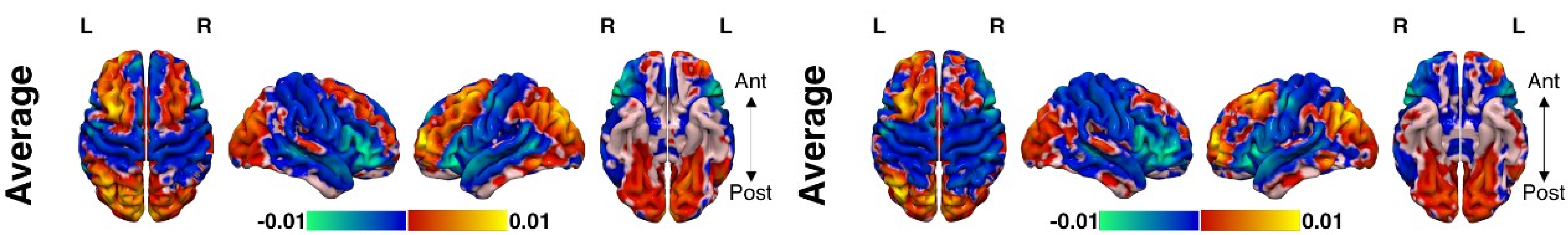
What goes on in your brain when you read and understand code?

Westley Weimer
University of Michigan



(a) Code Comprehension vs. Prose Review

(b) Code Review vs. Prose Review

Goal and Skepticism

- This talk provides a gentle introduction to a new and exciting area!
  - What has been done? What may be possible?

- Setup: assume you are skeptical
- Lay out logical arguments and citations
- But also highlight the costs

Outline (30-40 minutes)

- Program Comprehension
- Selected Psychology Results
- Neuroimaging Explained
- Software Engineering + Medical Imaging
- Costs and Challenges
- Call to Arms

Chain Of Reasoning (1/3)

- We want to reduce the costs associated with software development and maintenance
  - Leaves more resources to add features, improve quality, increase assurance and trust, etc.
- Program comprehension is a dominant SE activity
  - Amdahl's Law suggests we optimize it
- Abstractly, comprehension involves two big components: the human and the program

Chain of Reasoning (2/3)

- Improving the Program Aspect
  - Research for writing (e.g., style guidelines, comment content, commit messages, etc.)
  - Research for manipulating (e.g., synthesizing comments, code clones, refactoring, etc.)
  - Research for viewing (e.g., semantic search, better GUIs, debugging, specification mining, etc.)
- Historically, this is the “default” approach

Chain of Reasoning (3/3)

- Improving the Human Aspect
  - Training and pedagogy (novice → expert)
  - … ?
- By my count, of 41 ICPC papers this year
  - Only 4 target human cognition explicitly
  - “Replicating Novices”, “Live Programming”, “Cognitive Load”, and “CodersMUSE”
- cf. Streetlight Effect
  - Looking at the human side hasn't been as feasible

The Human Aspect Matters

- Early study of 12 industrial developers found order-of-magnitude individual variations [H. Sackman, W. J. Erikson and E. E. Grant. *Exploratory Experimental Studies Comparing Online and Offline Programming Performance*. Communications of the ACM, 1968.]

Metric	Poorest	Best	Ratio
Debugging Hours Algebra	170	6	28:1
Debugging Hours Maze	26	1	26:1
CPU Seconds Algebra	3075	370	8:1
CPU Seconds Maze	541	50	11:1
Code Writing Hours Algebra	111	7	16:1
Code Writing Hours Maze	50	2	25:1
Program Size Algebra	6137	1050	6:1
Program Size Maze	3287	651	5:1
Run Time Algebra	7.9	1.6	5:1
Run Time Maze	8.0	0.6	13:1

Stopwatch and Scoresheet

- We can conduct controlled experiments
- Sackman et al. study just measured time and accuracy
  - e.g., array-sorting program vs. list-sorting program
- Can measure effect of a feature
- But cannot always explain why
  - Often critical for generalizing & recommendations
- Self-reporting may not be reliable
  - How does psychology research address this?

