AFQ-Browser: Supporting reproducible human neuroscience research through browser-based visualization tools

Jason D. Yeatman[1,2*], Adam Richie-Halford[3], Josh K. Smith[4], Ariel Rokem[5*]

[1]Institute for Learning & Brain Sciences, [2]Department of Speech and Hearing Sciences, [3]Department of Physics, [4]Department of Chemical Engineering, [5]eScience Institute, University of Washington, Seattle, WA.

* Corresponding Authors:

Jason, D. Yeatman, PhD
Institute for Learning & Brain Sciences
Portage Bay Building, Box 357988
University of Washington
Seattle, WA 98195, USA
E-mail: jyeatman@uw.edu

Ariel Rokem, PhD
eScience Institute
WRF Data Science Studio
Physics/Astronomy Tower (PAT), 6th Floor
3910 15th Ave NE
University of Washington
Seattle, WA 98195, USA
E-mail: arokem@gmail.com

Abstract

Human neuroscience research faces several challenges with regards to reproducibility. While scientists are generally aware that data sharing is an important component of reproducible research, it is not always clear how to usefully share data in a manner that allows other labs to understand and reproduce published findings. Here we describe a new tool, AFQ-Browser, that builds an interactive website as a companion to a published diffusion MRI study. Because AFQ-browser is portable -- it runs in any modern web-browser -- it can facilitate transparency and data sharing. Moreover, by leveraging new web-visualization technologies to create linked views between different dimensions of a diffusion MRI dataset (anatomy, quantitative diffusion metrics, subject metadata), AFQ-Browser facilitates exploratory data analysis, fueling new scientific discoveries based on previously published datasets. In an era where Big Data is playing an increasingly prominent role in scientific discovery, so will browser-based tools for exploring high-dimensional datasets, communicating scientific discoveries, sharing and aggregating data across labs, and publishing data alongside manuscripts.

Introduction

Fueled by technical advances in modern web browsers, and by the development of open-source software libraries for interactive visualization, browser-based data visualizations have been playing an increasingly prominent role in communicating data on topics ranging from news events, climate patterns, election results, and public health concerns, as well as research findings from a broad range of scientific research disciplines [1]. JavaScript libraries like *D3* [2] and *threejs* [3] rival most platform-specific visualization software libraries in terms of plotting and rendering capabilities, and support interactive data visualization and exploration.

As a consequence of the development of these general-purpose visualization tools, many scientific disciplines have further developed tools based on these libraries for the visualization of discipline-specific data types in the browser. In the field of neuroscience there are several different libraries devoted to visualization of brain imaging data. Examples include BrainBrowser [4], XTK [5], Mango [6] and Fiberweb [7] which provide application programming interfaces (APIs) for programmers to create sophisticated applications that visualize three-dimensional brain structure with overlaid analysis results. For example, using BrainBrowser, Sherif and colleagues show that analysis of functional connectivity can be efficiently performed on a dataset of millions of brain maps in the 1TB MACACC dataset [4]. A browser-based visualization of the cortical surface and a series of widgets are used to initiate server-side computations that return analysis results to the browser in the form of a light-weight three-dimensional rendering of the brain. The scope of the BrainBrowser project is comprehensive: It includes functions to flexibly handle most commonly used data formats, and can be adapted to visualize the results of many different types of analyses. Other tools, including XTK, and Mango also provide software developers with substantial flexibility to design applications that interact with remote data-sets and bring them to users in their own web-browsers. These new tools are ushering in an era of Big Data in human neuroscience, and have laid the technical infrastructure for visualizing the breadth of commonly used medical imaging data types.

In the present work, we leverage these technical developments towards a more specific goal: To build a graphical user interface (GUI) that is designed to visualize results from diffusion-weighted magnetic resonance imaging (dMRI) studies that employ one specific analysis package, Automated Fiber Quantification (AFQ) [8]. AFQ is an open-source toolbox for performing quantitative analysis of white matter fiber tracts in the human brain. AFQ is widely used across clinical and basic science applications ranging from brain development and aging [9–14], autism spectrum disorders [15,16], major depressive disorder [17,18], head trauma [19–21], retinal disease [22], amyotrophic lateral sclerosis [23], surgical planning [24], and dyslexia [13,25]. This narrow focus on designing a web-based GUI that is integrated with a specific analysis approach, rather than a general-purpose visualization toolbox, confronts two major challenges in the study of human brain connectivity: (1) scientific reproducibility and (2) exploration of high dimensional data. The intentionally narrow focus makes it possible to design a robust system that can be used by

researchers without technical expertise in JavaScript and web visualization. Instead, we designed a simple command-line interface (AFQ-Browser) to allow researchers to rapidly visualize and explore data on their own computers and to publish results to the web.

Scientific reproducibility: Because AFQ-browser is portable -- it runs in any modern web-browser -- it can be used to facilitate transparency and data sharing. The field of human neuroscience faces several specific challenges with regards to reproducibility. Scientists are generally aware that data sharing is integral to reproducible research, but it is not always clear how to usefully share data in a manner that allows other labs to understand and reproduce published findings. There is a spectrum of data sharing practices, each presenting its own challenges.

On one end of the spectrum, raw data is often unwieldy, large and complex, and access to it by itself, though very useful [26], does not guarantee reproducibility. For example, reproducing results from raw data requires access to the full series of computations that was used in the analysis and, in many cases, simply running these analyses requires substantial computational resources. Computational time and data size, can present a serious barrier that prevents many scientists from attempting to reproduce a published finding [27]. On the other end of the spectrum, tables, graphs and scatter plots that typically appear in journal articles reflect an author's interpretation of the data, but do not suffice for meaningful reproducibility of the results, or exploration of alternative theories. A related issue is that the analysis of raw medical imaging data requires substantial domain expertise -- knowledge about the biology, anatomy and physiology of the system, the physics of the experimental signal generation process, and the domain-specific file formats used to store the data. This presents a barrier for researchers in computer science and statistics to apply innovations in their fields to the analysis of human brain data and to crosscheck the methodological assumptions of published work.

Here, we propose that sharing dimensionally-reduced portions of dMRI data, together with rich interactive data visualizations, lends itself not only to replication of original results, but to immediate and straight-forward extensions of these results, even in the hands of researchers in other disciplines. Ideally, this intermediate form of data sharing would supplement the release of raw data, but also might be appealing to researchers who wish to more completely communicate their findings, but are not ready to release the full collection of raw data from an ongoing study, or worry about privacy concerns associated with raw data. AFQ-browser automatically organizes data analyzed with AFQ into tables of "tidy" data [28]. The software facilitates rapid publication of both the visualization, and these data, as an openly available website.

Exploratory data analysis and linked view visualizations: Data visualization and exploration plays an integral role in scientific inquiry, even beyond communicating results from statistical tests of an *a priori* hypothesis. The statistician John Tukey coined the term "exploratory data

analysis" to describe the process of data analysis through iterative processing, probing and visualization of datasets [29]. Tukey argued for a sharp distinction between exploratory and confirmatory data analysis (or hypothesis testing), and posited that scientists should strive to obtain multiple datasets allowing them to explore a high-dimensional system, and develop hypotheses through exploratory data analysis, before performing the formal statistical tests to confirm or reject their hypothesis based on an independent dataset. In complex systems, with non-linear relationships, exploratory data analysis and visualization can be essential for clarifying patterns that might have been obscured in a conventional statistical analysis [30].

