Data mining in neuroimaging
Some whys and one how

Tal Yarkoni
University of Texas at Austin
Overview

1. Panic, panic, everywhere!
2. Just kidding
3. Neurosynth demo
4. Neurosynth demo

What is data mining?

- “...the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.” --wiki
dedia
- Much of what you’ve learned about already falls under data mining
- Often contrasted with hypothesis-driven research
- Until recently, data mining had a bad reputation in psychology/neuroscience
  - “Going on a fishing expedition...”
Why?

• It’s fun!
• Neuroimaging data is expensive; we should maximize its utility
• You can learn interesting and unexpected new things
  • Bottom-up and top-down approaches are complementary
• It’s probably wishful thinking to think you can avoid it

The promise of fMRI

Fig. 1, Canli et al (2001)
Look at those scatterplots!

Extraversion  

Neuroticism

Canli et al (2001)

In 2001...

- The neural correlates of personality (and many other things) seemed:
  - Highly localized
  - Extremely strong
  - Theoretically predictable
- The answers were just around the corner
What happened?

- Have these findings been replicated?
    - Some are conceptual replications
    - Localization varies
    - Measures vary
    - Some contradictory findings
    - Many unreported non-replications

- Depends on how you define replication
  - Arguably no finding in personality neuroscience has been replicated (Yarkoni, 2014)

- Why?

Power is often very low
A simulation

**Real effects in population**

The effects of sampling...

**Observed effects in sample**

<table>
<thead>
<tr>
<th>N = 20</th>
<th>N = 40</th>
<th>N = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR = .05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p &lt; .001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p &lt; .05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The (hypothetical) truth

Effect size ($r$)

Voxel #

Population $r = 0.3$

$r = 0.3, N = 20, p < 0.001$
$r = 0.3 \ , \ N = 20 \ , \ p < 0.001$

Effect size ($r$) vs. Voxel #

Threshold $r = 0.66$

Population $r = 0.3$
A pretty good projection of reality
A less good projection
Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button1,2, John P. A. Ioannidis3, Claire Mokrysz1, Brian A. Nosek4, Jonathan Flint6, Emma S. J. Robinson2 and Marcus R. Munafò1

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

Big Data helps

- Increased power
- Better effect size estimation
- Helps us separate truly large effects from small effects that only seem large
- In practice, an interesting thing happens...
Increasing sensitivity through meta-analysis

Wager, Lindquist, & Kaplan (2007)

Kober et al. (2008)
Meta-analyze all the things!

- **Kim et al. (2011)** - Task-switching
- **Kim (2011)** - Subsequent memory
- **Engelmann et al. (2011)** - Smoking cue reactivity
- **Molenberghs et al. (2011)** - Mirror system
- **Chase et al. (2011)** - Drug craving
- **Tillisch et al. (2011)** - Rectal distension
- **Rehme et al. (2012)** - Movement after stroke
- **Fan et al. (2011)** - Empathy
- **Rehme et al. (2012)** - Movement after stroke
- **Brooks et al. (2011)** - Subliminal arousing stimuli
- **Swick et al. (2011)** - Go/NoGo & Stop-Signal
Dude, where’s my effect?

- A funny thing: as samples grow, effects seem to shrink
- Vul et al (2009b) reported a correlation of -0.62 between sample size and effect size
- This is the raison d’être of the funnel plot
- The largest fMRI studies invariably report much smaller effects
  - Predict binge drinking at age 16 from an enormous amount of data at age 14
  - History, demographics, personality, cognitive ability, fMRI activity, genetics, etc.
  - “At the optimum point in the AUC curve, 73% of non-binge drinkers and 66% of binge drinkers were correctly classified, significantly better than chance (P = 4.2 × 10−17) given a base rate of 45% binge drinkers.”

![Graph showing correlation coefficients](image-url)
Voxelwise genome-wide association study (vGWAS)


A key goal in imaging neuroscience is to discover specific effects of center, as detected by one-way ANOVA (across 24 centers, including 1,093 participants). Columns and rows depict voxels exhibiting age- and sex-related variations. Center, age, and sex were included in the analysis model. We then calculated the correlation between age and functional connectivity strength as a function of sample size. For each of the two regions illustrating the effect of age for the PCC see Fig. 2 that showed a negative correlation between age and RSFC with the PCC seed. In each voxel, we developed a novel method to address the multiple comparisons problem and computational burden associated with the unprecedented amount of data. No variant survived the strict significance criterion, but several genes worthy of further exploration were identified, including CSMD2 and CDY1B. These genes have high relevance to brain structure. This is the first voxelwise genome wide association study to our knowledge, and offers a novel method to discover genetic influences on brain structure.

