number of voxels present in the ROI defined in the Harvard-Oxford atlas.

Supplementary Figure 3



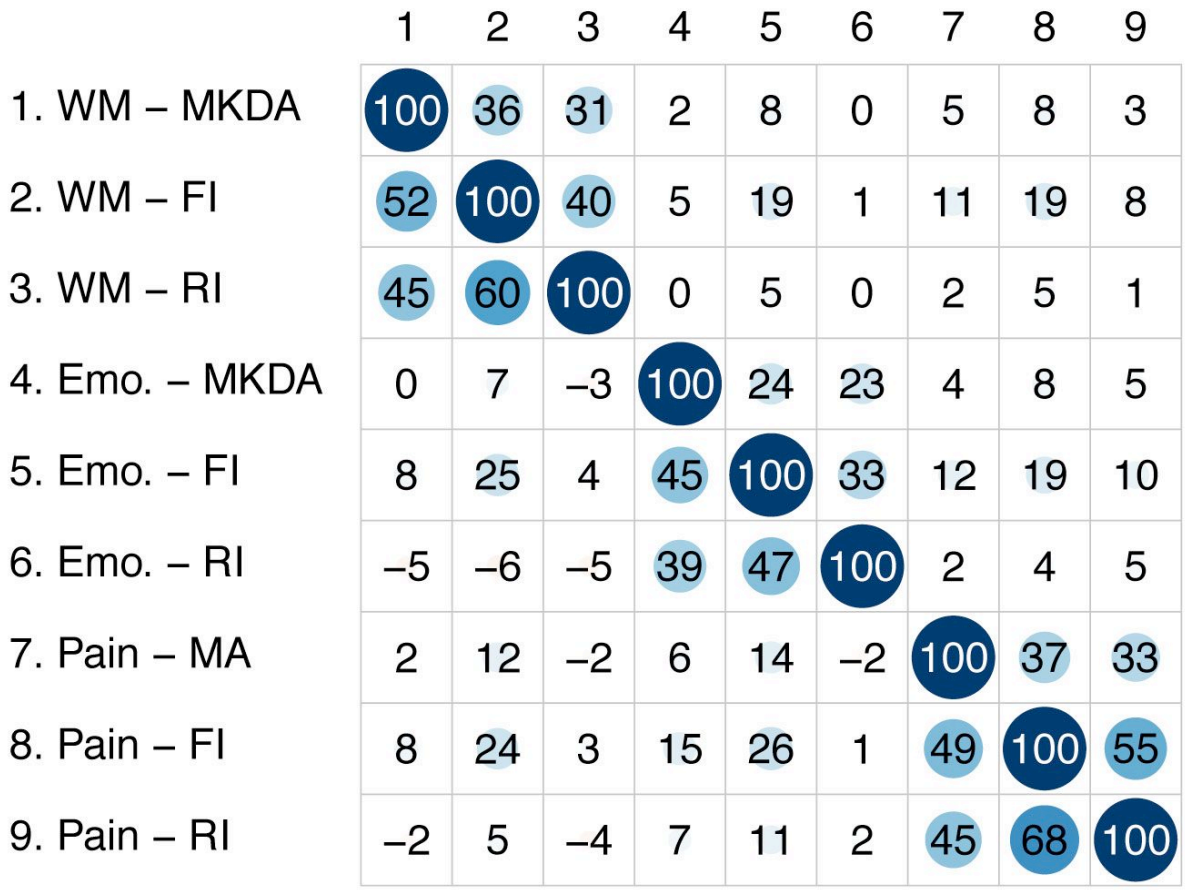
Category-specific activations in inferotemporal cortex identified with *NeuroSynth*. (A) Classification based on functional terms ('faces', 'places', and 'words'). (B) Classification based on putative functionally specialized cortical regions ('FFA', 'PPA', and 'VWFA'). Each voxel was assigned to the class with the highest posterior probability given observed activation at that voxel (e.g., in (A), observing activation in green voxels would imply a higher probability that the study was about places rather than faces or words).