Audience Participation: Hypothesis Time

- Do experts make fewer errors than novices?
- Do experts use a different approach than novices?

Example Implications

- If experts make fewer errors than novices …
- If experts use a different approach than novices …
- … we should teach care, have tools to automate error-prone tasks, have tools to catch mistakes, etc.
- … we should teach new approaches, have tools guide their steps, have tools identify when they can be applied, etc.

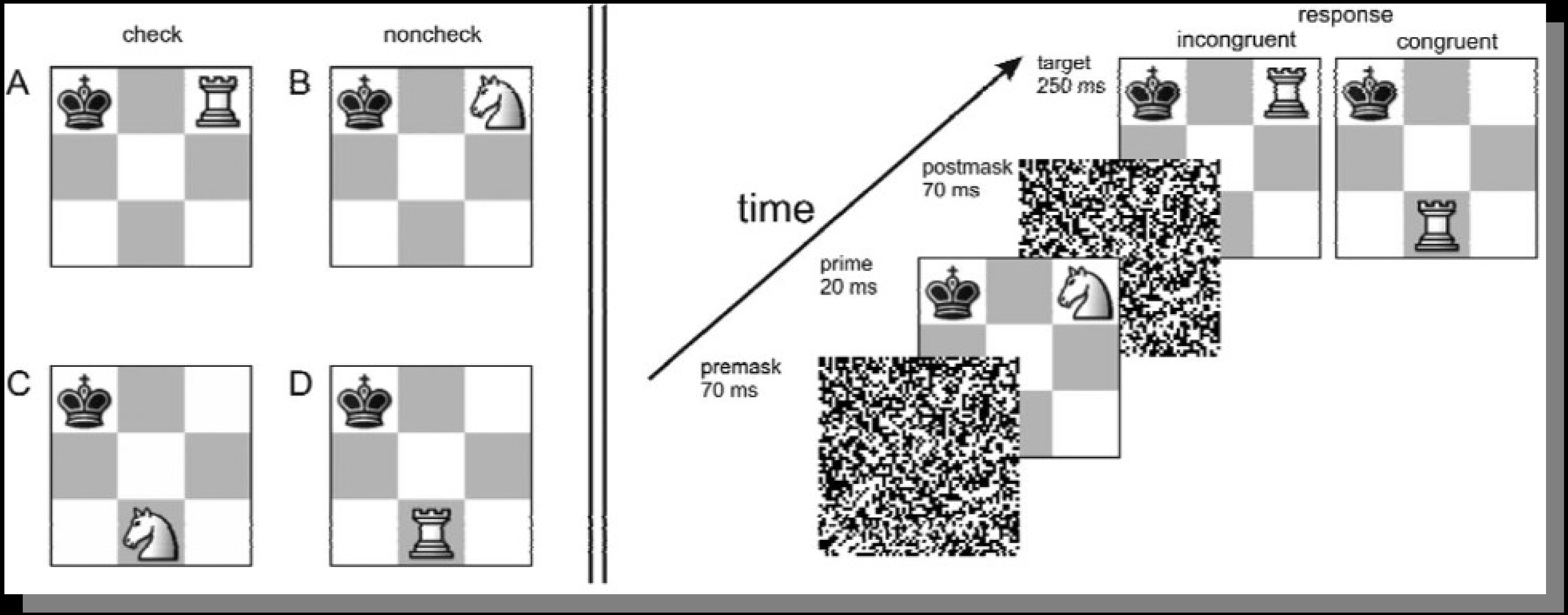
Expertise in Problem Solving

- Asked physics professors and undergrads to solve novice-level textbook problems
- "… aspect of novice problem solving is not only that they commit more errors than experts but that, even when they do solve a physics problem correctly, their approach is quite different …"
  - Experts solve equations in tightly-connected chunks and use far fewer diagrams

[M. Chi, R. Glaser and E. Rees. *Expertise in Problem Solving*. Advances in the Psychology of Human Intelligence, 1982.]