High-dimensional datasets, such as Tract Profiles of white matter tissue properties measured with dMRI [8], in conjunction with behavioral and demographic measures in large samples of subjects, pose a fundamental challenge for data visualization. A solution pioneered by astronomy, genomics and other fields that were early to embrace "Big Data" was the development of tools implementing linked views of a data set, where interaction with a visualization of one dimension evokes a change in another visualization of the same data [31]. By interactively exploring the relationships among different dimensions of a dataset, a researcher can develop an understanding of the principles that characterize the system without specifying an *a priori* model of the complex relationships that are present in the high-dimensional data. Drawing inspiration from other disciplines that have already realized the power of tools that implement linked view visualization for exploring high-dimensional data, we capitalize on new, open source JavaScript libraries to create the first platform-independent graphical user interface for exploratory data analysis of high-dimensional dMRI datasets.

In summary, we present here a software tool that visualizes results from the analysis of dMRI data with AFQ and facilitates exploratory data analysis through the implementation of linked views of the data. The system also facilitates reproducible research by making it easy for researchers to publish these visualizations and the underlying data as an interactive website. The publication of these results and data will allow researchers, stakeholders, and other members of the general public to explore large, important datasets through a web-browser, without having to download the data or execute a complex and unwieldy processing pipeline. By satisfying the need for both exploratory data analysis and data sharing, AFQ-Browser support a virtuous cycle where public data is increasingly valuable and easy to share, and there are new opportunities to aggregate large datasets across laboratories.

Methods

The AFQ-Browser software

Automated Fiber-tract Quantification (AFQ) is a software package for quantitative analysis of white matter fiber tracts [8]. The AFQ software is a fully automated pipeline that takes in

diffusion MRI data and returns Tract Profiles of diffusion properties (or other quantitative MRI parameters) sampled along the trajectory of 24 major white matter fiber tracts. Fiber tracts are identified in an individual's native space, and the diffusion properties are sampled at points along the trajectory of each tract, thereby representing the data for each tract as a vector of measurements. For groups of subjects, data for a tract is represented by a matrix of values where each row corresponds to a subject and each column corresponds to a node along the tract. This pipeline can be thought of as a dimensionality reduction technique, whereby the data from hundreds of thousands of voxels gets summarized in terms of features (fiber tracts) that have a known anatomy, and are important for specific aspects of cognitive function. Based on this dimensionality reduction and alignment into the individual participant's anatomy, groups of subjects can be compared in terms of these features, individuals can be compared to groups, and supervised and unsupervised learning techniques can be applied to link white matter biology to cognition in health and disease. But even this lower-dimensional view of the diffusion data can become unwieldy as datasets grow larger, and as there is an increasingly complex collection of subject meta-data characteristics (e.g., behavioral measures, demographics, disease state, etc.) that might be linked to the underlying biological measurements. Hence, data visualization that allows for linked views across different dimensions of the dataset is essential.

AFQ-Browser takes the output of the AFQ MATLAB pipeline, and generates a browser-based visualization of the results. The AFQ MATLAB analysis pipeline produces a standard AFQ object, stored as a MATLAB .mat/hdf5 file. This file contains a structure array data-structure, with Tract Profiles for all the diffusion properties that were calculated from the dMRI data, for all tracts and subjects. The AFQ file also contains a field for *metadata*: Subject level characteristics, such as age, clinical diagnosis, or scores on psychometric tests are saved into this field. A command line function, *afqbrowser-assemble*, extracts all this information from the AFQ .mat file and writes out the hierarchically nested structured array as a series of .csv and .json files, stored in tidy formats [28]. This command line application then organizes the various AFQ-Browser files into a fully-functioning AFQ-Browser website: A template of HTML and JavaScript scripts, and CSS styling are arranged into the appropriate folder structure, and the data are placed in a data folder from which the application files read it into the browser. A second command line function, *afqbrowser-run*, launches a static web-server on the user's computer with AFQ-Browser running for this dataset. Navigating a web browser to the returned url (defaulting to https://localhost:8080), will open the visualization.

Linked visualization

The browser based GUI has four panels (**Figure 1**): BUNDLE LIST; (b) ANATOMY; (c) BUNDLE DETAILS; (d) SUBJECT METADATA. The visualization is linked across the panels in four ways.

- First, color is used to identify each fiber tract (here referred to as "Bundle") across the Bundle List, 3D Brain visualization, and Bundle Details plots. We use the categorical Tableau-20 color scheme (https://www.tableau.com/about/blog/2016/7/colors-upgrade-tableau-10-56782). Clicking on a tract in the Bundle List, or 3D Brain, will highlight that tract in both panels and open up the corresponding line plot showing diffusion properties of that tract for each subject.
- Second, the Tract Profiles from each individual subject in the Bundle Details panels are linked to their metadata. Selecting a Tract Profile (line) in the Bundle Details plot will highlight that subject's row of the metadata table, and selecting a row of the metadata table will highlight that subjects Tract Profile in the plot. A subject of interest can be selected based on their metadata to visualize their Tract Profiles relative to the group of other subjects, or a Tract Profile of interest can be selected to compare their metadata against the group of other subjects.
- Third, columns in the Metadata table are linked to mean lines in the Bundle Details plots. Clicking a column will sort the metadata table based on the data in that field, and subjects will be divided into N groups by binning the data (the number of groups can be defined in a control bar). Each bin will be assigned a color and this color will be used for the rows of the metadata table, the mean lines in the Bundle Details plot, and the individual subject lines in the plot. Each time a new column in the metadata is selected, the mean lines are updated in the plot, and the rows are sorted and colored appropriately in the metadata table. This feature provides an efficient tool to slice a large data set across different dimensions, examine how different subject characteristics relate to diffusion measures, determine subjects that are outliers within a group, and determine how different manners of grouping produce changes across different white matter fiber tracts.
- Fourth, the spatial dimension (x-axis) of the Bundle Details plots is linked to the fiber tracts in the 3D brain visualization. Manual selection (*brushing* [32]*)* of a range of nodes in the Bundle Details plot, enabled by toggling on the *brushable tracts* feature in a control bar, highlights the corresponding region of the fiber tract in the 3D brain. This feature allows a user to link statistics, group differences, or quantitative comparisons of an individual subject back to their brain anatomy.

