Data science in neuroscience: Generating insight from rich, complex and messy data

Tuesday, Jun 19: 8:00 AM - 9:15 AM
1485
Symposium
Tuesday - Symposia AM

Following astronomy, particle physics, and genetics, massive data collection is currently becoming a game changer in neuroscience and medicine (House of Common, UK, 2016; National Research Council, USA, 2013). There is an always-larger interest in and pressure for data sharing, open access, and consortiums that build "big data" repositories for the healthy and diseased brain. For instance, UK Biobank is a longitudinal population study dedicated to the genetic and environmental influence on mental and other disorders. 500,000 enrolled volunteers have undergone an extensive battery of clinical diagnostics from brain scans to bone density with a >25 year follow-up. In the US, the Precision Medicine Initiative announced in 2015 to even profile 1,000,000 individuals. There is now an unprecedented, rapidly growing opportunity to provide principled answers to human brain function and its disturbances in mental disease.

What is currently changing is the questions that can be asked about a given brain phenomenon quantified in data. Indeed, the more brain recordings are available, the more can be learned about the brain given adequate statistical models. However, the more variables per observation are to be analyzed, the bigger the modeling challenges become at the statistical and computational level. This will require a symbiotic interplay between neuroscientific reasoning styles and statistical reasoning styles (Abbott, 2016; Goodman, 2016). Successful exploitation of the data wealth will require a new generation of computationally and statistically trained neuroscientists with a shift in data analysis practices (McKinsey Global Institute, 2011 and 2016). Indeed, brain sciences have been identified as the most data-rich among all medical specialties, with a dominant contribution from medical imaging techniques such as functional MRI (Nature Editorial, 2016). Due to the complexity of the patterns that need to be detected in neurobiology, a human cannot provide explicit, fine-detailed brain mechanisms. Instead, data science algorithms have the potential to become widespread tools that distill information from large data sets to turn unstructured data accumulation into structured knowledge.

Objective

1) What are the challenges and opportunities of growing datasets with rich meta-information?
2) What scientific questions and statistical approaches lend themselves to the data-rich setting?

Target Audience

The target broad target audience ranges from early PhD students to senior investigators intending to work in the 'big-data' setting.

Co Organizer

Danilo Bzdok, Department of Psychiatry

Organizer

Bertrand thirion, Parietal Team, INRIA/Neurospin Saclay

Presentations

Learning from neuroimaging and clinical data: a multiple-source machine learning approach for mental health disorders (index.cfm?do=ev.viewEv&ev=1636)

Over the last decade machine learning techniques have been successfully applied to clinical neuroimaging data leading to a growing body of research focused on diagnosis and prognosis of mental health disorders (Kloppel et al. 2012). However, most of studies combining neuroimaging and machine learning techniques have focused on binary classification problems using a single neuroimaging modality, i.e. they summarize the clinical assessment into a single measure (e.g. diagnostic classification) and the output of the models is limited in most cases to a binary decision (e.g. healthy/patient). Considering that current psychiatric categories fail to align with findings from clinical neuroscience and genetics and have not been predictive of treatment response (Insel et al., 2010), there is clearly a limitation to what these models can provide in terms of improving diagnosis and prognosis of mental health disorders. With the technological advances enabling acquisition of large volumes of patient data, new machine learning models that can combine information from neuroimaging techniques with complementary knowledge from clinical assessments and general patient information have the potential to identify reliable biological markers and improve patients' characterization in psychiatry. In this talk I discuss how multiple source machine learning models (such as Partial Least Square and Canonical Correlation Analysis) can be used to learn a latent space capturing associations between multiple sources of data which has the potential to identify subgroups of patients based on multivariate brain-behaviour associations and people at high/low risk of developing mental health disorders (Monteiro et al, 2016). Monteiro JM, Rao A, Shawe-Taylor J, Mourão-Miranda J; Alzheimer's Disease Initiative. A multiple hold-out framework for Sparse Partial Least Squares. J Neurosci Methods. 2016 Klöppel S, Abdulkadir A, Jack CR Jr, Koutsouleris N, Mourão-Miranda J, Vemuri P. Diagnostic neuroimaging across diseases. Neuroimage. 2012 Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang, P. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry. 2010

