Decoding the Representation of Code in the Brain: An fMRI Study of Code Review and Expertise

Benjamin Floyd, Tyler Santander, **Westley Weimer**
University of Virginia
University of Michigan

University of Michigan

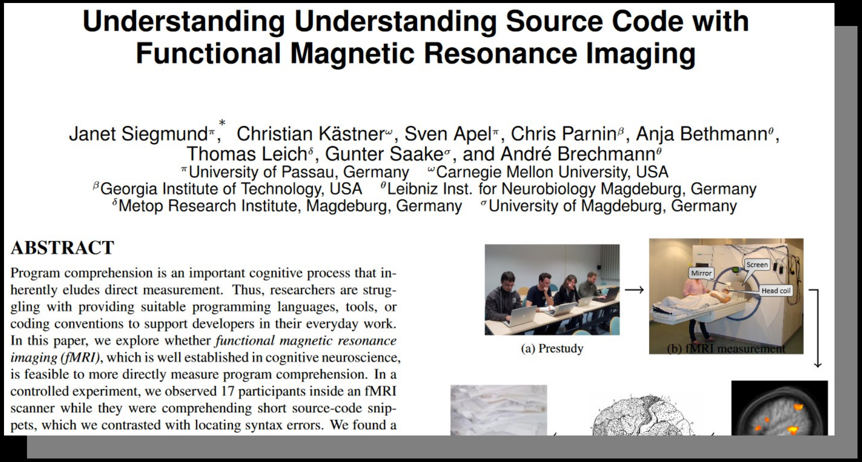
EECS ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

- Looking to grow in PL/SE over next few years

- Have your senior PhD students contact me

“Understanding Understanding Source Code” (ICSE 2014)

- Described an fMRI study framework for SE
- Found five brain regions associated with code comprehension
- Encouraged future fMRI+SE research



Understanding Understanding Source Code with Functional Magnetic Resonance Imaging

Janet Siegmund[π],[*] Christian Kästner[ω], Sven Apel[π], Chris Parnin[β], Anja Bethmann[θ],
Thomas Leich[δ], Gunter Saake[σ], and André Brechmann[θ]
[π]University of Passau, Germany [ω]Carnegie Mellon University, USA
[β]Georgia Institute of Technology, USA [θ]Leibniz Inst. for Neurobiology Magdeburg, Germany
[δ]Metop Research Institute, Magdeburg, Germany [σ]University of Magdeburg, Germany

ABSTRACT

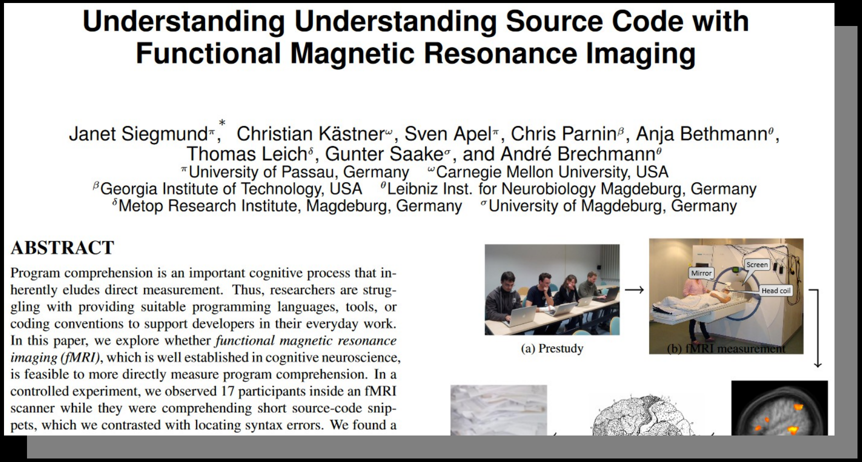
Program comprehension is an important cognitive process that inherently eludes direct measurement. Thus, researchers are struggling with providing suitable programming languages, tools, or coding conventions to support developers in their everyday work. In this paper, we explore whether *functional magnetic resonance imaging (fMRI)*, which is well established in cognitive neuroscience, is feasible to more directly measure program comprehension. In a controlled experiment, we observed 17 participants inside an fMRI scanner while they were comprehending short source-code snippets, which we contrasted with locating syntax errors. We found a

(a) Prestudy

(b) fMRI measurement

“Understanding Understanding Source Code” (ICSE 2014)

- Described an fMRI study framework for SE
- Found five brain regions associated with code comprehension
- Encouraged future fMRI+SE research
- Today: *Understanding* 'Understanding Understanding Source Code' ?