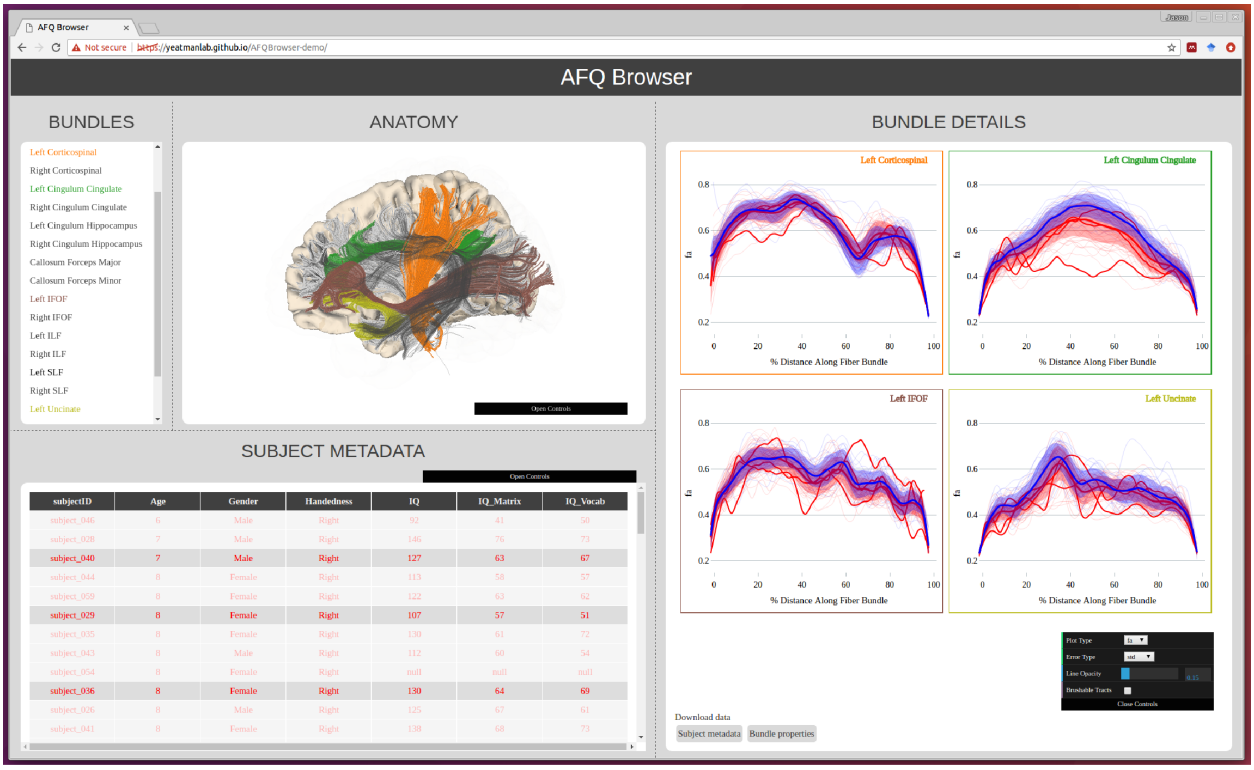

Figure 1: AFQ-Browser v1.0. *The BUNDLES panel displays the names of the tracts and the colors are linked to the ANATOMY and BUNDLE DETAILS panels. Selecting a tract in the BUNDLES or ANATOMY panel will display the Tract Profile in the BUNDLE DETAILS panel. Selecting an individual subject's Tract Profile will highlight that subject in the SUBJECT METADATA panel. Selecting a column of SUBJECT METADATA groups subjects based on this measure. In the example, subjects are grouped based on age and means and standard deviations are shown in the BUNDLE DETAILS panel.*

Publishing data for reproducible science

A single command, *afqbrowser-publish*, packages the entire website, including both data and visualization into a git repository, and uploads this repository to GitHub. This script automatically creates a website with this data, hosted on the repository's "GitHub Pages" website, so that it can be viewed by anyone through a web-browser. The published website also includes a link that allows users to download the .csv files that contain the information that is displayed, for additional computational exploration through other tools (e.g., by reading the data into scripts that implement machine learning algorithms). The only requirement is that the user has a GitHub account and *afqbrowser-publish* will create the public repository, build the webpage, and launch the web server through GitHub.

Saving the browser state

Reproducing results that are generated by a graphical user interface (GUI) can be problematic since figures are generated based on a series of user inputs (i.e., mouse clicks and key presses). To solve this problem, we have built a "save browser state" function into AFQ-Browser. This function saves a settings file that will load a specific browser state when AFQ-Browser is launched. Hence, a discovery made through a series of operations in the GUI can be communicated without a lengthy description of the series of user inputs.

Installation of AFQ-Browser

The current version of AFQ-Browser can be cloned from the GitHub repository: https://github.com/YeatmanLab/AFQ-Browser. The current stable release can be found on the Python Package Index (https://pypi.python.org/pypi/AFQ-Browser) and it can be installed, together with all of its dependencies, on any machine with Python and the pip package manager simply by calling: pip install AFQ-Browser.

Results

Generating new discoveries from old datasets

Publishing data in a convenient format supports reproducibility and fuels new scientific discoveries. For example, examining the published data from Yeatman, Wandell and Mezer (2014) [14] in a running instance of AFQ-Browser (http://YeatmanLab.github.io/AFQBrowser-demo), we can reproduce the previously reported finding that, in terms of mean diffusivity (MD), the arcuate fasciculus demonstrates more developmental change than the corticospinal tract (CST). When the sample is binned into three age groups, both the arcuate and CST show highly significant changes, but the magnitude of change between childhood and adulthood is larger for the arcuate, than the CST (**Figure 2b**). By switching the plot to fractional anisotropy (FA) rather than MD, another effect, not reported in the original manuscript, can be observed. While the arcuate shows the expected pattern of results - FA values increase with development - the CST shows the opposite pattern of developmental change. For the CST, the three age groups have equivalent FA values for the first half of the tract, but adults have lower FA values than young adults or children between nodes 50 and 80 (**Figure 2b**). At first this finding might seem counter-intuitive: FA typically increases with development as axons become more densely packed and myelinated. But in this case the developmental decline in FA occurs in the centrum semiovale, a portion of the Tract Profile where FA drops substantially due to crossing fibers. The developmental decline in FA is therefore likely to reflect development of the fiber tracts that cross through this portion of the CST, rather than changes in CST axons per se. This interpretation makes sense given that the superior longitudinal fasciculus (SLF), one of the

tracts crossing through this region of the CST, is believed to continue developing into young adulthood. This interpretation of the developmental changes in FA in regions of crossing fibers offers some clarity to other reports of declining FA values in the young adult brain, but also requires a more thorough investigation in an independent dataset.

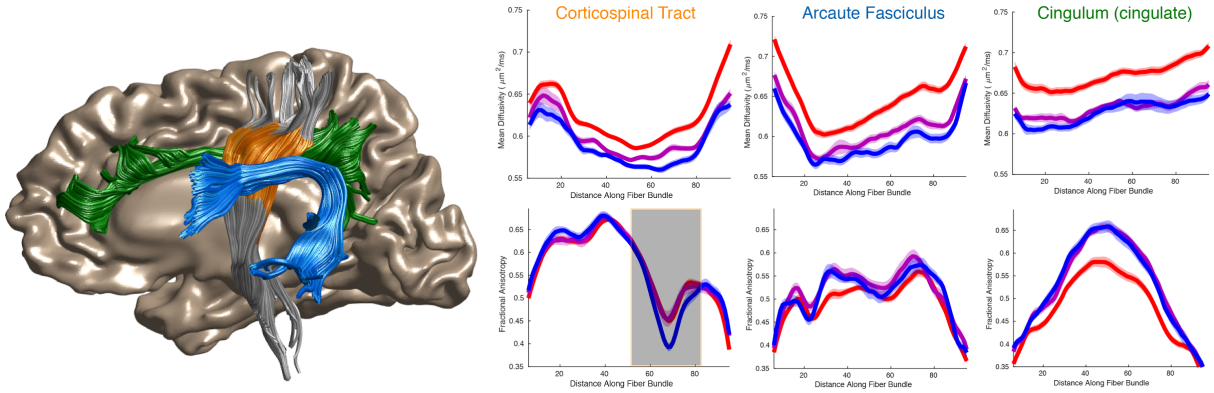

Figure 2: Development of the corticospinal tract, arcuate fasciculus and cingulum. *Tract Profiles of Mean Diffusivity (top) and Fractional Anisotropy (bottom) are shown for the left hemisphere corticospinal tract (CST, orange), arcuate fasciculus (blue) and cingulum (green). Splitting the group by age, and selecting 3 bins, displays mean lines three groups: 8-15 (red), 15-30 (purple), 30-50 (blue). For the CST, there is a region that shows a decrease in FA with development, and this location of the tract is highlighted on the plot using the "brushable tracts" feature (shaded gray box). The linked view in the anatomy displays the portion of the CST that is brushed in the plot demonstrating that this effect occurs in the anatomical portion of the CST known as the centrum semiovale, adjacent to the arcuate fasciculus. This linked visualization provides a connection between the data plots and the 3D Anatomy. Data and MATLAB code available at https://github.com/YeatmanLab/AFQ-Browser_data (see Figure2_Development.m) and running AFQ-Browser instance at: https://YeatmanLab.github.io/AFQBrowser-demo/)*