Supplementary Figure 4



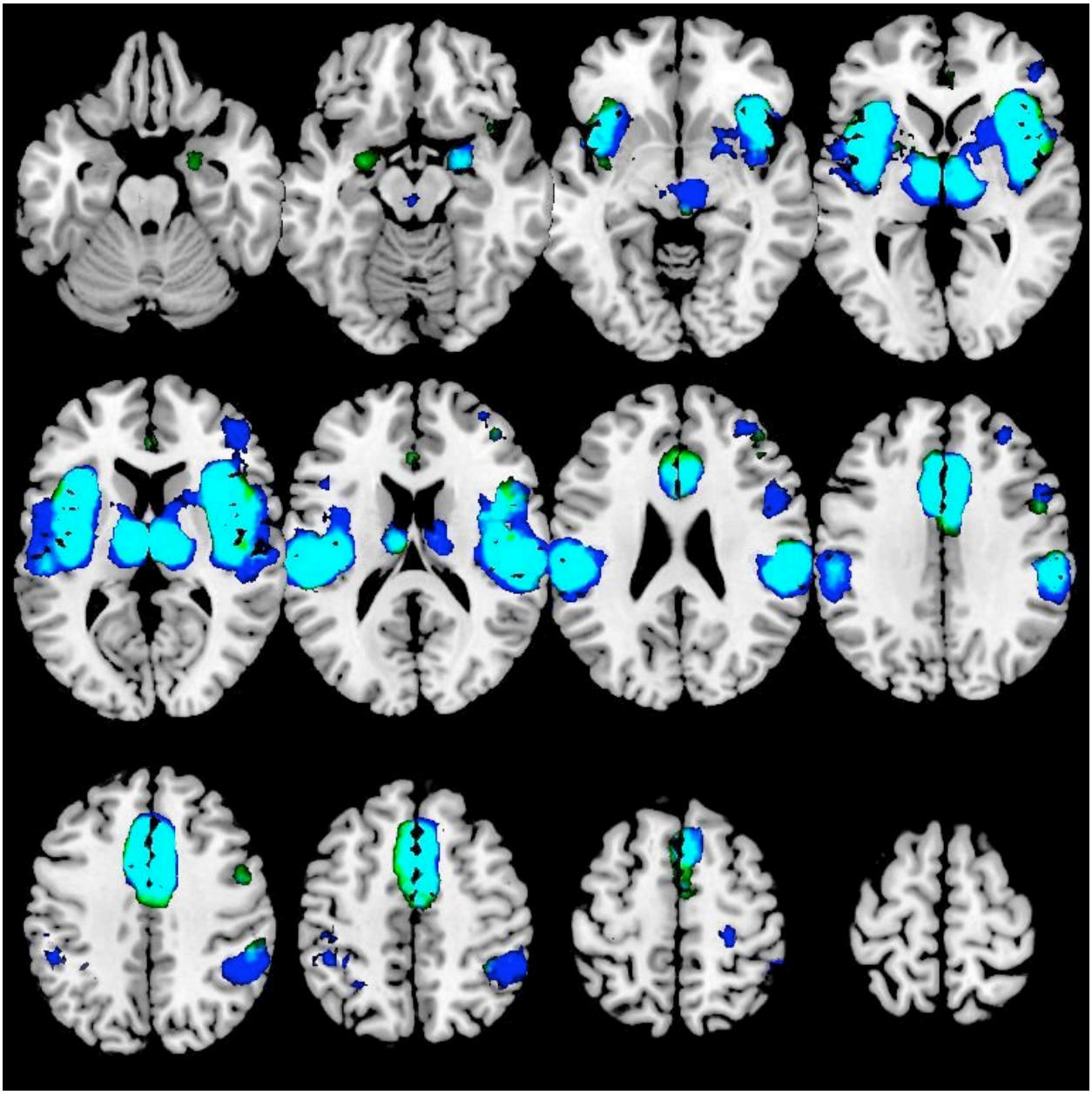
Whole-brain reverse inference meta-analysis maps for the terms 'working memory', 'emotion', and 'pain' when articles are coded strictly based on occurrence of terms in article title rather than in the full article text. Note the close similarity (along with apparent decreased sensitivity) relative to the full analyses reported in the text (Figure 2), suggesting that different sections of published articles carry broadly similar information.

Supplementary Figure 5



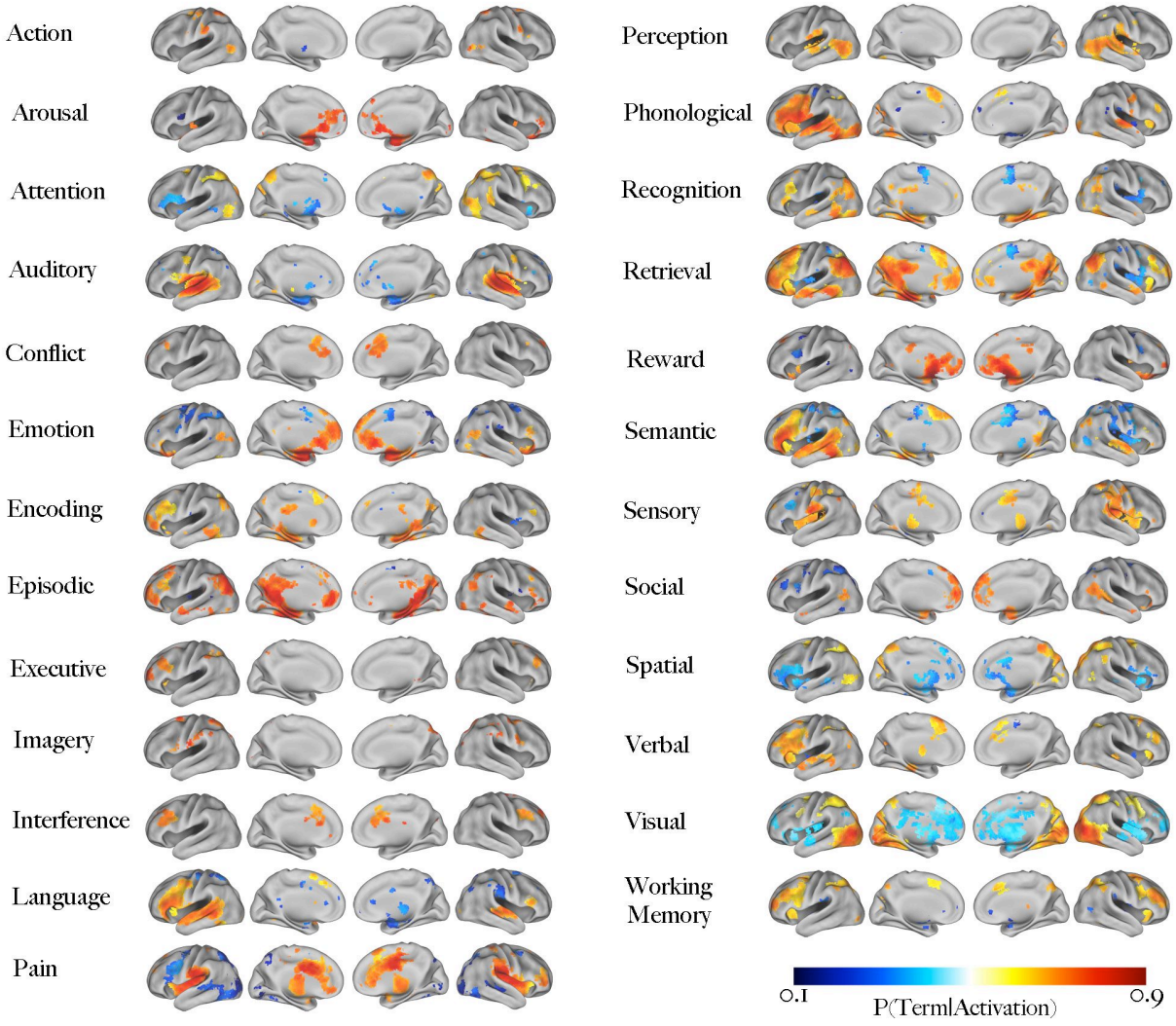
Correlogram displaying pair-wise similarities between manually-generated meta-analysis and mega-analysis maps (Fig. 2A) and automatically-generated forward inference maps (Fig. 2B) and reverse inference maps (Fig. 2C) for the domains of WM, emotion, and pain. For each pair of maps, similarity was computed by binarizing both maps (i.e., distinguishing between active and inactive voxels) and computing the Pearson correlation (lower triangle) or Jaccard index (upper triangle) across all voxels. Decimals are omitted for legibility. MKDA = multi-level kernel density analysis; FI = forward inference; RI = reverse inference; MA = mega-analysis (for pain, the map reflects pooled estimates from 5 different pain studies rather than an MKDA meta-analysis; see ref[3]). Note the high similarity coefficients for maps within the same domain and low coefficients for pairs of maps from different domains.

