Chapter 1
Motivation

1.1 DSPA Mission and Objectives

This textbook is based on the Data Science and Predictive Analytics (DSPA) course taught by the author at the University of Michigan. These materials collectively aim to provide learners with a solid foundation of the challenges, opportunities, and strategies for designing, collecting, managing, processing, interrogating, analyzing, and interpreting complex health and biomedical datasets. Readers that finish this textbook and successfully complete the examples and assignments will gain unique skills and acquire a tool-chest of methods, software tools, and protocols that can be applied to a broad spectrum of Big Data problems.

The DSPA textbook vision, values, and priorities are summarized below:

- **Vision**: Enable active learning by integrating driving motivational challenges with mathematical foundations, computational statistics, and modern scientific inference.
- **Values**: Effective, reliable, reproducible, and transformative data-driven discovery supporting open science.
- **Strategic priorities**: Trainees will develop scientific intuition, computational skills, and data-wrangling abilities to tackle big biomedical and health data problems. Instructors will provide well-documented *R*-scripts and software recipes implementing atomic data filters as well as complex end-to-end predictive big data analytics solutions.

Before diving into the mathematical algorithms, statistical computing methods, software tools, and health analytics covered in the remaining chapters, we will discuss several *driving motivational problems*. These will ground all the subsequent scientific discussions, data modeling techniques, and computational approaches.

I. D. Dinov, *Data Science and Predictive Analytics*,
https://doi.org/10.1007/978-3-319-72347-1_1

1.2 Examples of Driving Motivational Problems and Challenges

For each of the studies below, we illustrate several clinically relevant scientific questions, identify appropriate data sources, describe the types of data elements, and pinpoint various complexity challenges.

1.2.1 Alzheimer's Disease

- Identify the relation between observed clinical phenotypes and expected behavior.
- Prognosticate future cognitive decline (3–12 months, prospectively) as a function of imaging data and clinical assessment (both model-based and model-free machine learning prediction methods will be used).
- Derive and interpret the classifications of subjects into clusters using the harmonized and aggregated data from multiple sources (Fig. 1.1).

1.2.2 Parkinson's Disease

- Predict the clinical diagnosis of patients using all available data (with and without the unified Parkinson's disease rating scale (UPDRS) clinical assessment, which is the basis of the clinical diagnosis by a physician).
- Compute derived neuroimaging and genetics biomarkers that can be used to model the disease progression and provide automated clinical decisions support.
- Generate decision trees for numeric and categorical responses (representing clinically relevant outcome variables) that can be used to suggest an appropriate course of treatment for specific clinical phenotypes (Fig. 1.2).

Data Source	Sample Size/Data Type	Summary
ADNI Archive	Clinical data: demographics, clinical assessments, cognitive assessments; Imaging data: sMRI, fMRI, DTI, PiB/FDG PET; Genetics data: Illumina SNP genotyping; Chemical biomarker: lab tests, proteomics. Each data modality comes with a different number of cohorts. Generally, $\mathbf{200} \leq N \leq \mathbf{1200}$. For instance, previously conducted ADNI studies with $N > \mathbf{500}$ [doi: 10.3233/JAD-150335, doi: 10.1111/jon.12252, doi: 10.3389/fninf.2014.00041].	ADNI provides interesting data modalities, multiple cohorts (e.g., early-onset, mild, and severe dementia, controls) that allow effective model training and validation NACC Archive.

Fig. 1.1 Outline of an Alzheimer's disease case-study

Data Source	Sample Size/Data Type	Summary
PPMI Archive	Demographics: age, medical history, sex; Clinical data: physical, verbal learning and language, neurological and olfactory (University of Pennsylvania Smell Identification Test, UPSIT) tests, vital signs, MDS-UPDRS scores (Movement Disorder; Society-Unified Parkinson's Disease Rating Scale), ADL (activities of daily living), Montreal Cognitive Assessment (MoCA), Geriatric Depression Scale (GDS-15); Imaging data: structural MRI; Genetics data: lllumina ImmunoChip (196,524 variants) and NeuroX (covering 240,000 exonic variants) with 100% sample success rate, and 98.7% genotype success rate genotyped for APOE e2/e3/e4. Three cohorts of subjects; Group 1 = {de novo PD Subjects with a diagnosis of PD for two years or less who are not taking PD medications}, N1 = 263; Group 2 = {PD Subjects with Scans without Evidence of a Dopaminergic Deficit (SWEDD)}, N2 = 40; Group 3 = {Control Subjects without PD who are 30 years or older and who do not have a first degree blood relative with PD}, N3 = 127.	The longitudinal PPMI dataset including clinical, biological, and imaging data (screening, baseline, 12, 24, and 48 month follow-ups) may be used conduct model-based predictions as well as model-free classification and forecasting analyses.