Understanding Understanding Source Code with Functional Magnetic Resonance Imaging

Janet Siegmund$^{\pi}$,* Christian Kästner$^{\omega}$, Sven Apel$^{\pi}$, Chris Parnin$^{\beta}$, Anja Bethmann$^{\theta}$,
Thomas Leich$^{\delta}$, Gunter Saake$^{\sigma}$, and André Brechmann$^{\theta}$
$^{\pi}$University of Passau, Germany $^{\omega}$Carnegie Mellon University, USA
$^{\beta}$Georgia Institute of Technology, USA $^{\theta}$Leibniz Inst. for Neurobiology Magdeburg, Germany
$^{\delta}$Metop Research Institute, Magdeburg, Germany $^{\sigma}$University of Magdeburg, Germany

ABSTRACT

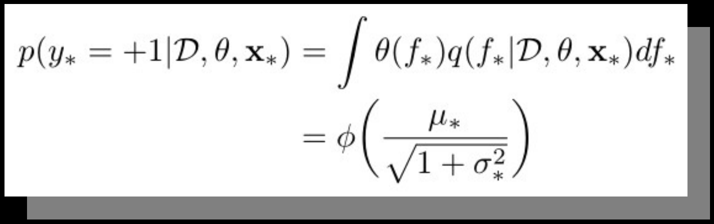
Program comprehension is an important cognitive process that inherently eludes direct measurement. Thus, researchers are struggling with providing suitable programming languages, tools, or coding conventions to support developers in their everyday work. In this paper, we explore whether *functional magnetic resonance imaging (fMRI)*, which is well established in cognitive neuroscience, is feasible to more directly measure program comprehension. In a controlled experiment, we observed 17 participants inside an fMRI scanner while they were comprehending short source-code snippets, which we contrasted with locating syntax errors. We found a

(a) Prestudy

(b) fMRI measurement

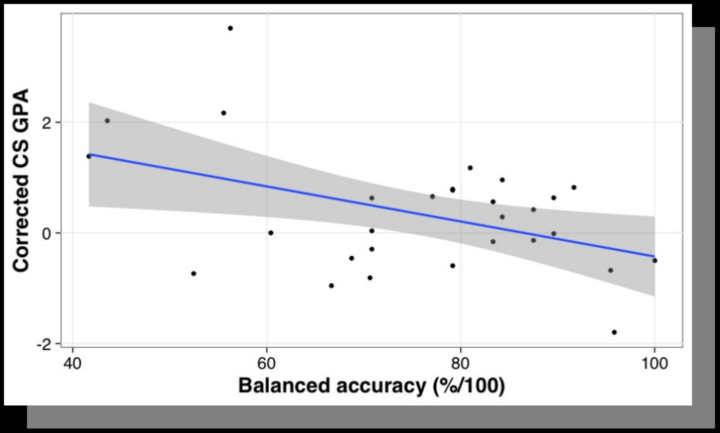
Special Note – This Talk

- Advertisement for the paper
  - Elide analysis details for time
  - Confidence in results



$$p(y_* = +1|\mathcal{D},\theta,\mathbf{x}_*) = \int \theta(f_*)q(f_*|\mathcal{D},\theta,\mathbf{x}_*)df_*$$
$$= \phi\left(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\right)$$

- Motivation and Background
- Experiment and Results



- Call to Arms

Expertise

- Individual differences in programming and debugging time, as well as program efficiency, can vary up to 28:1
- Novices and experts solve physics problems with different efficiency and categorize them differently
- Medical imaging studies have found neural correlates of expertise/learning in golf, juggling, London taxi navigation, etc.
  - Could this apply to CS?