Supplementary Figure 6



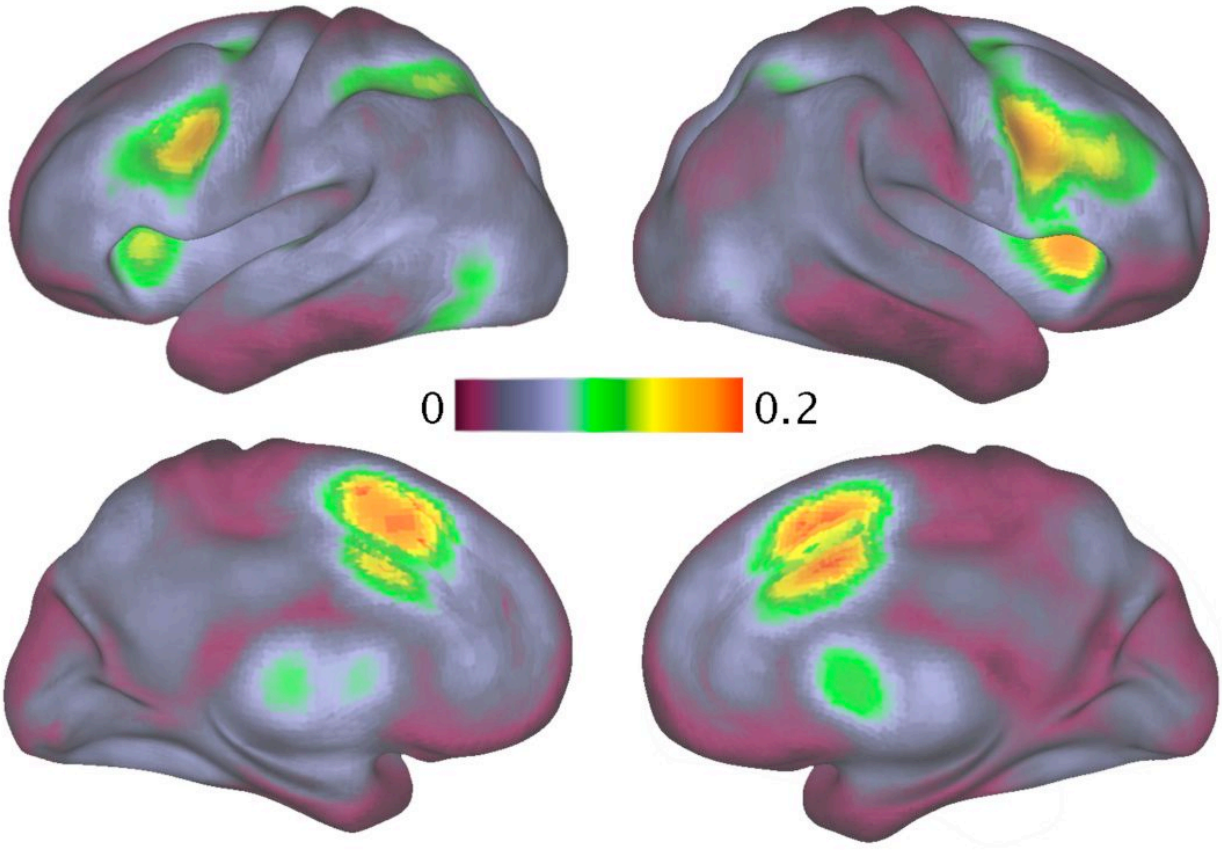
Overlap between meta-analyses based on automatically-coded vs. manually-coded pain data. Green: regions associated with the term 'pain' in a fully-automated (forward inference) meta-analysis (slices correspond to surface rendering displayed in Figure 2B). Blue: MKDA[4] meta-analysis results for a manually validated subset of 66 studies drawn from the automatically extracted dataset that contained valid contrasts between pain and a baseline condition. Cyan: overlap of green and blue. To facilitate direct comparison, results of both analyses are thresholded at the same level ($z = 5$). Across voxels, the correlation coefficient (maps unthresholded) and Jaccard similarity index (maps binarized at the $z = 5$ threshold) were 0.84 and 0.65, respectively.