Article info

Article history:
Received 31 August 2009
Revised 21 January 2010
Accepted 11 February 2010
Available online 17 February 2010

Abstract

The structure of the human brain is highly heritable, and is thought to be influenced by many common genetic variants, many of which are currently unknown. Recent advances in neuroimaging and genetics have allowed collection of both highly detailed structural brain scans and genome-wide genotype information. This wealth of information presents a new opportunity to find the genes influencing brain structure. Here we explore the relation between 448,293 single nucleotide polymorphisms in each of 31,622 voxels of the entire brain across 740 elderly subjects (mean age ± s.d.: 75.52 ± 6.82 years; 438 male) including subjects with Alzheimer’s disease, Mild Cognitive Impairment, and healthy elderly controls from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). We used tensor-based morphometry to measure individual differences in brain structure at the voxel level relative to a study-specific template based on healthy elderly subjects. We then conducted a genome-wide association at each voxel to identify genetic variants of interest. By studying only the most associated variant at each voxel, we developed a novel method to address the multiple comparisons problem and computational burden associated with the unprecedented amount of data. No variant survived the strict significance criterion, but several genes worthy of further exploration were identified, including CSMD2 and CDY1B. These genes have high relevance to brain structure. This is the first voxelwise genome wide association study to our knowledge, and offers a novel method to discover genetic influences on brain structure.
This does not mean...

- ...that all fMRI research is grossly underpowered
- Many questions can be answered with 4, 14, or 40 subjects
- It depends on the kind of question
- But in most cases, the prior probability of huge effects is not high

Two possible interpretations

- The largest studies to date are kind of crappy; the small studies are better
- We’re engaging in a collective exercise in wishful thinking
Simple minds vs. muddle heads

“The two opposite errors to which psychologists, especially clinical psychologists, are tempted are the simpleminded and the muddleheaded … The simpleminded, due to their hypercriticality and superscientism and their acceptance of a variant of operationalist philosophy of science … tend to have a difficult time discovering anything interesting or exciting about the mind. The muddleheads, per contra, have a tendency to discover a lot of interesting things that are not so. I have never been able, despite my Minnesota “simpleminded” training, to decide between these two evils.”

—Paul Meehl

On fooling one’s self
Surely, you jest...

- Could we really be so wrong, collectively, about these things?
- Yes
- Some examples...

Deja vu all over again
Meta-analysis of genome-wide association studies for personality


1Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands; *National Institute on Aging, NIH, Baltimore, MD, USA; **Department of Psychology, University of Minnesota, Minneapolis, MN, USA; †Department of Psychiatry, University Medical Center Groningen, Groningen, The Netherlands; ‡Departments of Clinical Psychology and Psychiatry, Leiden University, Leiden, The Netherlands; §Department of Psychiatry, EMGO + Institute, Neuroscience Campus Amsterdam, VU University Medical Center Amsterdam, Amsterdam, The Netherlands; ‡Department of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; ‡Estonian Biocentre, Tartu, Estonia; ‡Estonian Genome Project of University of Tartu, Tartu, Estonia; ‡Department of Psychiatry, Washington University School of Medicine, St Louis, MO, USA; ‡Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands; ‡Istituto di Neurogenetica e Neurofarmacologia, CNR, Monserrato, Cagliari, Italy; ‡Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; ‡Department of Psychology, University of Tartu, Tartu, Estonia; ‡Institute of Genetics and Biophysics, A Buzzati-Traverso—CNR, Naples, Italy; ‡Department of Psychiatry, University of Munich (LMU), Munich, Germany; ‡Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; ‡National Institute for Health and Welfare, Helsinki, Finland; ‡Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland; ‡Vasa Central Hospital, Vasa, Finland; ‡Unit of General Practice, Helsinki University Central Hospital, Helsinki, Finland; ‡Folkhälässan Research Centre, Helsinki, Finland; ‡Wellcome Trust Sanger Institute, Cambridge, UK; ‡Broad Institute of Harvard and MIT, Cambridge, MA, USA; ‡Department of Medical Genetics, University of Helsinki and Helsinki University Hospital, Helsinki, Finland; ‡Department of Psychology, Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK; ‡MRC Human Genetics Unit, The Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, UK; ‡Queensland Institute of Medical Research, Brisbane, QLD, Australia; ‡Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands; ‡Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA and ‡Department of Psychology, University of Helsinki, University of Helsinki, Helsinki, Finland