Localizing white matter lesions in patients with multiple sclerosis

Multiple Sclerosis (MS) is a degenerative disease of the white matter characterized by progressive loss of myelin. Even though measures such as MD and FA are not specific to myelin, dMRI is still a promising technique for detecting and monitoring white matter lesions in MS and quantifying results from drug trials targeting remyelination [33]. DMRI is sensitive to aspects of the disease that are not detectable with conventional imaging methods (T1, T2, FLAIR). Quantitative comparisons between MS patients and healthy control subjects have demonstrated differences in diffusion properties within "normal appearing white matter", or regions that do not show obvious lesions on a conventional MRI image. In longitudinal studies, these regions with diffusion differences are likely to progress into lesions, indicating the sensitivity of dMRI for detecting early signs of the disease and monitoring the benefit of drugs that aim to prevent the demyelination process [33–36].

One of the challenges for incorporating dMRI into clinical practice is the lack of user-friendly methods for visualizing results in a quantitative manner. For clinical applications, group comparisons have limited utility, because ultimately the goal is to detect abnormalities and make diagnoses at the level of the individual. For example, in the data previously published by Yeatman, Wandell, Mezer and colleagues [14,37], mean diffusivity (MD), radial diffusivity (RD), and fractional anisotropy (FA) values are significantly different in MS patients compared to controls for most tracts in the brain (https://jyeatman.github.io/AFQ-Browser-MSexample/). MD and RD show much greater sensitivity to group differences than FA: **Figure 3** show group means and standard errors for MD, RD and FA along the corticospinal tract, posterior callosum, inferior longitudinal fasciculus and arcuate fasciculus.

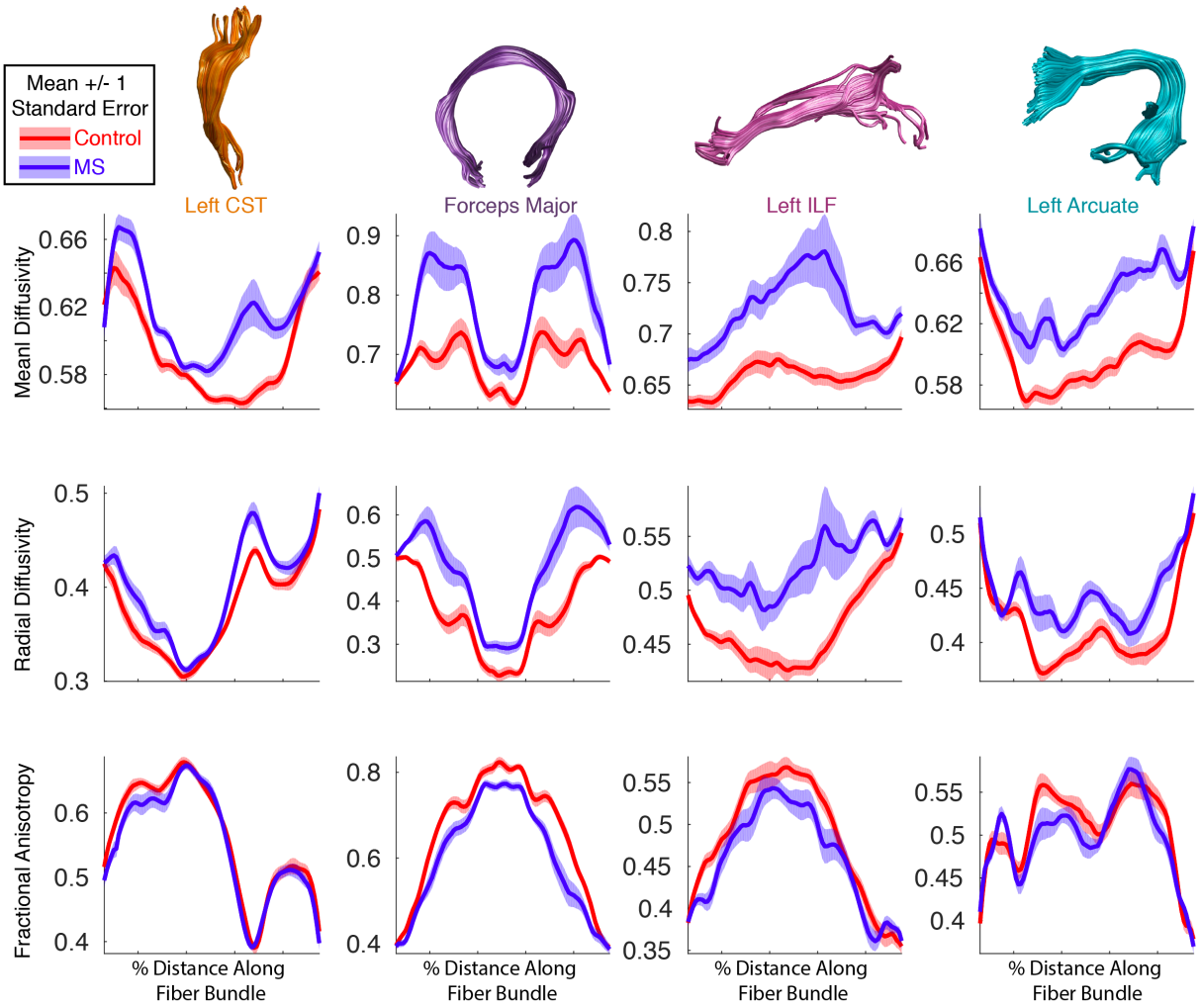

Figure 3: Group comparison between multiple sclerosis patients and healthy control subjects. *We observe highly significant (p<0.001) group differences in diffusion measures across many tracts. Mean diffusivity (top panel) and radial diffusivity (middle panel) show larger group differences than fractional anisotropy (bottom panel). Mean values +/- 1 standard error are shown for control subjects in red and multiple sclerosis (MS) patients in blue. (see Figure3_4_MultipleSclerosis.m and https://jyeatman.github.io/AFQ-Browser-MSexample/)*

Group comparisons demonstrate the sensitivity of the measure to the disease but don't provide diagnostic information about individual patients: each individual has tissue abnormalities in different parts of the brain, with some tracts showing diffusivity values in the normal range, others showing normal appearing white matter on a T1, but abnormalities in terms of diffusion metrics, and other tracts displaying major lesions. AFQ-Browser provides a simple and intuitive method to quantitatively compare an individual's white matter tissue properties to normative data from healthy brains by plotting each individual's Tract Profile in comparison to the normative distribution of healthy brains (means and standard deviations, **Figure 4**). Such a comparison can localize lesions to specific locations on a tract and quantify the extent of damage. Clinical data is a prime example of the utility of linked visualization: the links between quantitative plots of diffusion measures, tract anatomy, and subject metadata make it possible to quickly find a subject with a lesion, determine the location of the lesion and associate this information with clinical symptoms. While not as specific to myelin as other quantitative measurements such as R1 [14,38–41], we find that MD and RD are highly sensitive to MS lesions. For example the lesion shown in Figure 8 of [14] can be detected based on MD values that are 5.6 SD away from the norms, with a larger lesion in the left compared to the right occipital callosal connections (**Figure 4, subject_020**). In this lesion, RD values are slightly more sensitive showing a z-score of 6.2 and FA values are slightly less sensitive, with a z-score of -3.3 compared to healthy controls (**Figure 4**). For this patient, the large lesion on the ILF was more than 10 SD greater than the controls in terms of MD and RD. As more clinical datasets are aggregated in public repositories there will be new opportunities to explore the sensitivity and specificity of this type of individual comparison.