Supplementary Figure 7



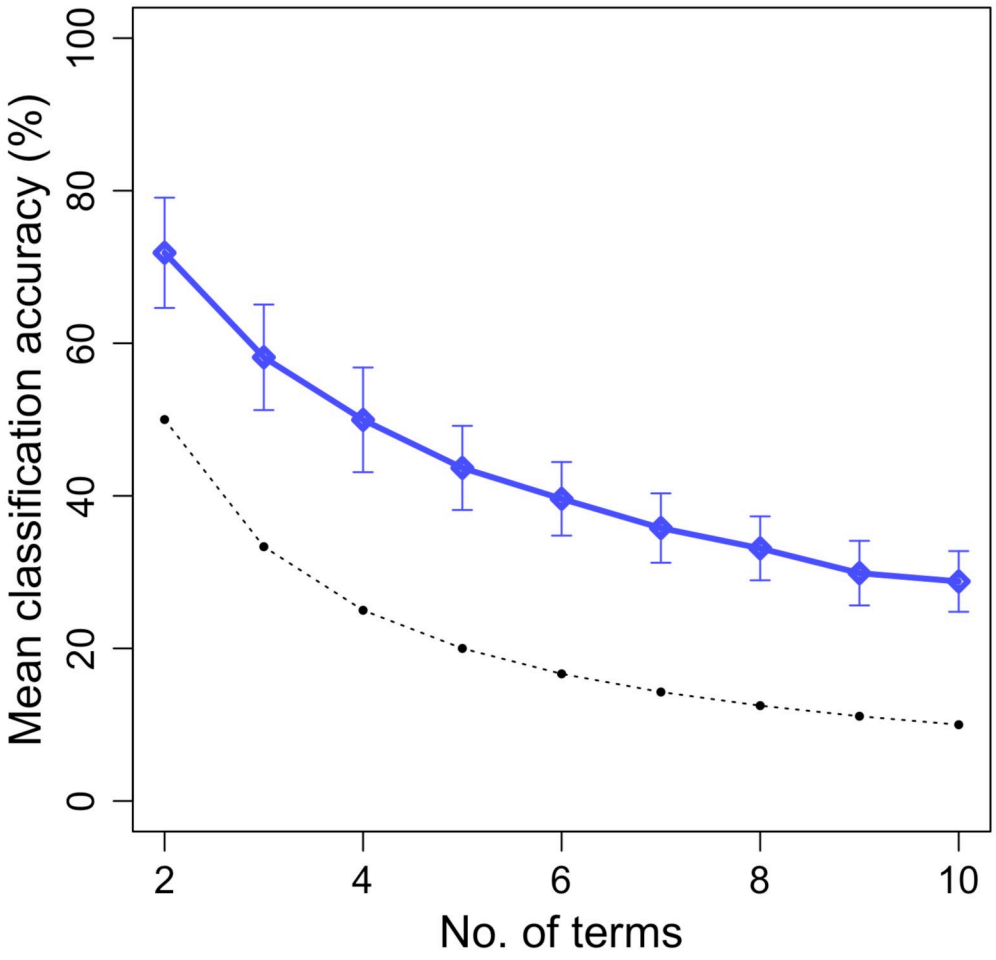
Whole-brain meta-analytic posterior probability maps for 25 key terms that occurred at high frequency (> 1 in 1,000 words) in at least 100 different studies in our database. Voxel values display the probability of the term occurring in a study given observed activation at that voxel (i.e., P(T|A)). To account for base differences in term frequencies, we assume uniform priors for all terms (i.e., equal 50% probabilities of Term and No Term). Activation in orange/red voxels implies a high probability that a term is present, and activation in blue voxels implies a high probability that a term is not present. Values are displayed only for voxels that are significant for a test of association between Term & Activation, with a whole-brain correction for multiple comparisons (FDR = .05). Data available at http://sumsdb.wustl.edu/sums/directory.do?id=8285126.