Playing Chess Unconsciously (1/2)

- Show experts and novices very brief ("subliminal priming") glimpses of relevant and irrelevant chess boards. Are you in check?



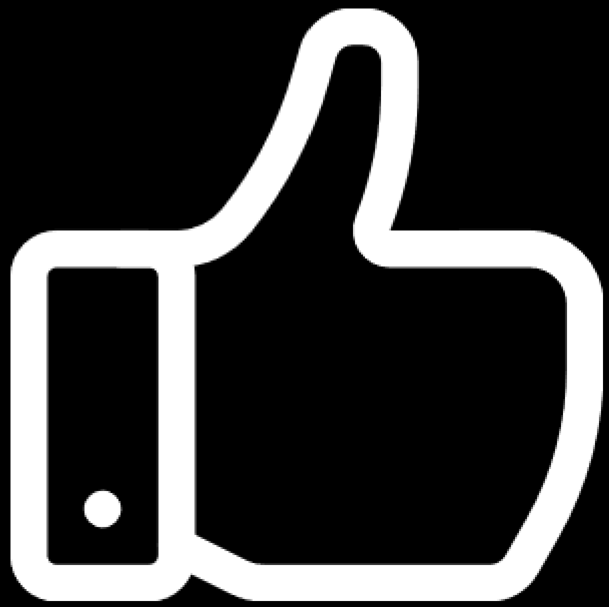
Playing Chess Unconsciously (2/2)

- Helps and hurts experts. No effect on novices.
- "practice might be an important prerequisite of subliminal processing (e.g., in the case of reading, letters are automatically integrated and form word "chunks") … experts' priming effects are brought about by acquired perceptual chunks that incorporate integrated features of chess pieces' identities and locations"

[A. Kiesel, W. Kunde, C. Pohl, M. Berner, and J. Hoffmann. *Playing Chess Unconsciously*. J. Experimental Psychology, 2009.] [W. Chase and H. Simon. *Perception in chess*. Cognitive Psychology, 1973.]

Audience Participation: Hypothesis Time

- When adults learn skills, do their brains physically change shape & organization ("new hardware")?
- When adults learn skills, do their memories change ("new software")?



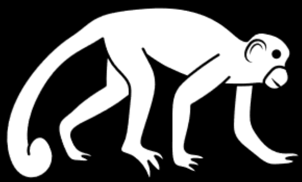
London Driver Study

- Use medical imaging to compare taxi drivers (memorizing 25,000 streets; some pass and others fails) and bus drivers over a few years
- "London taxi drivers not only have larger-than-average memory centers in their brains, but also that their intensive training is responsible for the growth"

[K. Woollett, H. Spiers and E. Maguire. *Talent in the taxi: a model system for exploring expertise*. Trans. Royal Society B, 2009.]

[U. Debarnot, M. Sperduti, F. Di Rienzo, and A. Guillot. *Experts bodies, experts minds: How physical and mental training shape the brain*. Frontiers in Human Neuroscience, 2014.]

Metabolic Efficiency



- Trained monkeys for 1-6 years on tasks
- "After extended practice, we observed a profound reduction of metabolic activity in M1 for the performance of internally generated compared to visually guided tasks. In contrast, measures of neuron firing displayed little difference during the two tasks. These findings suggest that the development of skill through extended practice results in a reduction in the synaptic activity" [N. Picard, Y. Matsuzaka, P. Strick. *Extended practice of a motor skill is associated with reduced metabolic activity in M1*. Nature Neuroscience. 2013.]

Age Effects on Brain Activity

- Older humans show more diffuse patterns of neural activity, recruiting other regions to help solve problems.
- "older adults showed weaker occipital activity and stronger prefrontal and parietal activity than younger adults … consistent with the view that sensory processing decline is a common cause in cognitive aging, and … reflect[ing] functional compensation."

[R. Cabeza, S. Daselaar, F. Dolcos, S. Prince, M. Budde and L. Nyberg. *Task-independent and Task-specific Age Effects on Brain Activity during Working Memory, Visual Attention and Episodic Retrieval*. Cerebral Cortex, 2004.]