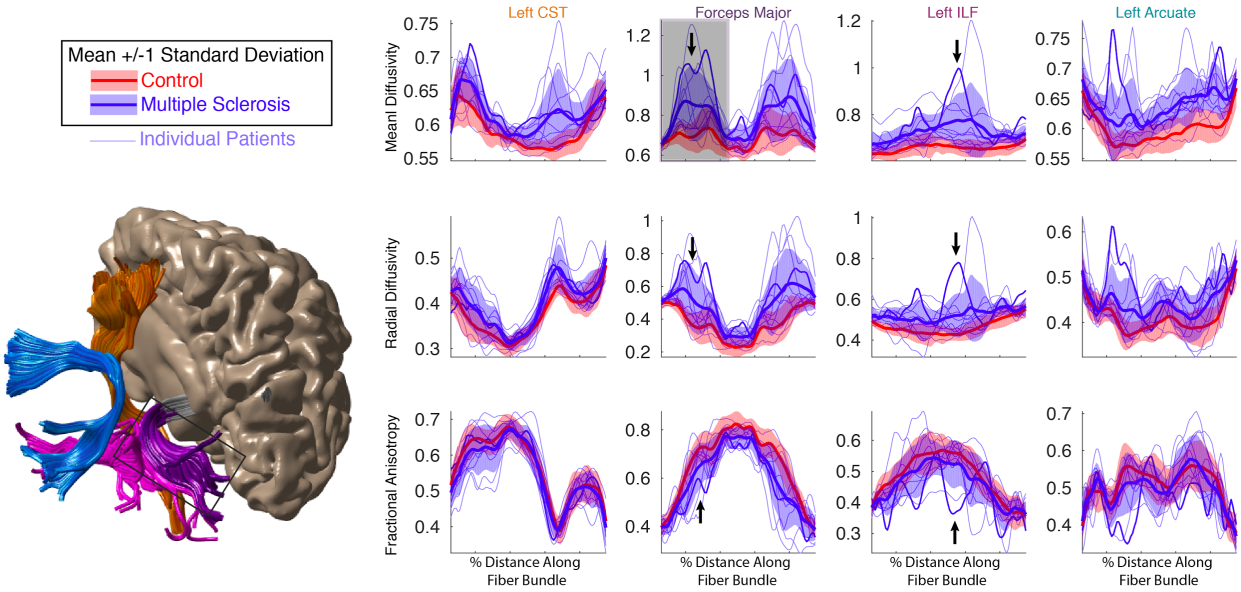

Figure 4: Localizing lesions in an individual's brain. *Individual MS patients (light blue lines) are plotted against the normal distribution (mean +/-1 1 standard deviation) of values in healthy*

control subjects. Lesions and diffuse abnormalities can be detected in individuals based on large deviations from the control subjects. The darker blue line is data from the patient shown in Figure 8 of Yeatman, Wandell and Mezer, 2014. By plotting standard deviations rather than standard errors, the large (>1SD) difference between MS patients and control subjects is apparent, as are the large deviations of specific patients from the normal distribution. (see Figure3_4_MultipleSclerosis.m and https://jyeatman.github.io/AFQ-Browser-MSexample/)

Detecting degeneration of the corticospinal tract in amyotrophic lateral sclerosis (ALS)

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease in which progressive degeneration of upper and lower motor neurons leads to atrophy, weakness and loss of muscle control. The time-course of disease progression varies substantially across patients, with some showing rapid degeneration and other showing a sporadic or gradual decline. Due to the heterogeneous presentation of clinical symptoms in ALS, early diagnosis can be challenging and the disease can go undetected in many patients until they present with severe symptoms. Hence, the development of quantitative and automated methods for diagnosis and disease monitoring has been a major focus within clinical neuroimaging research. Diffusion MRI, holds promise as a tool to detect the early stages of neural degeneration and corroborate behavioral assessments. Group analyses have consistently demonstrated significant reductions in FA, increases in RD and increases in MD in the corticospinal tract [23,42]. Group comparisons provide information about the average pattern of disease progression but ultimately the goal of clinical neuroimaging research is to develop techniques that have sufficient sensitivity and specificity to be applicable at the individual level. A recent study used AFQ and a random forest classifier to develop an automated diagnosis system to classify subjects as healthy or diseased based on dMRI measures [23]. They achieved 80% classification accuracy (cross-validated) based on Tract Profiles of the corticospinal tract and reported that FA and RD at the level of the cerebral peduncle and posterior limb of the internal capsule were the most informative diffusion properties. These effects can be visualized in AFQ browser by binning the subjects based on disease diagnosis (https://yeatmanlab.github.io/Sarica_2017/**, Figure 5**). As reported by Sarica et. al, the mean RD and FA values in this region of the CST are more than 1SD different in ALS patients compared to controls (node 40, arrow **Figure 5**). Even though a multivariate classification strategy (random forests) is used to achieve good diagnostic accuracy, visualization of individual Tract Profiles demonstrates that a majority of patients (75%) deviated by more than 1SD from control values within the right CST at the level of the cerebral peduncle.

The goal of most clinical neuroimaging studies is to detect regions of the brain that are affected by the disease. While not a central focus of clinical research, there is also scientific importance to clearly establishing regions of the brain that are not affected by the disease. Based on the previously published data in Sarica et al. [23], we can investigate the specificity of the effects to the CST and determine whether there are there are any tracts that can be established as

control regions not affected by the disease. We find that the CST is the only tract that shows large (>1SD) differences between patients and controls in terms of RD and FA values. While there are a few regions that show small differences (depending on the statistical threshold), the specificity of the effects to the CST is striking. For example, many tracts including the forceps major and forceps minor of the corpus callosum, and the left and right inferior fronto-occipital fasciculus show nearly identical distributions of values between patients and controls (**Figure 5**).