Functional Magnetic Resonance Imaging (fMRI)

- Noninvasive way to study the neurobiological substrates of cognitive functions *in vivo*
- Which parts of the brain are in use?
- Your brain needs energy but does not store it
- So can track where oxygen is consumed
  - Oxygenated and deoxygenated hemoglobin have different magnetic properties that can be detected
  - Millimeter scale (>> EEG or PET, etc.)
  - Blood-oxygen level dependent (BOLD) signal

A Study in Contrasts

- A subject might be doing multiple things
  - e.g., reading code *and* being nervous
- How can we tell if an observed pattern of activation corresponds to one activity?
- Experimental design and control
  - Task A = "reading code + nervous + …"
  - Task B = "reading prose + nervous + …"
- The contrast A-B shows patterns of brain activation that *vary* between the stimuli/tasks

High-Level Question

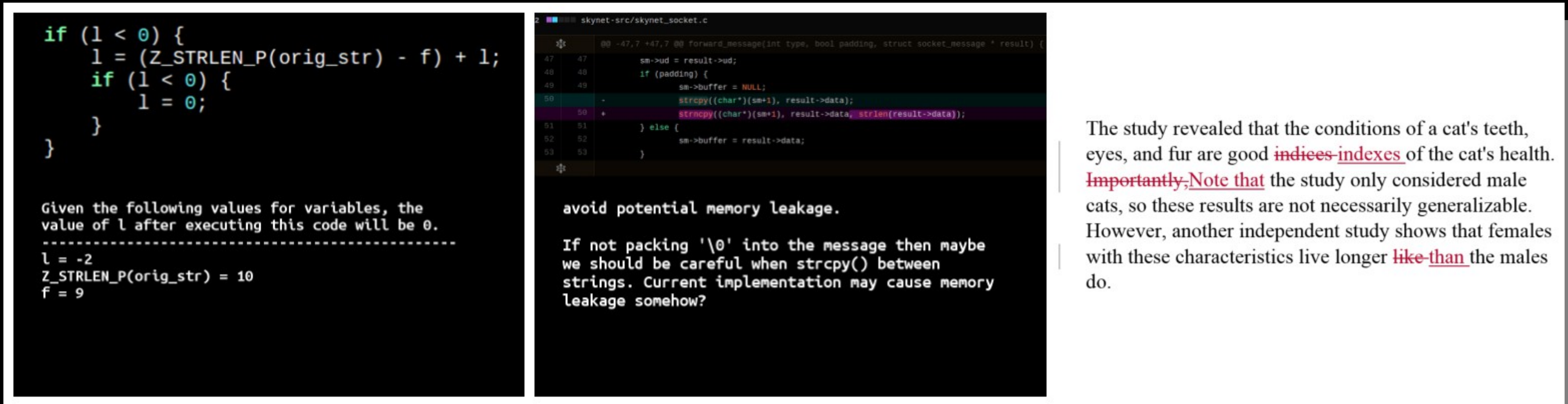
Is reading code more like doing math or more like reading prose?

Code Review and Comprehension

- Developers spend more time understanding and comprehending code than any other activity
  - NASA: understanding > correctness for reuse
- Code review is a de facto standard
  - “Should we accept this commented patch?”
  - Mandated in Facebook, Google, etc.
  - One of the most effective techniques in software development

Experimental Design: 3 Tasks

- Code Comprehension
- Code Review (top 100 GitHub repos)
- Prose Review (College Board SAT, etc.)



```
if (l < 0) {
    l = (Z_STRLEN_P(orig_str) - f) + l;
    if (l < 0) {
        l = 0;
    }
}
```

```
Given the following values for variables, the
value of l after executing this code will be 0.
-----------------------------------------------
l = -2
Z_STRLEN_P(orig_str) = 10
f = 9
```

(a) Code Comprehension

```
skynet-src/skynet_socket.c
@@ -47,7 +47,7 @@ forward_message(int type, bool padding, struct socket_message * result) {
        sm->ud = result->ud;
        if (padding) {
                sm->buffer = NULL;
-               strcpy((char*)(sm+1), result->data);
+               strncpy((char*)(sm+1), result->data, strlen(result->data));
        } else {
                sm->buffer = result->data;
        }
```

```
avoid potential memory leakage.

If not packing '\0' into the message then maybe
we should be careful when strcpy() between
strings. Current implementation may cause memory
leakage somehow?
```