The Case For Medical Imaging

- We want to improve SE performance
- Human brains have a major effect on SE
- We want to understand and influence humans
    - Humans *represent* and *perceive* knowledge differently with expertise
    - Humans process more *efficiently* with expertise
    - Human brains *physically* change with expertise
- How can we make novices more like experts?

Outline

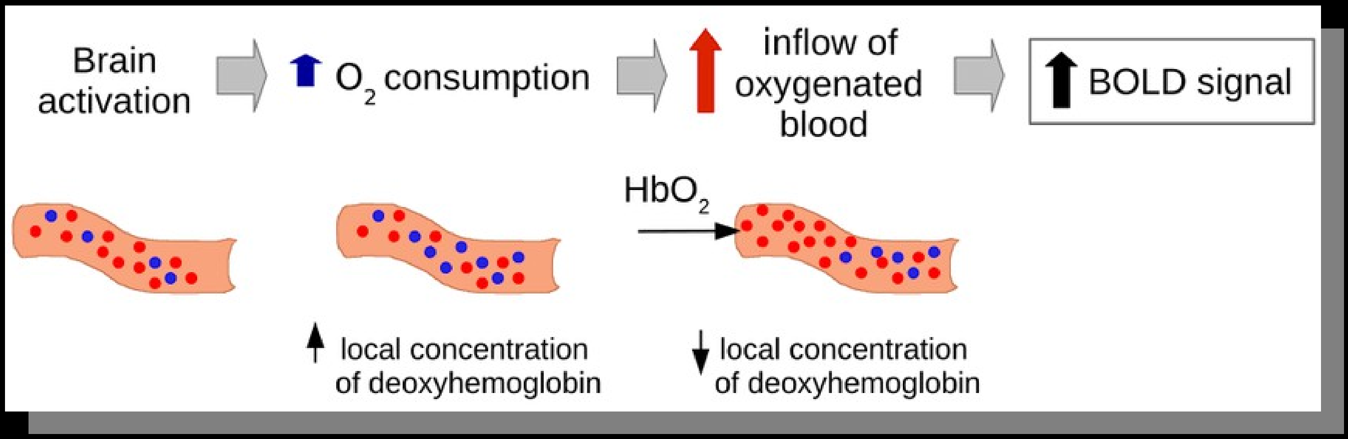
- Program Comprehension
- Selected Psychology Results
- Neuroimaging Explained
- Software Engineering + Medical Imaging
- Costs and Challenges
- Call to Arms

Medical Imaging Introduction

- Many here are familiar with eye tracking
- Other techniques: EEG, PET, fMRI, fNIRS, …
  - All are non-invasive, in vivo
- fMRI and fNIRS are gaining popularity
  - Earlier: sub-vocalization, etc. [C. Parnin. Subvocalization – toward hearing the inner thoughts of developers. ICPC 2011]
  - Now: allow sampling the whole brain (not just surface) rapidly (scale of seconds) with high spatial resolution (scale of millimeters)
- How do fMRI and fNIRS work?

Shadows on the Wall

- Your brain uses energy but does not store it
  - Thinking consumes oxygen from your blood
- We can track oxygen-rich vs. -poor blood
  - Blood Oxygen Level Dependent (BOLD) Signal



- 1,500+ fMRI pubs per year in Psychology
  - vs. 10 in last 5 years in Software Engineering

SE + Medical Imaging

- Survey recent papers that use such technologies to understand the brain
- Most focus on program comprehension
- Highlight one exciting result from each

Understanding Understanding: Is Reading Code Like Reading Prose?