Fig. 1.2 Outline of a Parkinson's disease case-study

Data Source	Sample Size/Data Type	Summary
MAWS Data / UMHS EHR / WHO AWS Data	Scores from Alcohol Use Disorders Identification Test-Consumption (AUDIT-C), including dichotomous variables for any current alcohol use (AUDIT-C, question 1), total AUDIT-C score > 8, and any positive history of alcohol withdrawal syndrome (HAWS).	~1,000 positive cases per year among 10,000 adult medical inpatients, % RAWS screens completed, % positive screens, % entered into MAWS protocol who receive pharmacological treatment for AWS, % entered into MAWS protocol without a completed RAWS screen.

Fig. 1.3 Outline of a substance use case-study

1.2.3 Drug and Substance Use

- Is the Risk for Alcohol Withdrawal Syndrome (RAWS) screen a valid and reliable tool for predicting alcohol withdrawal in an adult medical inpatient population?
- What is the optimal cut-off score from the AUDIT-C to predict alcohol withdrawal based on RAWS screening?
- Should any items be deleted from, or added to, the RAWS screening tool to enhance its performance in predicting the emergence of alcohol withdrawal syndrome in an adult medical inpatient population? (Fig. 1.3)

Data Source	Sample Size/Data Type	Summary
ProAct Archive	Over 100 clinical variables are recorded for all subjects including: Demographics: age, race, medical history, sex; Clinical data: Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS), adverse events, onset_delta, onset_site, drugs use (riluzole). The PRO-ACT training dataset contains clinical and lab test information of 8,635 patients. Information of 2,424 study subjects with valid gold standard ALSFRS slopes will be used in out processing, modeling and analysis.	The time points for all longitudinally varying data elements will be aggregated into signature vectors. This will facilitate the modeling and prediction of ALSFRS slope changes over the first three months (baseline to month 3).

Fig. 1.4 Outline of an amyotrophic lateral sclerosis (Lou Gehrig's disease) case-study

1.2.4 Amyotrophic Lateral Sclerosis

- Identify the most highly significant variables that have power to jointly predict the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function).
- Provide a decision tree prediction of adverse events based on subject phenotype and 0–3-month clinical assessment changes (Fig. 1.4).

1.2.5 Normal Brain Visualization

The SOCR Brain Visualization tool (http://socr.umich.edu/HTML5/BrainViewer) has preloaded sMRI, ROI labels, and fiber track models for a normal brain. It also allows users to drag and drop their data into the browser to visualize and navigate through the stereotactic data (including imaging, parcellations, and tractography) (Fig. 1.5).

1.2.6 Neurodegeneration

A recent study of Structural Neuroimaging in Alzheimer's disease (https://www.ncbi.nlm.nih.gov/pubmed/26444770) illustrates the Big Data challenges in modeling complex neuroscientific data. Specifically, 808 ADNI subjects were divided into 3 groups: 200 subjects with Alzheimer's disease (AD), 383 subjects with mild cognitive impairment (MCI), and 225 asymptomatic normal controls (NC). Their sMRI data were parcellated using BrainParser, and the 80 most important neuroimaging biomarkers were extracted using the global shape analysis pipeline workflow. Using a pipeline implementation of Plink, the authors obtained 80 SNPs highly associated with the imaging biomarkers. The authors observed significant