Supplementary Figure 8



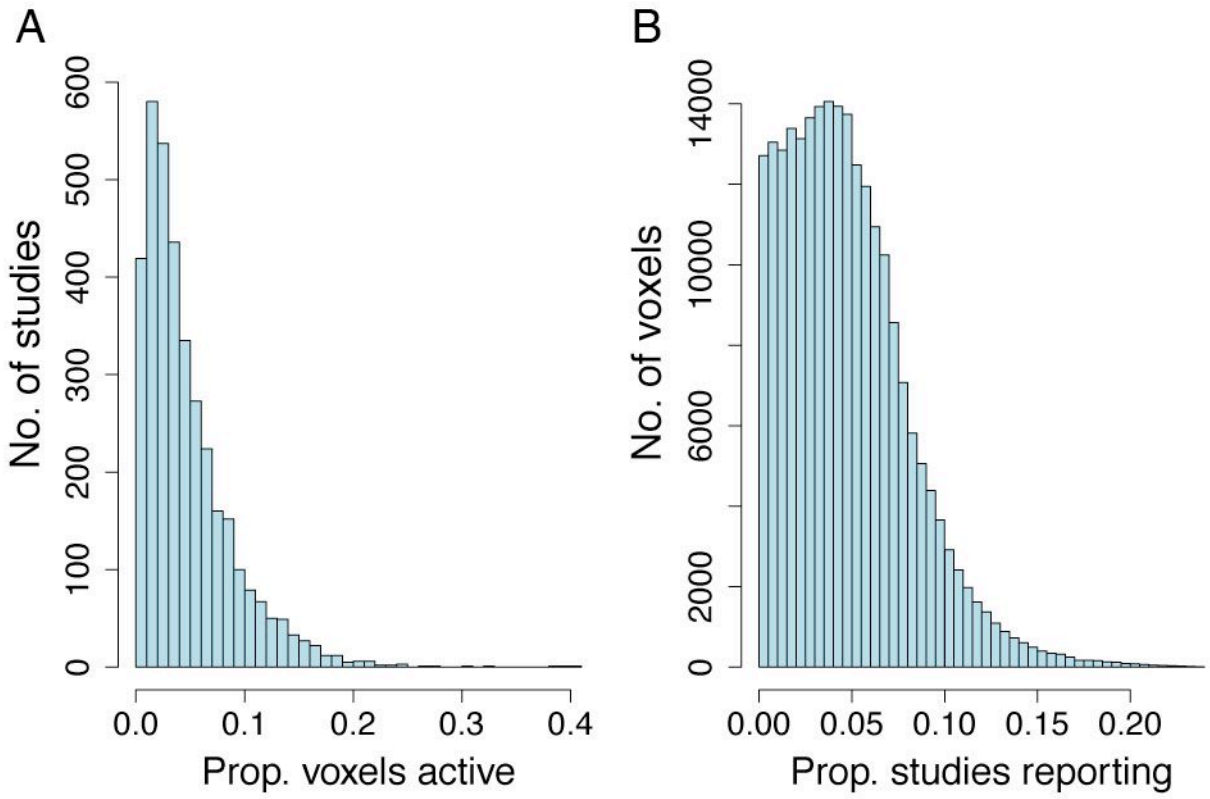
Mean probability of activation at each brain voxel across all 3,489 studies in the database. Frontoparietal regions implicated in cognitive and attentional control were consistently activated at a higher rate than other regions, highlighting the degree to which these areas are nonselectively active (see also Figure 1 in ref.[5]).

Supplementary Figure 9



Accuracy of the naive Bayes classifier as a function of the number of terms being classified. *N*-sized subsets of terms were repeatedly sampled at random from a set of 25 high-frequency terms (cf. Figures 4 and S3). Blue: mean classification accuracy for one hundred random draws. Accuracy was averaged across classes rather than studies to avoid capitalizing on base rate differences between terms. Errors bars reflect the standard deviation. Black: performance level that would be expected by chance.

Supplementary Figure 10



Histograms showing frequency distributions of (A) proportion of voxels reported active in each study and (B) proportion of all studies reporting activity at each voxel. Note that all coordinates reported in articles are convolved with a 10 mm sphere (see Methods); thus, voxels are considered active if they fall within 10 mm of a reported focus.

Supplementary Note

In this note we report additional validation analyses of the NeuroSynth framework. The first section reports a series of supplemental analyses validating the automated coordinate extraction algorithm. The second section reports analyses validating and extending the automated content coding. Throughout both sections, we discuss residual limitations of the NeuroSynth framework and potential directions for future research.

Validation of automated coordinate extraction

As noted in the main text, automated coordinate extraction is susceptible to a number of potential problems that could affect the resulting data quality. First, false positives could occur—that is, the software might incorrectly classify information in a table as an activation focus when it actually represented an entirely different type of information. Second, different software packages and research groups report foci in different stereotactic spaces, resulting in potential discrepancies in the anatomical locations represented by the same set of coordinates across different studies. Third, studies could differ widely in the rigorousness of their experimental and statistical methods, the size of their samples, and the quality of their results, potentially adding noise to the database. Fourth, the software did not discriminate activations from deactivations, and made no attempt to label or categorize foci according to the type of contrast (e.g., task vs. fixation, condition A vs. B, etc.). To address these issues, we conducted a series of additional analyses.

Convergence with manually coded coordinates

To assess the accuracy of the automatic coordinate extraction software, we first compared a set of automatically extracted coordinates with a manually coded set of foci drawn from all studies published in the 2006 and 2007 volumes of Cerebral Cortex and available from SumsDB. Results demonstrated that the automated extraction procedure worked extremely well overall. Eighty-four percent (2929 / 3501) of the foci in the SumsDB reference set were successfully detected by the parser. Inspection of the missing coordinates revealed that in the vast majority of cases, the source of the error was invalid HTML in the table specification. Although we were able to adjust the parser to correctly handle many of these errors, some were idiosyncratic and could only have been handled on a case-by-case basis, which we deemed logistically impractical.

Of the 3334 foci extracted by the automated parser, 405 (12%) were not found in the SumsDB reference set. Careful inspection revealed that 299 of these represented genuine activation foci that were absent from SumsDB; only 106 were 'true' false alarms (though some of these were borderline cases—e.g., foci that were valid, but were from MEG or VBM studies rather than fMRI). Thus, these results suggested an extremely low false positive rate of approximately 3%. This reflects the fact that the parser was designed to be conservative—that is, we deliberately calibrated the software to err on the side of caution

(i.e., to discard any foci that appeared at all questionable). Crucially, there was no reason to expect either false negatives or false positives to produce a systematic bias in our analyses, because the coordinate extraction and semantic tagging procedures were entirely independent of one another. While we are currently working to improve the extraction procedure by developing a more sophisticated parser that uses machine learning techniques to improve its performance with experience, the present results suggest that the current implementation already allows only a small proportion of invalid data into the database.

Contrast-level validation of automatically extracted coordinates

Accurate extraction of coordinates from published articles is necessary but not sufficient for an automated meta-analysis to produce accurate results. If the coordinates extracted by the parser reflect irrelevant experimental contrasts or occur within invalid tables, a meta-analysis could potentially produce null or even misleading results. Because our parser currently lacks the ability to automatically code experimental contrasts, we sought to quantify the loss of signal (if any) associated with the use of a strictly automated approach.