- First fMRI study of program comprehension
- Studied small code snippets, finding five brain regions with distinct activation patterns, all of which are relevant to working memory, attention, and language processing
- Later: "evidence that data-flow-based code complexity metrics (but not control-flow-based metrics) rest on valid assumptions"

[J. Siegmund, C. Kästner, S. Apel, C. Parnin, A. Bethmann, T. Leich, G. Saake, A. Brechmann. *Understanding understanding source code with functional magnetic resonance imaging*. ICSE 2014] [N. Peitek, J. Siegmund, S. Apel, C. Kästner, C. Parnin, A. Bethmann, T. Leich, G. Saake, A. Brechmann. *A look into programmers' heads*. Trans. Software Engineering, 2018]

Semantic Cues and Beacons

- fMRI study of beacons and comprehension:
  - "found evidence of semantic chunking during bottom-up comprehension and lower activation of brain areas during comprehension based on semantic cues, confirming that beacons ease comprehension" [J. Siegmund, N. Peitek, C. Parnin, S. Apel, J. Hofmeister, C. Kästner, A. Begel, A. Bethmann, and A. Brechmann. Measuring Neural Efficiency of Program Comprehension. FSE 2017.]
- Simultaneous study: fMRI and eye tracking
  - Confirms semantic recall of programming plans by linking brain activation & eye fixation on beacons
    [N. Peitek, J. Siegmund, C. Parnin, S. Apel, J. Hofmeister, and A. Brechmann. Simultaneous Measurement of Program Comprehension with fMRI and Eye Tracking: A Case Study. ESEM 2018.]

Code Review and Expertise

- fMRI study comparing Github pull requests to code comprehension to prose revisions
  - “the neural representations of programming and natural languages are distinct. Our classifiers can distinguish between these tasks based solely on brain activity … expertise matters: greater skill accompanies a less-differentiated neural representation”
    - Expert brains treat programming languages more like natural languages

[B. Floyd, T. Santander, W. Weimer. *Decoding the representation of code in the brain: an fMRI study of code review and expertise*. ICSE 2017]

Bug Detection

- fMRI studies of finding bugs via code inspection, comprehension, decision making
  - “insula activity levels were critically related to the quality of error detection … Activity in this salience network (SN) region evoked by bug suspicion was predictive of bug detection precision, suggesting that it encodes the quality of the behavioral evidence”

[J. Castelhano, I. Duarte, C. Ferreira, J. Duraes, H. Madeira, M. Castelo-Branco. *The role of the insula in intuitive expert bug detection in computer code: an fMRI study.* Brain Imaging and Behavior, 2018.]

[J. Duraes, H. Madeira, J. Castelhano, C. Duarte and M. C. Branco. *WAP: Understanding the Brain at Software Debugging*. ISSRE 2016.]

Objective Comprehension Difficulty: Variables & Obfuscation

- NIRS studies of difficult code comprehension
  - "significant differences in brain activity were observed at a task that requires memorizing variables to understand a code snippet" [Y. Ikutani and H. Uwano. *Brain activity measurement during program comprehension with NIRS*. Software Engineering, Artificial Intelligence, Networking and Parallel / Distributed Computing, 2014.]
  - "showed high cerebral blood flow while understanding strongly obfuscated programs (requiring high mental workload)" [T. Nakagawa, Y. Kamei, H. Uwano, A. Monden, K. Matsumoto, D. German. *Quantifying programmers' mental workload during program comprehension based on cerebral blood flow measurement: A controlled experiment*. ICSE NIER, 2014.]

Subjective Comprehension Difficulty

- Study code comprehension in novices and professionals with eye tracking, EEG (and electrodermal)
  - "we could predict task difficulty and programmer level of expertise with 64.9 and 97.7% precision and 68.6 and 96.4% recall, respectively"
    [S. Lee, D. Hooshyar, H. Ji, K. Nam, H. Lim. *Mining biometric data to predict programmer expertise and task difficulty.* Cluster Computing 2017] [S. Lee, A. Matteson, D. Hooshyar, S. Kim, J. Jung, G. Nam, H. Lim. *Comparing programming language comprehension between novice and expert programmers using EEG analysis.* Bioinformatics and Bioengineering, 2016.]
  - Cheap biometrics predict expertise & difficulty
    [I. Crk, T. Kluthe, A. Stefik. *Understanding programming expertise: an empirical study of phasic brain wave changes.* CHI 2016] [T. Fritz, A. Begel, S. Müller, S. Yigit-Elliott, M. Züger. *Using psycho-physiological measures to assess task difficulty in software development.* ICSE 2014]