Personality can be thought of as a set of characteristics that influence people's thoughts, feelings and behavior across a variety of settings. Variation in personality is predictive of many outcomes in life, including mental health. Here we report on a meta-analysis of genome-wide association studies for Conscientiousness. In silico replication did not, however, show significant associations of the top SNPs with Openness and Conscientiousness, although the direction of effect of the KATNAL2 SNP on Conscientiousness was consistent in all replication samples. Larger scale GWA studies and alternative approaches are required for confirmation of KATNAL2 as a novel gene affecting Conscientiousness.
Hypothesis-driven tests in neuroimaging

• Paradigmatic case is focusing on a single ROI for theoretical reasons
• This is often held up as an example of the right way to do science
• Is it?
  • How many ROIs were actually tested?
  • Were the ROIs truly a priori?
  • Correction for multiple comparisons?
• How would you know?
Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling

Leslie K. John¹, George Loewenstein², and Drazen Prelec³
¹Marketing Unit, Harvard Business School; ²Department of Social & Decision Sciences, Carnegie Mellon University; and ³ Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology

Abstract

Cases of clear scientific misconduct have received significant media attention recently, but less flagrantly questionable research practices may be more prevalent and, ultimately, more damaging to the academic enterprise. Using an anonymous elicitation format supplemented by incentives for honest reporting, we surveyed over 2,000 psychologists about their involvement in questionable research practices. The impact of truth-telling incentives on self-admissions of questionable research practices was positive, and this impact was greater for practices that respondents judged to be less defensible. Combining three different estimation methods, we found that the percentage of respondents who have engaged in questionable practices was surprisingly high. This finding suggests that some questionable practices may constitute the prevailing research norm.

Keywords

professional standards, judgment, disclosure, methodology

Received 5/20/11; Revision accepted 10/20/11
The first principle

One example: Millikan measured the charge on an electron by an experiment with falling oil drops, and got an answer which we now know not to be quite right. It’s a little bit off, because he had the incorrect value for the viscosity of air. It’s interesting to look at the history of measurements of the charge of the electron, after Millikan. If you plot them as a function of time, you find that one is a little bigger than Millikan’s, and the next one’s a little bit bigger than that, and the next one’s a little bit bigger than that, until finally they settle down to a number which is higher.

Why didn’t they discover that the new number was higher right away? It’s a thing that scientists are ashamed of--this history--because it’s apparent that people did things like this: When they got a number that was too high above Millikan’s, they thought something must be wrong--and they would look for and find a reason why something might be wrong. When they got a number closer to Millikan’s value they didn’t look so hard. And so they eliminated the numbers that were too far off, and did other things like that. ...

The first principle is that you must not fool yourself--and you are the easiest person to fool. So you have to be very careful about that. After you’ve not fooled yourself, it’s easy not to fool other scientists. You just have to be honest in a conventional way after that.

--Richard Feynman, Cargo Cult Science

You are the easiest person to fool

“ For example, the maximal correlation between right superior PFC activation and choice performance was $r = -.75$. The magnitude of this correlation may surprise those unfamiliar with recent imaging studies that find similarly strong correlations between isolated brain activity and overt behavior across a wide variety of contexts ... We would suggest that given the magnitude of the correlations obtained, it is not unreasonable to entertain the possibility that complex cognitive functions are localized to specific regions.”

Yarkoni et al. (2005), p. 552
Big Data helps

- It’s more difficult to fool yourself when you have large N’s
- Fudging tends to take place on the margins
  - With enough data, there are no margins
- Large datasets are more likely to be public, transparent (e.g., fc1000)
- Not a panacea, of course

The importance of being specific
The problem of reverse inference

- Reasoning from brain activity to mental states is hard (Poldrack, 2006)

- $P(\text{State}|\text{Activation})$ is not the same as $P(\text{Activation}|\text{State})$

- Saying activation X implies state S1 is simultaneously saying that X implies the absence of states S2, S3, S4 ...