As it was not feasible to manually validate the entire database of over 3,000 studies, we focused on a single psychological domain for which we were able to directly compare automated and manual results. Specifically, we used the NeuroSynth framework to identify 265 studies that used either pain-related terms ('pain', 'painful', 'painfully', 'nociceptive', or 'noxious') or touch-related terms ('touch', 'touched', 'touching', or 'tactile') at high frequency, comprising 8246 activation foci. We then manually inspected and validated all 265 studies. Following inspection, 163 studies (62%) were retained as valid studies that directly investigated pain and/or touch processing (the majority of the 102 excluded studies were relevant to pain or touch—e.g., empathy for pain, real vs. sham acupuncture, etc.—but did not include one or more contrasts directly contrasting relevant pain or touch conditions (e.g., pain vs. rest, touch vs. rest, etc.). Of the 5395 foci automatically extracted from the 162 common studies, 3944 (73%) passed validation and were included in the manually coded dataset (representing 48% of all automatically extracted coordinates).

It is important to note that the majority of the excluded foci were extracted correctly (i.e., the parser identified the correct numerical coordinates), and were excluded because the associated contrast did not meet the stringent criteria of the manual coding. For instance, a large proportion of foci reported pain-related activations at the single-subject (rather than group) level, or reflected contrasts that were only tangentially relevant (e.g., pain empathy vs. rest, heat vs. cold pain, etc.). Although exclusion of these foci from the manual database was clearly the appropriate course, it is important to note that the benefits of an automated approach accrue primarily through large-scale aggregation, and the effects of a decrease in data quality could potentially be offset or even outweighed by the increase in data quantity. That is, excluded studies and coordinates could still contribute useful information to a meta-analysis in the event that the increase in precision and stability of the meta-analytic results provided by larger amounts of data outweighs the increase in noise.

As an empirical test, we compared the (forward inference) pain meta-analysis map produced using the NeuroSynth framework (Figure 2B) with a focused meta-analysis of 66 studies drawn from the manually validated dataset that reported an experimental contrast between painful stimulation and a baseline condition (N = 66 studies reporting pain vs. rest). As Supplementary Figure 6 illustrates, the two approaches produced strikingly similar results (correlation coefficient = .84 across voxels), with partial overlap in virtually all brain regions active in either map. Thus, these results suggest that for broad content terms associated with hundreds of studies and thousands of reported foci, an automated approach can produce largely the same results as a manual approach. However, there is no question that manual coding will continue to be necessary in many if not most cases (e.g., if narrower states such as "pain empathy" are to be distinguished from broad ones such as "pain"), and much of our current work focuses on improving the extraction of metadata that can support more fine-grained automated coding of the experimental contrasts associated with individual activation foci.

Quantification of reported activation increases versus activation decreases

Related to the lack of automated contrast coding is another potential concern that the results generated by our framework inherently confound reported increases and decreases in activation. Because the coordinate parser has no sense of directionality (i.e., it cannot distinguish between task > rest and rest > task), some proportion of the coordinates that reflect reported activation decreases are inevitably treated as activation increases, potentially biasing the results. Because the extent of this problem depends largely on the proportion of total coordinates that constitute activation decreases rather than increases, we sought to quantify the balance between reported increases and decreases in the neuroimaging literature. In lieu of a full manual coding of the entire database, we used the manually-validated pain and touch dataset described above, which included a standardized coding of the contrast corresponding to each valid activation (e.g., pain > rest, high pain > low pain, rest > touch, etc.).

The results of the manual coding indicated that activation decreases were reported much less often than activation increases for all major contrasts. For instance, for comparisons between pain and rest (1863 coordinates), 94% of coordinates were increases (pain > rest) and only 6% were decreases; for comparisons between touch and rest (1293 coordinates), 95% were increases; and for comparisons between high and low pain (548 coordinates), 83% were increases. These findings suggest that, at least for the tested domains of pain and touch, activation decreases constitute a relatively small proportion of activations reported in tables in published neuroimaging articles, and thus appear to exert minimal influence on meta-analytic results (cf. Supplementary Figure 6). Nonetheless, since the extent to which this conclusion holds may vary across psychological domains, future efforts should seek to develop automated ways of coding increases versus decreases or activations versus deactivations—at least for a subset of contrasts that can be relatively easily identified (e.g., those involving terms like 'rest', 'baseline', etc.).

Automated stereotactic space detection and coordinate transformation

The Automated Coordinate Extraction software made no attempt to identify the stereotactic space in which coordinates from different studies were reported, or to correct for between-study differences in spaces[6,7]. It is unlikely that such space differences heavily influenced our results, because (a) the great majority of studies (approximately 75 – 80%) are reported using an MNI-based space, and (b) the spatial specificity of meta-analytic results is already limited by the use of a 10 mm smoothing kernel and the marked heterogeneity in preprocessing procedures used in different studies. Nonetheless, the ability to detect and correct for space differences in an automated way would undoubtedly help reduce error by maximizing overlap between coordinates. In an effort to implement an automated space correction procedure, we have begun to develop an algorithm that (a) identifies the stereotactic space used in each study based on the usage of key terms within the article text, and (b) uses an existing affine transformation developed by Lancaster and colleagues[2,7] to convert coordinates from different spaces to a common reference space.