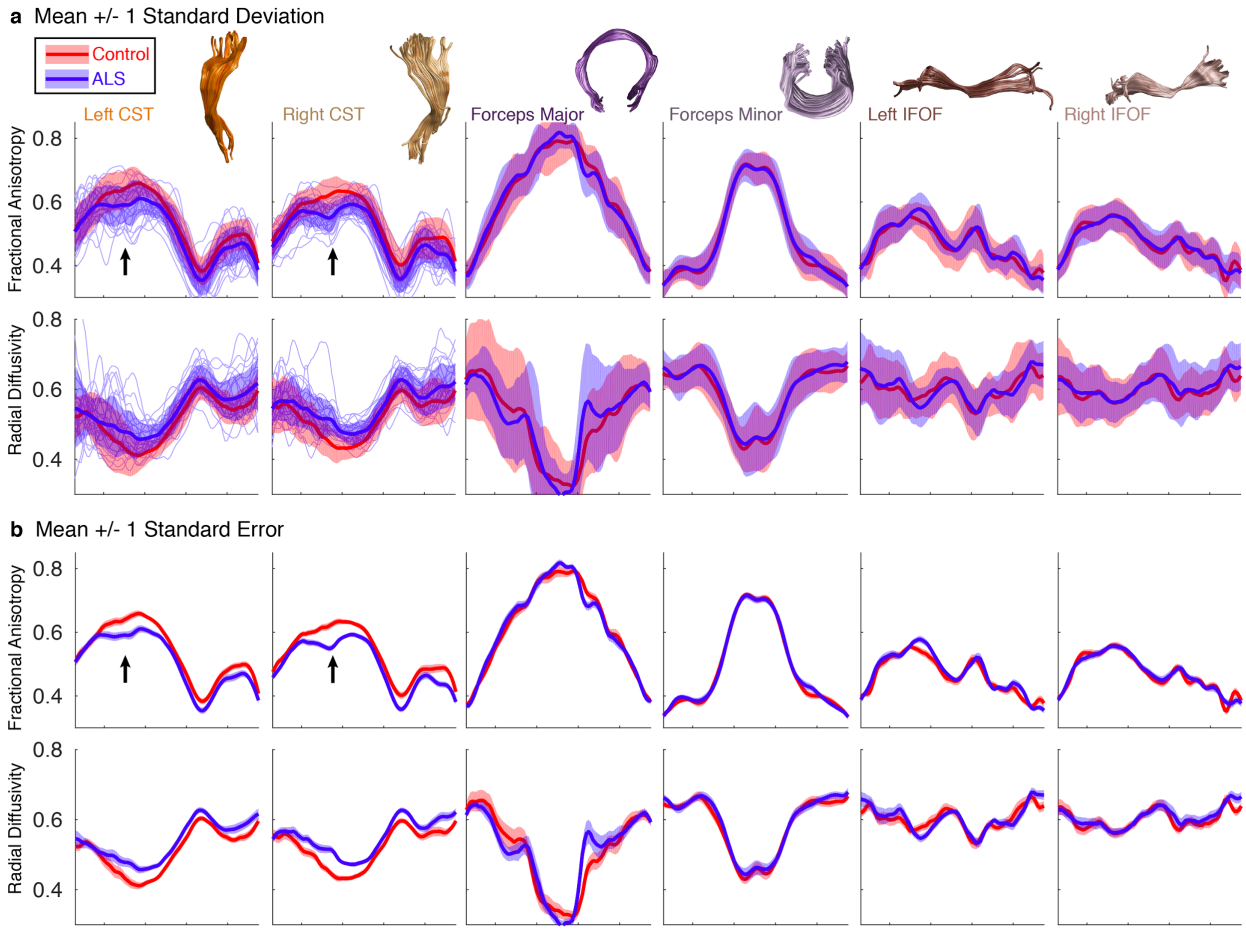

Figure 5: Amyotrophic lateral sclerosis patients show isolated degeneration of the corticospinal tract. *(a) Means and standard deviations of FA and RD values are shown for ALS patients (blue) and control subjects (red). Individual patients are displayed as light blue lines for the CST. At the level of the cerebral peduncle, patients differ from controls by more than 1 standard deviation (black arrow). No other tracts show this large effect. (b) Means and standard errors are shown for ALS and control subjects to indicate regions of significant group differences. Group differences are relatively specific to the CST. (See Figure5_ALS.m and https://YeatmanLab.github.io/Sarica_2017/)*

Removing barriers for interdisciplinary collaboration: Informed features shared as "tidy data"

Statistics, machine learning, and data science are making impressive strides in the development of general-purpose methods for the interpretation of data across a variety of scientific fields [43]. One of the current barriers to a broader application of these methods is the extraction of useful analysis features from unstructured data sets that contain large, heterogeneous, noisy measurements, saved in obscure domain-specific or proprietary formats, that require special software, and arcane preprocessing steps. Brain imaging data is a paradigmatic case of this state of affairs: measurements are typically large, on the order of several gigabytes per individual, signal-to-noise ratio is low, and differences in 3D brain structure between individuals make naive image processing of the original measurement fraught. One of the major strengths of the AFQ approach is that it extracts features from brain imaging data based on domain-specific knowledge: quantitative measurements of tissue properties for well-defined anatomical segments of the white matter connections in an individual's brain that contain the major tracts. This reduces the dimensionality of the data substantially, while still retaining rich, complex information about an individual's neuroanatomy.

AFQ-Browser provides these domain-relevant features in a format that will be familiar to many machine learning and statistics practitioners: Tables with observations as rows, and variables as columns. This format, known as "tidy data" [28] is the universal exchange format of data science. The data are converted by the AFQ-Browser software and stored in ubiquitous text-based formats: CSV and JSON files. Separate tables are available for node-by-node estimates of the diffusion properties along the length of the fiber groups, and for the subject metadata, and these tables can be merged in an unambiguous manner through a shared subject ID variable. These files can be read using the standard data science tool-box: Software libraries such as the Python pandas library [44], or using the R statistical language [45]. Once data are read into tables, data processing and visualization with tools such as Seaborn (https://seaborn.pydata.org/) or ggplot (http://ggplot.yhathq.com/) are also straightforward. Furthermore, very few steps are required to apply machine learning techniques to the data, using tools such as the scikit-learn library [46], and results such as classifier weights can be easily interpreted with respect to known brain anatomy. An example of such an analysis is presented in **Figure 6**, using the same data as in **Figure 5**.

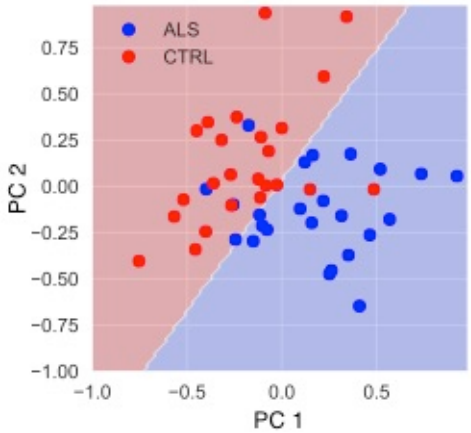

Figure 6: Classification of ALS patients based on FA in the corticospinal tracts. *The Tract Profiles in the two CSTs are submitted to a Principal Components Analysis -- the first two PCs form the dimensions of this plot (accounting for about 50% of the variance in the data). The data is separately used to train a support vector machine classifier, with a polynomial kernel. The classification boundary is shown here in the space of first two PCs. This classifier performs at 88% accuracy (cross-validated) in discriminating patients from controls. The Jupyter Notebook containing all steps of the analysis is shared here: https://github.com/yeatmanlab/AFQ-Browser_data/blob/master/AFQ-Browser_ALSexample/Figure6.ipynb*

Discussion

We have developed a new visualization tool for the quantitative analysis of diffusion MRI data in the web browser. The goals of this work were, first, to support scientific reproducibility by removing barriers to public data release and, second, to capitalize on new technologies for linked visualization that facilitate exploratory data analysis. AFQ-Browser makes it possible to create an interactive website of companion data for a manuscript with a single command (*afqbrowser-publish*). While, ultimately, we advocate for releasing all the raw data and analysis code associated with published work [26], we also maintain that releasing derived measures (Tract Profiles) is a major step in the right direction and will allay the concerns that many scientists feel about giving up control of difficult to collect data sets. Ideally, this practice will serve as a stepping-stone to further data and code sharing.