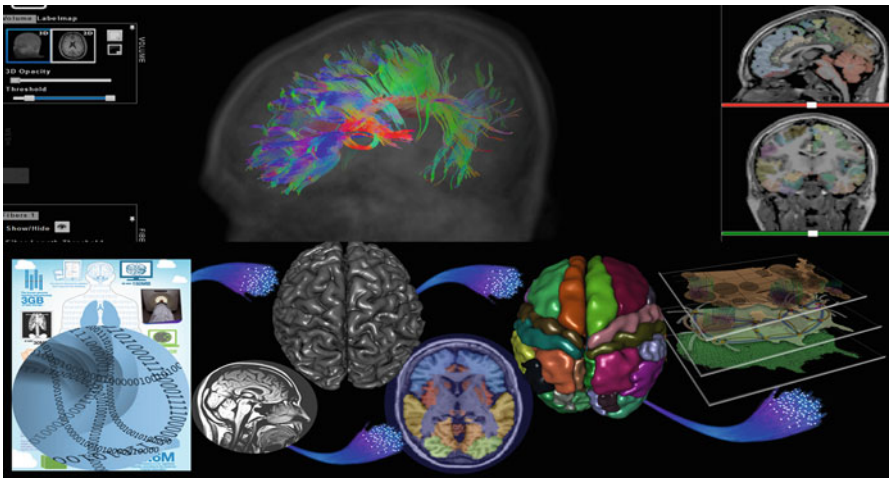
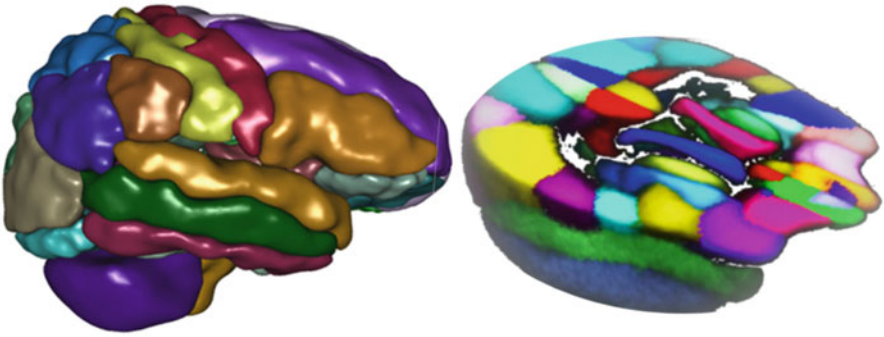


Fig. 1.5 Interactive 3D brain visualization

correlations between genetic and neuroimaging phenotypes in the 808 ADNI subjects. These results suggest that differences between AD, MCI, and NC cohorts may be examined by using powerful joint models of morphometric, imaging, and genotypic data (Fig. 1.6).

1.2.7 Genetic Forensics: 2013–2016 Ebola Outbreak

This Howard Hughes Medical Institute (HHMI) disease detective activity illustrates the genetic analysis of sequences of Ebola viruses isolated from patients in Sierra Leone during the Ebola outbreak of 2013–2016. Scientists track the spread of the virus using the fact that most of the genome is identical among individuals of the same species, most similar for genetically related individuals, and more different as the hereditary distance increases. DNA profiling capitalizes on these genetic differences particularly in regions of noncoding DNA, which is DNA that is not transcribed and translated into a protein. Variations in noncoding regions have less impact on individual traits. Such changes in noncoding regions may be immune to natural selection. DNA variations called **short tandem repeats (STRs)** are comprised on short bases, typically 2–5 bases long, that repeat multiple times. The repeat units are found at different locations, or loci, throughout the genome. Every STR has multiple alleles. These allele variants are defined by the **number of repeat units** present or by the **length of the repeat sequence**. STRs are surrounded by nonvariable segments of DNA known as flanking regions. The STR allele in Fig. 1.7 could be denoted by "6", as the repeat unit (GATA) repeats 6 times, or as 70 base pairs (bps) because its length is 70 bases in length, including the starting/ending flanking regions. Different alleles of the same STR may correspond to different number of GATA repeats, with the same flanking regions.