At present, our algorithm distinguishes only between the two most common stereotactic spaces—namely, the MNI-based space used by default in SPM and FSL, and the Talaraich & Tournoux[1] (T88)-based space used by default in AFNI and BrainVoyager. The algorithm assigns the label MNI or T88 to a study in the event that one or more keywords associated predominantly with one space is used at least once in the article text AND there are no occurrences of words predominantly associated with the opposite space. For instance, the occurrence of terms such as 'MNI', 'SPM', and 'FSL' in the absence of any terms like 'Talairach', 'AFNI', or 'BrainVoyager' would be taken to imply that data were reported in MNI space, and vice versa for T88 space (the exception is that the term 'Talairach' is not taken as evidence *for* the use of Talairach space, because many researchers use the term to refer generically to all stereotactic coordinate systems). If the algorithm detects competing evidence (e.g., the terms 'BrainVoyager' and 'MNI' are both used), as might happen when researchers use BrainVoyager software with a non-default template), the label 'UNKNOWN' is assigned (9% of all studies), and no transformation is applied.

To validate the accuracy of our algorithm, we selected a random subset of 100 studies from the database and manually examined each one to identify the originating space. Inspection revealed that the automated space detection algorithm performed relatively well overall. Fifty-eight of 66 (88%) of studies in MNI space were correctly labeled MNI (2 were labeled T88, and 6 unknown), and 15 of 31 (48%) of studies in T88 space were correctly labeled T88 (12 were labeled MNI, and 4 unknown). The relatively high false negative rate for T88 studies was attributable largely to the fact that 10 studies conducted analyses in MNI space but reported transformed coordinates in T88 space. While the results reported here are preliminary, and do not take MNI-to-T88 transformation into account, we anticipate that modifying the algorithm to account for such transformations will be relatively straightforward, as a relatively small set of terms appear to be highly diagnostic (e.g., the terms 'Brett', 'Lancaster', 'converted', or 'transformed' in close proximity to the term 'Talairach'). Thus, we expect a final version of the automated correction algorithm to perform with high accuracy once fully integrated with the NeuroSynth framework. However, it is important to note that our current approach only distinguishes between MNI

and T88 spaces; it does not distinguish between more subtle differences in template (e.g., SPM99, SPM05, and FSL all use slightly different templates by default), and is deliberately conservative, making no attempt to categorize ambiguous studies (9% of studies are labeled unknown). Ongoing work aims to directly address these limitations; in the interim, we expect that relatively limited effort will be required to manually inspect and label the small proportion of studies that use an unclassified space.

Once space labels are assigned, coordinate can be automatically converted between stereotactic spaces using existing affine transformations. Because most studies (approximately two-thirds) report data in MNI space rather than T88, we converted coordinates from T88 studies to MNI rather than the converse so as to minimize error induced by transformation. We used an inverted version of the previously validated icbm_spm2tal affine transformation developed by Lancaster et al[2]. (We chose the SPM version of the transformation rather than the FSL or pooled versions because SPM is by far the most commonly used neuroimaging software package, and the majority of coordinates in published articles are consequently reported for an SPM-based template. Future extensions will support software-specific space detection and transformation.)

To validate the automated application of the Lancaster transformation, we compared the results obtained for studies in TAL space before and after transformation relative to studies in MNI space. Supplementary Figure 1 presents sample results for a reverse inference meta-analysis of the term 'amygdala' with (top) and without (bottom) the transformation. The results demonstrate that differences between MNI and T88 in the spatial localization of the amygdala are substantially reduced following transformation of T88 coordinates, though they remain noticeable, particularly along the rostral/ventral axis. (It is presently unclear whether the residual differences reflect sampling error due to the use of mutually exclusive studies, the misidentification of some T88 studies as MNI studies, or fundamental limitations of the affine transformation, which cannot account for non-linear differences.)

To quantitatively assess the effects of the coordinate transformation algorithm on our meta-analysis results, we automatically generated new meta-analysis maps for 30 common terms drawn from the sets in Figure 5 and Supplementary Figure 2, conducting separate analyses for studies reporting coordinates in MNI space and in T88 space. We then computed the correlation coefficient between the MNI and T88 maps across all voxels both before and after applying the Lancaster transformation. As expected, for virtually all terms (28 of 30), a stronger correlation was observed post-transformation (mean $r = .66$) than pre-transformation (mean $r = .60$; paired t-test, $p < .001$), demonstrating that it is possible to detect and compensate for differences in stereotactic space to a significant extent in an automated way.

Validation and extension of automated content coding

At present, automated coding of article contents is based exclusively on a lexical approach, which assumes that usage rates of individual words provide a reasonable proxy for more effortful manual coding of the psychological processes investigated by individual

neuroimaging studies. The results presented in the main text provide strong support for this assumption, as it would not have been possible to successfully classify study-level and subject-level data if the meta-analysis maps did not accurately reflect stable mappings between cognitive and neural states. Nonetheless, to ensure the accuracy and robustness of the lexical approach we conducted additional validation analyses detailed below.

Convergence with anatomically-defined regions

First, we demonstrated that the lexical approach could recapture conventional boundaries between distinct anatomical regions reasonably accurately. We compared the maps generated using lexical mapping for key anatomical terms (e.g., 'amygdala', 'hippocampus', and 'parahippocampal') with the regional boundaries found in the widely used Harvard-Oxford anatomical atlas[8]. The anatomical labels we used for the lexical meta-analyses were derived from the PubBrain neuroanatomical ontology (pubbrain.org); we selected only those terms that occurred at a high (> 50 studies) frequency, excluding very broad terms (e.g., frontal lobe, telencephalon, etc.). Supplementary Figure 2 displays boundaries for selected regions as defined by the lexical analysis versus the Harvard-Oxford atlas. Because the posterior probability maps could be thresholded arbitrarily, we restricted each lexical ROI to the same number of voxels present in the ROI defined in the Harvard-Oxford atlas. The resulting lexically-defined ROIs were reasonably similar to the corresponding anatomical regions, even for relatively small structures (e.g., the amygdala). Note that dissimilarities between the two maps are not solely attributable to errors in the automatic extraction procedure, as researchers often use anatomical labels somewhat idiosyncratically.