- But how would we know?

- No individual study, no matter how big, can tell us
Meta-analysis doesn’t help

- Molenberghs et al. (2011) - Mirror system
- Engelmann et al. (2011) - Smoking cue reactivity
- Smiling et al. (2011) - Subsequent memory

- Chase et al. (2011) - Drug craving
- Fan et al. (2011) - Empathy
- Swick et al. (2011) - Go/NoGo & Stop-Signal

- Rehme et al. (2012) - Movement after stroke
- Brooks et al. (2011) - Subliminal arousing stimuli
All regions are not equal(ly active)

E.g., anterior insula is "selectively" activated by...

- Disgust
- Pain
- Empathy
- Risk
- Interoceptive awareness
- Autonomic arousal
- Self-reflection
- Task-switching
- Conscious error perception
- Response inhibition
- Speech production
- Sustained attention
- Etc...
Conventional meta-analysis is not enough

- We need comprehensive coverage
- Must directly compare studies of one task with many others
- Not easy to do manually
- Large-scale horizontal data mining is required

The good news...

- New resources and methods are rapidly emerging that can help us overcome these problems
- Massive datasets spanning thousands of subjects and/or many tasks
- New multivariate approaches that improve sensitivity and/or specificity
500 Subjects Data Release

Announcing a new data release by the WU-Minn HCP Consortium.

The WU-Minn Human Connectome Project Consortium (WU-Minn HCP) is pleased to release imaging and behavioral data for more than 500 subjects. Registration is required.

How to Download HCP Public Data Releases

The Human Connectome Project

Mapping the human brain is one of the great scientific challenges of the 21st century. The Human Connectome Project (HCP) is tackling a key aspect of this challenge by elucidating the neural pathways that underlie brain function and behavior. Deciphering this complex wiring diagram will reveal much about what makes us uniquely human and what makes every person different from all others.

The consortium led by Washington University, University of Minnesota, and Oxford University (the WU-Minn HCP consortium) is comprehensively mapping human brain circuits in a target number of 1200 healthy adults using cutting-edge methods of noninvasive neuroimaging. It will yield invaluable information about brain connectivity, its relationship to behavior, and the contributions of genetic and environmental factors to individual differences in brain circuitry and behavior.

Starting with the first quarterly (Q1) data release (March, 2013), HCP datasets are being made freely available to the scientific community. Four imaging modalities are used to acquire data with unprecedented resolution in space and time. Resting-state functional MRI (fMRI) and diffusion imaging (dMRI) provide information about brain connectivity. Task-based fMRI provides

1000 Functional Connectomes Project

INDI

International Neuroimaging Data-Sharing Initiative

Learn more  Request Access

How-To’s  Prospective

FCP Scripts  Retrospective

Forums  Summer of Sharing

Share that brain!
### Data Sets

**ADD A DATASET**

<table>
<thead>
<tr>
<th>Study</th>
<th>Investigators</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balloon Analog Risk-Tasking Task</td>
<td>Tim Schonberg</td>
<td>0001001</td>
</tr>
<tr>
<td>Classification learning</td>
<td>Arun, A.R.</td>
<td>0001002</td>
</tr>
<tr>
<td>Rhyme judgment</td>
<td>Xue, G.</td>
<td>0001003</td>
</tr>
<tr>
<td>Mixed-gambit task</td>
<td>Tim S.M.</td>
<td>0001005</td>
</tr>
<tr>
<td>Living-naming declar with plans or</td>
<td>Kimura</td>
<td>0001006</td>
</tr>
<tr>
<td>motor neuron test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stop-signal task with spoken &amp;</td>
<td>Xue, G.</td>
<td>0001007</td>
</tr>
<tr>
<td>manual responses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stop-signal task with unconditioned</td>
<td>Arun, A.R.</td>
<td>0001009</td>
</tr>
<tr>
<td>and conditional stopping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The generality of self-control</td>
<td>Jessica Cohen</td>
<td>0001009</td>
</tr>
<tr>
<td>Classification learning and tone-sensing</td>
<td>Feeney, K.</td>
<td>00010011</td>
</tr>
<tr>
<td>Classification learning and stop-signal (1 year test-retest)</td>
<td>Hsu-Jackson</td>
<td>0001017</td>
</tr>
<tr>
<td>Classification learning and stop-signal (1 year test-retest)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-language repetition priming</td>
<td>Alzare, R.</td>
<td>00010061</td>
</tr>
<tr>
<td>Classification learning and reversal</td>
<td>Poldreck, R.A.</td>
<td>0001052</td>
</tr>
</tbody>
</table>
One particular approach

- There are many resources for large-scale fMRI data mining
- You’ve already learned about several in this course
- But since I’m giving this particular talk...
Large-scale automated synthesis of the literature

• The data we need to establish many brain-mind mappings already exists
• > 20,000 fMRI articles have been published
• Can we synthesize this literature in an automated way?