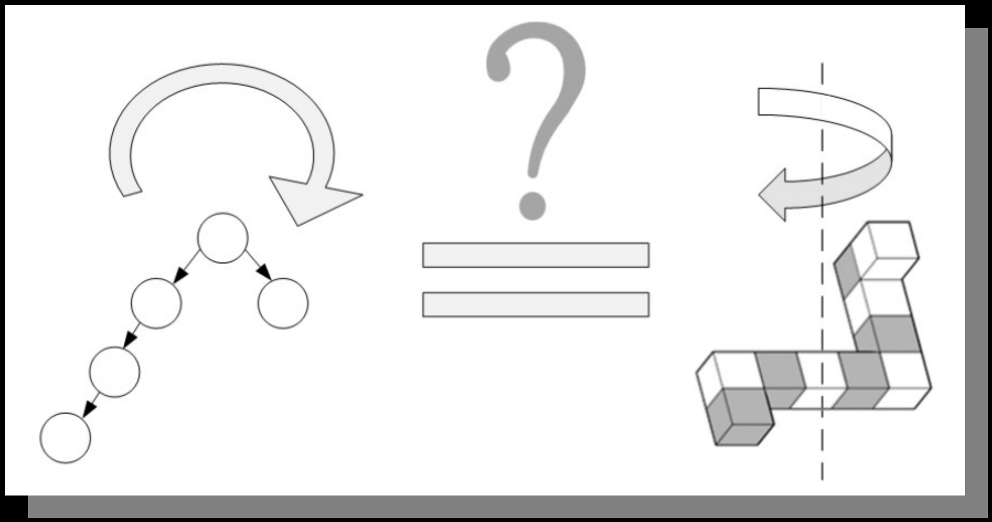
Readability and Cognitive Load

- fNIRS and eye tracking study of source code lexicon (e.g., identifier name) effects
  - “using fNIRS and eyetracking devices, developers’ cognitive load can be accurately associated with identifiers in source code and text”
  - “The existence of linguistic antipatterns in the source code significantly increases the cognitive load experienced by participants”

[S. Fakhoury, Y. Ma, V. Arnaoudova, O. Adesope. *The effect of poor source code lexicon and readability on developers’ cognitive load.* ICPC, 2018.]

Data Structures

- fMRI and fNIRS study of 76 participants
  - Data structure manipulations (rotating a tree) use the same parts of the brain as manipulating objects in the real world (spatial ability)



[Y. Huang, X. Liu, R. Krueger, T. Santander, X. Hu, K. Leach, W. Weimer. *Distilling Neural Representations of Data Structure Manipulation using fMRI and fNIRS*. ICSE 2019]

Costs and Challenges (1/3)

- All of that sounds great, but …

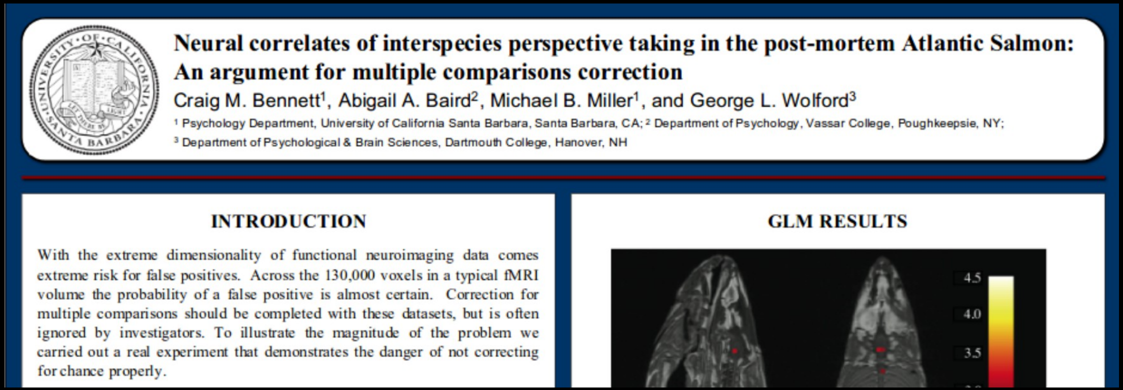
- IRB / Ethics Board
  - Brain Scans are HIPAA-protected data in the US
  - Must use full formal "medical" IRB