Identification of functionally-selective cortical regions

Second, we used the lexical approach to replicate previous findings of category-specific activation for visual object recognition in posterior cortical regions. The terms 'faces', 'words', and 'places' were strongly and selectively associated with activations in the putative fusiform face area (FFA[9]), visual word form area (VWFA[10]), and parahippocampal place area (PPA[11]), respectively (Supplementary Fig. 3, top). These mappings were further confirmed by searches for 'FFA', 'VWFA', and 'PPA', which aligned closely with the results of the category-based search (Supplementary Fig. 3, bottom).

Convergence with prior literature

Third, we used the lexical approach to generate whole-brain meta-analysis maps for 25 high-frequency terms corresponding to concepts that have been extensively studied in the fMRI literature (Supplementary Fig. 7). The results closely replicated numerous previous studies, with different sets of terms activating expected brain networks. For instance, the terms 'conflict', 'executive', 'interference', and 'working memory' most selectively activated

medial and lateral frontal regions implicated in working memory and executive control; language-related terms such as 'language', 'phonology', 'semantic', and 'verbal' were associated with strongly left-lateralized activations in temporal and ventrolateral prefrontal regions; modality-related terms such as 'visual', 'auditory', and 'sensory' were associated with selective activations in visual, auditory, and sensory cortices, respectively; and so on.

Title-based lexical analyses produce similar results

The empirical evidence reported above suggests that, at least for broad psychological domains, term-based meta-analyses can provide accurate and robust results that converge with prior findings. Nonetheless, it is clear that the present approach has a number of limitations that will be important to address in future work. One is that article coding is currently based solely on the frequency with which terms occurs anywhere in an article and does not take into account contextual information such as the location of a word or its relation to other nearby words. A priori, it is plausible that some article sections (e.g., title, abstract, table captions, or results) might carry more diagnostic information than others (e.g., the introduction or discussion), or that words should be weighted based on their immediate context (e.g., greater weight for words with proximal references to tables or figures). As a preliminary effort in this direction, we conducted a full set of meta-analyses identical to those reported in the text but based solely on the occurrence of words in article titles (rather than anywhere in the text). Despite the large reduction in number of studies associated with each term (most terms, the title-based searches returned only 10 – 20% the number of studies identified by the full-text search), the results were very similar to (if somewhat less sensitive than) those obtained using full-text searches (see Supplementary Fig. 4 for examples), suggesting that the information contained in the full text of articles overlaps closely with that conveyed by article titles.

In future work, we intend to further refine our approach by modeling terms in different article sections separately and empirically identifying weightings that maximize the sensitivity and specificity of the meta-analytic results. Along similar lines, one could also assign different weights to studies—for instance, assigning greater weight to studies that are perceived as more authoritative (e.g., having higher citation counts) or are authored by researchers known to work closely in a particular domain. Our hope is that the availability of the tools, data, and results introduced here will encourage other researchers to contribute to such efforts and implement innovative extensions.

Beyond individual terms

A second limitation of the present implementation is that it is based exclusively on counts of individual words or phrases, whereas the contents of articles are probably better captured using groups of words that naturally coalesce into coherent topics. To facilitate an eventual shift from term-based analyses to topic-based analyses, we have developed a rudimentary syntax for conducting analyses that involve combinations of multiple words.

Our code (available at http://neurosynth.org) embeds a parsing expression grammar[12], enabling users to recursively nest arbitrarily complex queries. For instance, the query:

"(disgust | sad* | anger | fear* | anx*) &~ (pain* | noxious | nocicept*)"

would select all studies that use one or more negative emotion terms (e.g., disgust, sad, sadness, anger, etc.) but NOT one or more pain-related terms (e.g., pain, painful, noxious, etc.). This approach facilitates dynamic analyses that go beyond individual words and allow users to conduct sophisticated meta-analyses targeting arbitrary word combinations. We have also begun to explore other approaches that attempt to model latent topics or clusters in the text of neuroimaging articles (e.g., using latent Dirichlet allocation or multidimensional scaling); however, such analyses are beyond the scope of the present article.

References

1. Talairach, J. & Tournoux, P. *Co-Planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging*. (Thieme: 1988).

2. Lancaster, J.L. et al. Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Human brain mapping* **28**, 1194-205 (2007).

3. Atlas, L.Y., Bolger, N., Lindquist, M.A. & Wager, T.D. Brain Mediators of Predictive Cue Effects on Perceived Pain. *Journal of Neuroscience* **30**, 12964 (2010).

4. Wager, T.D., Lindquist, M.A., Nichols, T.E., Kober, H. & Van Snellenberg, J.X. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage* **45**, 210-221 (2009).

5. Nelson, S.M. et al. Role of the anterior insula in task-level control and focal attention. *Brain Structure and Function* 1-12 (2010).

6. Van Essen, D.C. & Dierker, D.L. Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron* **56**, 209-225 (2007).

7. Laird, A.R. et al. Comparison of the disparity between Talairach and MNI coordinates in functional neuroimaging data: Validation of the Lancaster transform. *Neuroimage* **51**, 677-683 (2010).

8. Smith, S.M. et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**, S208-S219 (2004).

9. Kanwisher, N., McDermott, J. & Chun, M.M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* **17**, 4302-4311 (1997).

10. McCandliss, B.D., Cohen, L. & Dehaene, S. The visual word form area: expertise for reading in the fusiform gyrus. **7**, 293-299 (2003).

11. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598-601 (1998).

12. Ford, B. Parsing expression grammars. *ACM SIGPLAN Notices* **39**, 111-122 (2004).