(b) Code Review

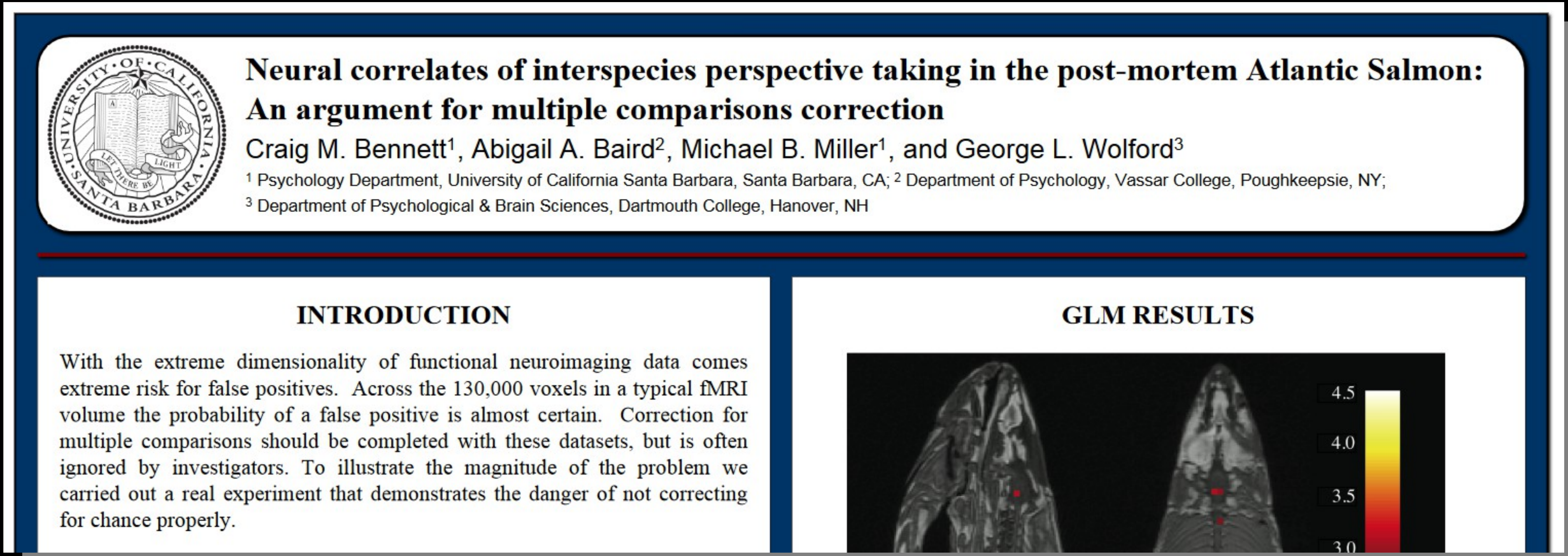
The study revealed that the conditions of a cat's teeth, eyes, and fur are good ~~indices~~ indexes of the cat's health. ~~Importantly,~~ Note that the study only considered male cats, so these results are not necessarily generalizable. However, another independent study shows that females with these characteristics live longer ~~like~~ than the males do.

(c) Prose Review

Experiment Setup and Data

- 29 grads and undergrads (38% women)
  - Right-handed, native English speakers, corrected-to-normal vision, IRB-HSR #18420, etc.
- Placed in fMRI, computer projection displayed via mirror
- A single participant completing four 11-minute runs produces 399,344,400 floating point numbers of data (153,594 voxels × 650 volumes × 4 runs)