Presenter

Janaina Mourao-Miranda, University College London

Connectome Coding (index.cfm?do=ev.viewEv&ev=1637)

Neural coding is the discipline devoted to understanding the relationship between the ongoing external environment (including stimuli and behavior) and ongoing internal neural activity (focusing on neural spike times). Connectome coding is a discipline on the frontiers of brain sciences, devoted to understanding the relationship between the past environment (including genetics and experiences) and the current neural anatomy (focusing on synaptic connections). In other words, connectome coding is about characterizing the structure-function relationship, for example, predicting from the structure of the brain how it functions, or vice versa. Specifically, connectome coding is a requisite stepping stone towards a mechanistic understanding of normal and abnormal behavior and cognition, as well as neurological and psychiatric disorders, learning, memory, and personality traits in terms of synaptic connectivity. Kiar, G., Bridgeford, E., Chandrashekhar, V., Mhembere, D., Burns, R., Roncal, W. G., & Vogelstein, J. (2017). A Comprehensive Cloud Framework for Accurate and Reliable Human Connectome Estimation and Meganalysis. bioRxiv, 188706. Athreya, A., Fishkind, D. E., Levin, K., Lyzinski, V., Park, Y., Qin, Y., ... & Vogelstein, J., Priebe, C. E. (2017). Statistical inference on random dot product graphs: a survey. arXiv preprint arXiv:1709.05454.

Presenter

Joshua Vogelstein, John Hopkins University

Learning from heterogeneous data to increase sample size: Why neuroscientists should care (index.cfm?do=ev.viewEv&ev=1638)

Whatever some people have told you “the more data you have the better” it is to draw strong scientific conclusions. However more data is not always easy: samples are expensive, number of cases is limited in a given cohort etc. To address this issue, it is a natural to pool data from different studies, cohorts, scanners etc. One problem that emerges when doing so is that the data easily end up with different statistical properties. The technical term for this is “distribution shift”: it can be a covariate shift, a variations in noise levels etc. In this talk I will give an overview of statistical techniques, such as domain adaptation, transfer learning or multi-task learning, that have been proposed to deal with such issues. I will illustrate this using MEG and EEG data, including clinically relevant problems such as the problem of steep stage classification. Massias, M., Fercoq, O., Gramfort, A., & Salmon, J. (2017). Heteroscedastic Concomitant Lasso for sparse multimodal electromagnetic brain imaging. arXiv preprint arXiv:1705.09778. Chambon, S., Galtier, M., Arnal, P., Wainrib, G., & Gramfort, A. (2017). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. arXiv preprint arXiv:1707.03321.

Presenter

Alexandre Gramfort, Inria

Population neuroscience meets computer-age statistics: Extending what we can learn about the brain? (index.cfm?do=ev.viewEv&ev=1639)

Neuroimaging datasets are constantly growing in resolution, sample size, multi-modality, and meta-information complexity. The changing data reality may open the brain imaging field to a more data-guided machine-learning regime (e.g., structured sparsity, predictive phenotype modeling, multi-output learning). However, in everyday research practice analysis methods from the domain of classical statistics remain ubiquitous (e.g., ANOVA, Pearson correlation, Student's t-test). This talk will highlight key conceptual differences between neuroscientific knowledge generation based on traditional null-hypothesis testing and cross-validation. Alternative access to answer long-standing systems neuroscience questions will then be portrayed in three large-scale studies: 1) Translating structured sparsity penalization into brain imaging to jointly identify compact brain region and distributed network patterns in high-dimensional regression. 2) Combining latent factor discovery based on autoencoder architectures and statistical gains from multi-class transfer-learning to reveal predictive dimensions from heterogeneous data sources. 3) Exploiting sparse canonical correlation analysis to simultaneously explore hidden inter-individual differences of brain-behavior variation in high-level cognition. These examples in the Human Connectome Project, the UK Biobank, and other extensive brain-imaging repositories will illustrate how the currently increasing information granularity may shape our future data analysis practices. Bzdok, D., Krzywinski, M., Altman, N., 2017. Machine learning: A primer. Nature Methods. Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: Opportunities and challenges. Biological Psychiatry: CNNI. Bzdok, D., Yeo, B.T.T., 2017. Inference in the age of big data: Future perspectives on neuroscience. Neuroimage.

Presenter

Danilo Bzdok, Department of Psychiatry