Two problems

• Data standardization
• Semantic annotation

• Our approach: ignore both problems
Assumption 1: If it looks like a duck...

- Anything that looks like brain activation is brain activation
- Parser looks for x/y/z-like numbers in sequence within HTML tables

Assumption 2: The brain is a bag of words

- Each published article is an unordered list of words
- A study is about the words it uses with high frequency
What these assumptions buy us


Current state of the database

- 9,700 studies drawn from 50+ journals
- > 300,000 reported activations
- Represents almost any psychological state that can be indexed with words
- Considerable potential for growth
- Database is completely open
Mining for associations between text and brain activation in a functional neuroimaging database

Finn Årup Nielsen¹,², Lars Kai Hansen², Daniela Balslev¹,³

Pain vs. Pain

Blue: manual  Green: automated
Quantitative reverse inference
Classification of cognitive states

Decoding mental states in single subjects

- Can we identify cognitive states in individual subjects?
- Test ability to classify working memory, emotion, and pain

Data-driven insula parcellation

Chang, Yarkoni, Khaw & Sanfey (2013)
A tale of two inferences

![Diagram showing Forward Inference and Reverse Inference with various categories such as Emotion, Gustation, Face, Anxiety, Olfaction, Switching, Inhibition, Error Processing, Conflict, Feedback, Pain, Somatosensory, Motor, Music, Auditory, and Insula Region with Ventral, Dorsal, and Posterior regions highlighted.]

Chang, Yarkoni, Khaw & Sanfey (2013)

---

Everything is open

- [http://github.com/neurosynth/](http://github.com/neurosynth/)
- Available packages:
  - ACE: Automated coordinate extraction and very basic text mining (Python)
  - Core tools: Large-scale synthesis of fMRI data (Python)
  - Neurosynth viewer: JavaScript library for interactive 2D visualization (CoffeeScript)
- Everything on the website was generated using these tools
- Focus is on simple, large-scale analysis and ease of use
neurosynth.org

- A web interface that provides access to many of the data, tools, and results generated with Neurosynth
- beta.neurosynth.org: because every beta website needs its own beta website
- Has an API, so other applications can easily use Neurosynth data

Work in progress
Neurosynth has all kinds of limitations

• The data are very low quality
  • No concept of contrasts, deactivations, etc.
  • No ontology—every article is a bag of words

• No centralization of data

• Arguably the wrong kind of data

• Etc…

Improving data quality and annotation

• Revamped coordinate extraction package

• Text-mining is more than just word counting
  • E.g., “topic mapping” (Poldrack et al, 2012)

• Crowdsourcing
  • Can’t eliminate human effort
  • Can we build elegant interfaces that make contributing easy?
  • E.g., Roberto Toro’s brainspell.org
Towards images instead of coordinates

• We’re still relying on discrete coordinates
• We would prefer to operate on images (e.g., Salimi-Khorshidi et al., 2009)
• Can we incentivize people to share?

NeuroVault

• neurovault.org
• Gorgolewski et al.
• A public repository of whole-brain images
• Why use it?
  • Permanent URL w/ free storage
  • Interactive visualization
  • Easy sharing
  • Interfaces with other tools
Enables new applications

- NeuroVault has an open API
- Any image can be downloaded
- Opens the door to new applications
- For example...

Soon (hopefully)…

- Perform real-time meta-analyses in-browser
- Instantly share the results interactively
- Evaluate all details of others’ meta-analyses
- Annotate anything: voxels, contrasts, studies, etc.
Conclusions

• Large-scale data mining approaches are essential
  • They are the only way to solve many inferential problems
• The choice between more vs. “right” data is not an obvious one
• We live in a world full of enormous, rich, open datasets—use them!

Thanks to...

+ many others...
Neurosynth is hiring!