Dead Fish and Software Bugs



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

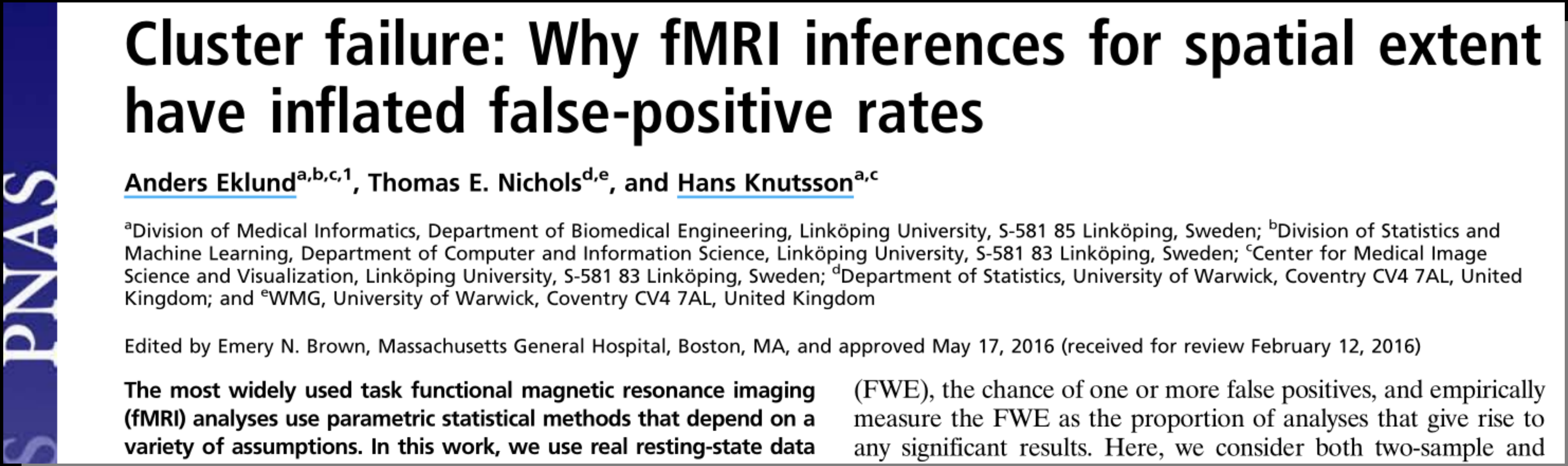
Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

GLM RESULTS



Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

[a]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; [b]Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; [c]Center for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; [d]Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and [e]WMG, University of Warwick, Coventry CV4 7AL, United Kingdom

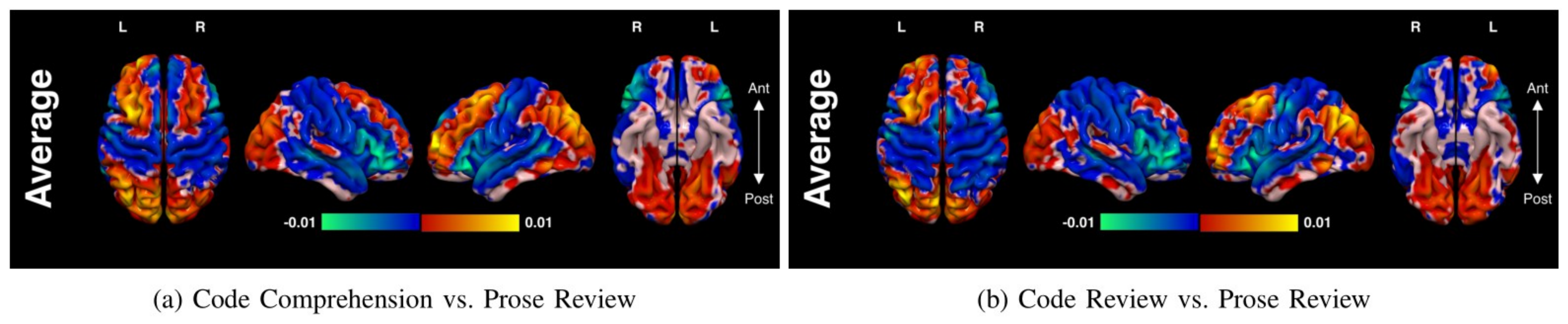
Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

The most widely used task functional magnetic resonance imaging (fMRI) analyses use parametric statistical methods that depend on a variety of assumptions. In this work, we use real resting-state data

(FWE), the chance of one or more false positives, and empirically measure the FWE as the proportion of analyses that give rise to any significant results. Here, we consider both two-sample and

Results: Mind Reading

- We can classify which task a participant is undertaking based solely on brain activity
  - Balanced accuracy 79%, p < .001
- These results suggest that Code Review, Code Comprehension, and Prose Review all have largely distinct neural representations