An additional benefit of releasing derived measures is that readers of a manuscript can easily explore dimensions of the data that were not reported in the publication. For example, it is not feasible to report results for every possible diffusion metric and it is common for a manuscript to focus on a single metric. In our previous work [14] we only reported modeling results for MD, R1 and MTV. A reader that is left wondering whether other metrics (e.g., RD and AD) would show the same pattern of results can now quickly answer this question through the companion website (http://yeatmanlab.github.io/AFQBrowser-demo). Not only is a companion website

more feasible than a supplement that includes every potential analysis, through AFQ-Browser researchers can extend published work and make new discoveries. For example, we have made three observations, that extend the findings reported in published datasets: (1) In regions of crossing fibers there are developmental declines in FA (**Figure 2**); (2) MS lesions can be detected in an individual, and localized on a tract, based on RD or MD but not FA (**Figure 4**); (3) White matter degeneration in ALS is highly specific to the corticospinal tract and many cortical association tracts are largely unaffected by the disease (**Figure 5**). While each of these discoveries is only an incremental contribution to what was reported in the original work, we contend that having datasets openly available online, with tools that facilitate data exploration, will fuel important new discoveries in human neuroscience.

Democratizing web-based visualization

We are not the first to create interactive web-based visualizations to accompany a manuscript. For example, the Allen Brain Institute has built a powerful GUI to explore large, multimodal genomics and physiology datasets (http://casestudies.brain-map.org/celltax). Friederici and colleagues built an interactive brain viewer to accompany a review paper on the neuroanatomy of language [47] so that readers could explore anatomy in a more detailed manner than is possible in a static figure (http://onpub.cbs.mpg.de/index.html). The BigBrain project [48] has released a high resolution atlas of the human brain histology that can be navigated based on custom WebGL code (https://bigbrain.loris.ca). Huth and colleagues used pyCortex [49] to build an interactive website to accompany recent work [50] on the structure of semantic maps in the human brain (http://gallantlab.org/huth2016/). There are numerous examples of beautiful interactive websites that labs have designed to accompany key studies, and interact with landmark datasets. However, these major achievements in browser-visualization are isolated to a few labs with high technical capabilities and the willingness to invest the time and resources required to design a custom website for a publication. AFQ-Browser fills an important gap by removing these constraints: a website can be published by running a single command (*afqbrowser-publish*) in a software package that can be installed automatically on any machine with Python (*pip install AFQ-Browser*), and the website is hosted for free through GitHub Pages. Thus, even labs with minimal resources and technical capabilities can communicate important scientific findings in an interactive format.

Is exploratory data analysis at odds with hypothesis testing?

Should we worry that we are supporting scientific transparency at the expense of artificially diminishing p-values? Traditionally, the field of cognitive neuroscience has approached data analysis with the goal of testing specific hypotheses. Thus, experiments and data collection are designed with a hypothesis in mind, and data analysis involves computing statistics to formally test this hypothesis. In hypothesis-driven science, data visualization is often viewed as separate

from the scientific investigation. However, with new imaging techniques, and large-scale data collection efforts, the field of human neuroscience sits at a transition point, where data mining is becoming appreciated as an increasingly important component of scientific discovery. Other scientific fields such as astronomy and genomics that have embraced Big Data, have discovered the critical role that data visualization can play in developing new theories [31]. As the field of human neuroscience transitions to an era of Big Data, tools like AFQ-Browser will become increasingly important as a way for scientists to interact with large datasets. As datasets grow, so will the importance of tools that can operate in the same manner on data stored on a personal computer in a laboratory, or on remote datasets stored in the cloud. Browser-based GUIs can fill this growing need.

However, we might also worry that in developing tools like AFQ-Browser, we are supporting reproducibility and data mining at the expense of "p-hacking" [51–53]. This is a valid concern and highlights the need for our standards on scientific rigor to evolve with the changing landscape of Big Data. For example, a lab might typically only conduct a limited number of statistical tests and, ideally, would correct p-values for each statistical test that was performed (not just the tests that were reported in the manuscript). But exploratory data analysis involves examining many possible processing pipelines and relationships between variables in a system [29]. The strength of tools like AFQ-Browser is the ease of exploring large datasets to identify relevant dimensions, and make data-driven discoveries that suggest a new hypothesis to test in future work. Data exploration is a critical component of hypothesis generation, and data mining tools should not be discarded over worries of p-hacking. But thoughtful consideration of statistical concerns is also paramount. Drawing a distinction between exploratory and confirmatory data analysis allays concerns over biased p-values by defining the central role of replication in scientific discovery. An observation that emerges from exploratory data analysis should be confirmed in an independent dataset. As more datasets become publicly available, confirmatory data analysis and independent replication will become standard practice. Tools like AFQ-Browser facilitate this goal of aggregating many independent datasets. Finally, Big Data should not be viewed as a replacement for small and careful, hypothesis driven investigations within a single laboratory. The field should strive for a balance between the innovative data-driven discoveries that can emerge from large public datasets, and the careful, targeted scientific investigations that a lab can undertake to definitively test a specific hypothesis.

Author Contributions

J.D.Y., A.R.H., J.K.S., and A.R. conceived the idea and designed the tool, A.R.H., J.K.S., and A.R. wrote the code, J.D.Y., A.R.H., J.K.S., and A.R. wrote the manuscript.

Acknowledgements

The work was funded through a grant by the Gordon & Betty Moore Foundation and the Alfred P. Sloan Foundation to the University of Washington eScience Institute. A.R.H.'s work was supported by the Department of Energy Computational Science Graduate Fellowship Program of the Office of Science and National Nuclear Security Administration in the Department of Energy under contract DE-FG02-97ER25308. We would like to thank Jeff Heer, for providing the original impetus for this work, as an assignment in his class on data visualization. We thank Parmita Mehta and Zac Lin for their work on the prototype of AFQ-Browser. Finally, we would like to thank the authors that contributed the public datasets discussed in this manuscript: Sarica A., Cerasa A., Valentino P., Trotta M., Barone S., Granata A., Nisticò R., Perrotta P., Pucci F., Quattrone A., Mezer A., Wandell B.A.

References

1. Tushar, A. & G Reich, N. flusight: interactive visualizations for infectious disease forecasts. *J. Open Source Softw.* (2017).

2. Bostock, M., Ogievetsky, V. & Heer, J. D^3: Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **17,** 2301–2309 (2011).

3. Cabello, R. Three. js. *URL https//github. com/mrdoob/three. js* (2010).

4. Sherif, T., Kassis, N., Rousseau, M.-É., Adalat, R. & Evans, A. C. BrainBrowser: distributed, web-based neurological data visualization. *Front. Neuroinform.* **8,** 89 (2014).

5. Hähn, D., Rannou, N., Ahtam, B., Ellen Grant, P. & Pienaar, R. Neuroimaging in the browser using the X Toolkit. *F1000Research* (2012).

6. Lancaster, J. L. *et al.* Automated analysis of fundamental features of brain structures. *Neuroinformatics* **9,** 371–380 (2011).

7. Ledoux, L.-P. *et al.* Fiberweb : diffusion visualization and processing in the browser. *Front. Neuroinform.* **11,** 54 (2017).

8. Yeatman, J. D., Dougherty, R. F., Myall, N. J., Wandell, B. A. & Feldman, H. M. Tract profiles of white matter properties: automating fiber-tract quantification. *PLoS One* **7,** e49790 (2012).

9. Teubner-Rhodes, S. *et al.* Aging-Resilient Associations between Arcuate Fasciculus Microstructure and Vocabulary Knowledge. *Snl* **36,** 7210–7222 (2015).