- Monetary Cost
  - fMRI can cost $600+ per participant

Costs and Challenges (2/3)

- Experimental Design
  - fMRI and fNIRS require a contrast-based controlled experimental setup. Informally:

    (Code_A + Breathing) – (Code_B + Breathing)
- Analysis and Modeling Expertise
  - Voluminous data require careful analysis to avoid spurious correlations and false discoveries



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

Costs and Challenges (3/3)

- Magnetic Interference
  - fMRI precludes the direct use of metal components, such as keyboards. Most studies involve a fixed set of pneumatic button presses.
- Hemodynamic Lag
  - BOLD signal limits: about 30 seconds per task
  - cf. dynamic AOI calculations for long-running eye-tracking studies

Call To Arms (1/2)

- We should propose, accept for publication, and evaluate new theories for programming and program comprehension
- "Recently, advances in cognitive neuroscience and brain imaging technology has provided new insight into the inner workings of the mind; unfortunately, theories such as program understanding have not been accordingly advanced." [C. Parnin. *A cognitive neuroscience perspective on memory for programming tasks*. Programming Interest Group 2010]

Call To Arms (2/2)

- For example, in psychology the heuristic-systematic model of information processing (HSM) formalizes how humans receive and process persuasive messages: humans try to save cognitive energy when reviewing (e.g., via sufficiency threshholds and heuristics).
  - We are now in a position to probe the details of this sort of model: “heuristics” (e.g., eye-tracking) and “energy” (e.g., medical imaging)

[S. Chen, K. Duckworth, S. Chaiken. *Motivated Heuristic and Systematic Processing*. Psychological Inquiry, 1989.]

Example Underexplored Areas

- Experts and novices cluster physics problems differently: is that true for SE?
- Does expert perceptual chunking, shown via chess endgame priming, also apply to SE?
- Do we find neural reorganization with training in SE, similar to findings for taxi drivers?
- Are metabolic efficiencies in SE experts more like those found in internally-generated actions or those found in visually-guided tasks?

Conclusion

- Generality: Why does this feature make comprehension easier or harder?
- Pedagogy: Will practicing this task help with that task?
- Training: What happens as novices become experts and the young become older?
- Support: What are the perceptual units of code?
- Medical imaging experiments
  - *require* resources and careful design
  - *provide* objective answers to subjective queries

Bonus Slides

Importance of Program Comprehension

- "Understanding code is by far the activity at which professional developers spend the most time." [Hallam. *What do programmers really do anyway?* Microsoft Tech Report.]
- Understanding is more important than functional correctness when choosing software. [NASA Software Reuse Working Group. *Software reuse survey.* NASA Tech Report.]
- Many companies mandate code review
  - Google Mondrian, Facebook Phabricator, etc.

View From The Gallery

- What could explain those individual differences?
- Other fields focus research on such areas
- Provide a high-level survey of a few interesting psychology results curated for relevance to program comprehension
  - Typically “classic” or “non-controversial” works
  - Motivates the study of the brain for SE (e.g., via medical imaging or psychology techniques)

Training Shapes The Brain

- Experiments with motor skills, imagery skills, and purely mental tasks …
- "The neural reorganizations that occur with expertise reflect the optimization of the neurocognitive resources to deal with the complex computational load … the disparity between the quality of the performance of novice and expert golfers lies at the level of the functional organization of neural networks during motor planning." [U. Debarnot, M. Sperduti, F. Di Rienzo, and A. Guillot. *Experts bodies, experts minds: How physical and mental training shape the brain.* Frontiers in Human Neuroscience, 2014.]