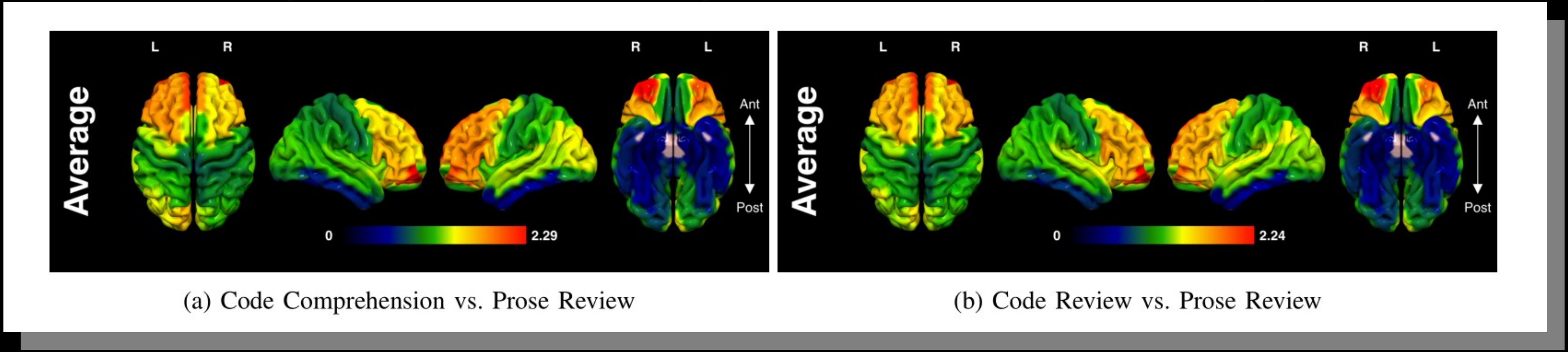


(a) Code Comprehension vs. Prose Review

(b) Code Review vs. Prose Review

Results: Can we relate tasks to brain regions?

- Near-perfect correspondence: $r=0.99$, $p<.001$



(a) Code Comprehension vs. Prose Review

(b) Code Review vs. Prose Review

- A wide swath of prefrontal regions known to be involved in higher-order cognition (executive control, decision-making, language, conflict monitoring, etc.) were highly weighted
  - Activity in those areas strongly drove the distinction between code and prose processing

Results: Can we relate expertise to classification accuracy?

- “Expertise” = (CS GPA) * (CS Credits Taken)
- How accurately our model distinguishes between Code Comprehension and Prose significantly predicted expertise ($r = -0.44$, $p=0.016$)
- The inverse relationship between accuracy and expertise suggests that, as one develops more skill in coding, the neural representations of code and prose are less differentiable. That is, programming languages are treated more like natural languages with greater expertise.

Costs and Reproducible Research

- Easy: recruiting
- Medium: equipment cost ($500/hour)
- Hard: IRB, HIPAA, experimental design
- All datasets and materials available online
    - Including IRB protocol application, recruitment materials, screening forms, training videos, visual stimuli, etc.
    - http://dijkstra.cs.virginia.edu/fmri/

Future Studies

- Social relationships (boss over shoulder)
- Patch provenance (cheating)
- Industrial expertise (replicate protocol)
- Writing code (fMRI-safe keyboard)
- Transcranial magnetic stimulation (read-write)

- Does any of this sound interesting? …

Call To Arms

- By what mechanism do humans experience consciousness?
  - “Extending the human subjective experience of consciousness over time” is a most important problem: “NP-Hard” in the sense that solving it would allow us to solve others. Is it solvable?
- I have funding and am looking for collaborators
  - Come talk to me

Conclusion

- These studies are still exploratory
  - The area is wide open for future work
- Neural representations of programming and natural languages are distinct
- Our classifiers distinguish them based solely on brain activity
  - The same brain locations distinguish these tasks
- Greater expertise accompanies a less-differentiated neural representation

Bonus Slides

Medical Imaging and CS

Future Potential

- Replace unreliable self-reporting
- Inform pedagogy
- Retrain aging engineers
- Guide technology transfer
- Understand expertise
- Foundational, fundamental understanding

Preprocessing and Overfitting

- A significant challenge in fMRI analysis is processing the data correctly
- We cannot naively build a model from 150,000 features and 100 labeled instances
- Align and unwarp data, coregistered with a high-resolution anatomical scan, generalized linear models, high pass filters, robust weighted least squares, multivariate Gaussian process classification, feature selection via Automated Anatomical Labeling atlas, kernel function, expectation propagation …

Taxi Driver Study

“We found that compared with bus drivers, taxi drivers had greater gray matter volume in mid-posterior hippocampi and less volume in anterior hippocampi. Furthermore, years of navigation experience correlated with hippocampal gray matter volume only in taxi drivers, with right posterior gray matter volume increasing and anterior volume decreasing with more navigation experience.”

- Maguire et al., London taxi drivers and bus drivers: a structural MRI and neuropsychological analysis.