10. Johnson, R. T. *et al.* Diffusion properties of major white matter tracts in young, typically developing children. *Neuroimage* **88,** 143–154 (2014).

11. Travis, K. E., Leitner, Y., Feldman, H. M. & Ben-Shachar, M. Cerebellar white matter pathways are associated with reading skills in children and adolescents. *Hum. Brain Mapp.* **36,** 1536–53 (2015).

12. Yeatman, J. D., Dougherty, R. F., Ben-Shachar, M. & Wandell, B. A. Development of white matter and reading skills. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E3045-53 (2012).

13. Wang, Y. *et al.* Development of Tract-Specific White Matter Pathways During Early Reading Development in At-Risk Children and Typical Controls. *Cereb. Cortex* bhw095 (2016). doi:10.1093/cercor/bhw095

14. Yeatman, J. D., Wandell, B. A. & Mezer, A. Lifespan maturation and degeneration of human brain white matter. *Nat Commun* **5,** 4932 (2014).

15. Libero, L. E., Deramus, T. P., Lahti, A. C., Deshpande, G. & Kana, R. K. Multimodal neuroimaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates. *Cortex* **66,** 46–59 (2015).

16. Fingher, N. *et al.* Toddlers later diagnosed with autism exhibit multiple structural abnormalities in temporal corpus callosum fibers. *Cortex.* (2017). doi:10.1016/j.cortex.2016.12.024

17. Sacchet, M. D. *et al.* Structural abnormality of the corticospinal tract in major depressive disorder. *Biol. Mood Anxiety Disord.* **4,** 8 (2014).

18. Sacchet, M. D. *et al.* Characterizing white matter connectivity in major depressive disorder: Automated fiber quantification and maximum density paths. *Proc. IEEE Int. Symp. Biomed. Imaging* **11,** 592–595 (2014).

19. Bahrami, N. *et al.* Subconcussive Head Impact Exposure and White Matter Tract Changes over a Single Season of Youth Football. *Radiology* **281,** 919–926 (2016).

20. Yeh, P.-H. *et al.* Compromised Neurocircuitry in Chronic Blast-Related Mild Traumatic Brain Injury. *Hum. Brain Mapp.* **38,** 352–369 (2017).

21. Main, K. L. *et al.* DTI measures identify mild and moderate TBI cases among patients with complex health problems: A receiver operating characteristic analysis of U.S. veterans. *NeuroImage Clin.* **16,** 1–16 (2017).

22. Ogawa, S. *et al.* White matter consequences of retinal receptor and ganglion cell damage. *Invest. Ophthalmol. Vis. Sci.* **55,** 6976–86 (2014).

23. Sarica, A. *et al.* The corticospinal tract profile in amyotrophic lateral sclerosis. *Hum. Brain Mapp.* **38,** 727–739 (2017).

24. Keller, S. S. *et al.* Preoperative automated fibre quantification predicts postoperative

seizure outcome in temporal lobe epilepsy. *Brain* **140,** 68–82 (2017).

25. Langer, N. *et al.* White Matter Alterations in Infants at Risk for Developmental Dyslexia. *Cereb. Cortex* bhv281 (2015). doi:10.1093/cercor/bhv281

26. Poline, J.-B. *et al.* Data sharing in neuroimaging research. *Front. Neuroinform.* **6,** 9 (2012).

27. Wandell, B. A., Rokem, A., Perry, L. M., Schaefer, G. & Dougherty, R. F. Data management to support reproducible research. (2015).

28. Wickham, H. Tidy Data. *J. Stat. Softw.* **59,** (2014).

29. Tukey, J. W. *Exploratory data analysis*. *Addison Wesley, Reading* (Addison-Wesley Pub. Co, 1977).

30. Matejka, J. & Fitzmaurice, G. Same Stats, Different Graphs. in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* 1290–1294 (ACM Press, 2017). doi:10.1145/3025453.3025912

31. Goodman, A. A. Principles of high-dimensional data visualization in astronomy. *Astron. Nachr.* **333,** 505–514 (2012).

32. Heer, J. & Shneiderman, B. A taxonomy of tools that support the fluent and flexible use of visualizations. *Interact. Dyn. Vis. Anal.* **10,** 1–26 (2012).

33. Mallik, S., Samson, R. S., Wheeler-Kingshott, C. A. M. & Miller, D. H. Imaging outcomes for trials of remyelination in multiple sclerosis. *J. Neurol. Neurosurg. Psychiatry* **85,** 1396–1404 (2014).

34. Roosendaal, S. D. *et al.* Regional DTI differences in multiple sclerosis patients. *Neuroimage* **44,** 1397–1403 (2009).

35. Vrenken, H. *et al.* Altered diffusion tensor in multiple sclerosis normal-appearing brain tissue: cortical diffusion changes seem related to clinical deterioration. *J. Magn. Reson. Imaging* **23,** 628–636 (2006).

36. Sbardella, E., Tona, F., Petsas, N. & Pantano, P. {DTI} Measurements in Multiple Sclerosis: Evaluation of Brain Damage and Clinical Implications. *Mult. Scler. Int.* **2013,** 671730 (2013).

37. Mezer, A. *et al.* Quantifying the local tissue volume and composition in individual brains with magnetic resonance imaging. *Nat. Med.* **19,** 1667–72 (2013).

38. Dick, F. *et al.* In vivo functional and myeloarchitectonic mapping of human primary auditory areas. *J. Neurosci.* **32,** 16095–105 (2012).

39. Weiskopf, N., Mohammadi, S., Lutti, A. & Callaghan, M. F. Advances in MRI-based computational neuroanatomy: from morphometry to in-vivo histology. *Curr. Opin. Neurol.* **28,** 313–22 (2015).

40. Lutti, A., Dick, F., Sereno, M. I. & Weiskopf, N. Using high-resolution quantitative mapping of R1 as an index of cortical myelination. *Neuroimage* 1–13 (2013). doi:10.1016/j.neuroimage.2013.06.005

41. Stüber, C. *et al.* Myelin and iron concentration in the human brain: A quantitative study of MRI contrast. *Neuroimage* (2014). doi:10.1016/j.neuroimage.2014.02.026

42. Li, J. *et al.* A meta-analysis of diffusion tensor imaging studies in amyotrophic lateral sclerosis. *Neurobiol. Aging* **33,** 1833–1838 (2012).

43. Tarca, a L., Carey, V. J., Chen, X. W., Romero, R. & Draghici, S. Machine learning and its applications to biology. *PLoS Comput. Biol.* **3,** e116 (2007).

44. McKinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. (2011).

45. R Core Team. R: A language and environment for statistical computing. (2013).

46. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

47. Friederici, A. D. The Brain Basis of Language Processing: From Structure to Function. *Physiol. Rev.* **91,** (2011).

48. Amunts, K. *et al.* BigBrain: An Ultrahigh-Resolution 3D Human Brain Model. *Science (80-.).* **340,** (2013).

49. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* **9,** 23 (2015).

50. Huth, A. G., Heer, W. A. De, Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532,** 453–458 (2016).

51. Gellman, A. & Lokem, E. The Statistical Crisis in Science Data-dependent analysis—a 'garden of forking paths'—explains why many statistically significant comparisons don't hold up. *Am. Sci.* **102,** 460 (2014).

52. Gellman, A. The problems with p-values are not just with p-values. *Am. Stat.* (2016).

53. Wasserstein, R. L. & Lazar, N. A. The ASA's Statement on *p* -Values: Context, Process, and Purpose. *Am. Stat.* **70,** 129–133 (2016).