A: Individual brain parcellation B: LPBA40 atlas

Index	Volume Intensity	ROI Name	Index	Volume Intensity	ROI Name
1	21	L superior frontal gyrus	29	65	L inferior occipital gyrus
2	24	R middle frontal gyrus	30	164	R putamen
3	50	R precuneus	31	61	L superior occipital gyrus
4	181	cerebellum	32	30	R middle orbitofrontal gyrus
5	47	L angular gyrus	33	42	R postcentral gyrus
6	122	R cingulate gyrus	34	27	L precentral gyrus
7	83	L middle temporal gyrus	35	32	R lateral orbitofrontal gyrus
8	90	R lingual gyrus	36	121	L cingulate gyrus
9	81	L superior temporal gyrus	37	31	L lateral orbitofrontal gyrus
10	91	L fusiform gyrus	38	92	R fusiform gyrus
11	44	R superior parietal gyrus	39	45	L supramarginal gyrus
12	66	R inferior occipital gyrus	40	88	R parahippocampal gyrus
13	87	L parahippocampal gyrus	41	22	R superior frontal gyrus
14	162	R caudate	42	29	L middle orbitofrontal gyrus
15	85	L inferior temporal gyrus	43	68	R cuneus
16	182	brainstem	44	62	R superior occipital gyrus
17	43	L superior parietal gyrus	45	33	L gyrus rectus
18	28	R precentral gyrus	46	48	R angular gyrus
19	23	L middle frontal gyrus	47	64	R middle occipital gyrus
20	89	L lingual gyrus	48	84	R middle temporal gyrus
21	41	L postcentral gyrus	49	49	L precuneus
22	86	R inferior temporal gyrus	50	67	L cuneus
23	163	L putamen	51	161	L caudate
24	26	R inferior frontal gyrus	52	165	L hippocampus
25	102	R insular cortex	53	166	R hippocampus
26	25	L inferior frontal gyrus	54	82	R superior temporal gyrus
27	46	R supramarginal gyrus	55	63	L middle occipital gyrus
28	34	R gyrus rectus	56	101	L insular cortex

Fig. 1.6 Indices of the 56 regions of interest (ROIs): A and B – extracted by the BrainParser software using the LPBA40 brain atlas

1.2.8 Next Generation Sequence (NGS) Analysis

Whole-genome and exome sequencing include essential clues for identifying genes responsible for simple Mendelian inherited disorders. A recent paper proposed methods that can be applied to complex disorders based on population genetics.

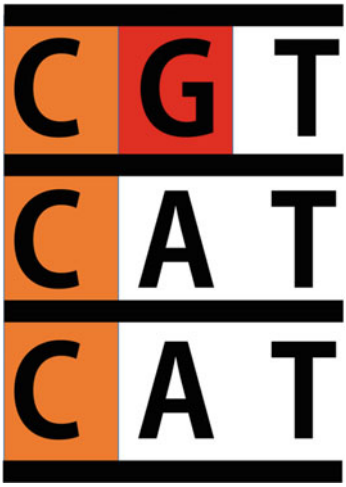


Fig. 1.7 Snippet of the Ebola STR genomic sequence

Next generation sequencing (NGS) technologies include bioinformatics resources to analyze the dense and complex sequence data. The Graphical Pipeline for Computational Genomics (GPCG) performs the computational steps required to analyze NGS data. The GPCG implements flexible workflows for basic sequence alignment, sequence data quality control, single nucleotide polymorphism analysis, copy number variant identification, annotation, and visualization of results. Applications of NGS analysis provide clinical utility for identifying miRNA signatures in diseases. Enabling hypotheses testing about the functional role of variants in the human genome will help to pinpoint the genetic risk factors many diseases (e.g., neuropsychiatric disorders).

1.2.9 Neuroimaging-Genetics

A computational infrastructure for high-throughput neuroimaging-genetics (doi: https://doi.org/10.3389/fninf.2014.00041) facilitates the data aggregation, harmonization, processing, and interpretation of multisource imaging, genomic, clinical, and cognitive data. A unique feature of this architecture is the graphical user interface to the Pipeline environment. Through its client-server architecture, the Pipeline environment provides a graphical user interface for designing, executing, monitoring, validating, and disseminating complex protocols that utilize diverse suites of software tools and web services. These pipeline workflows are represented as portable Extensible Markup Language (XML) objects, which transfer the execution instructions and user specifications from the client user machine to remote pipeline servers for distributed computing. Using Alzheimer's and Parkinson's data, this study provides examples of translational applications using this infrastructure (Figs. 1.8 and 1.9).

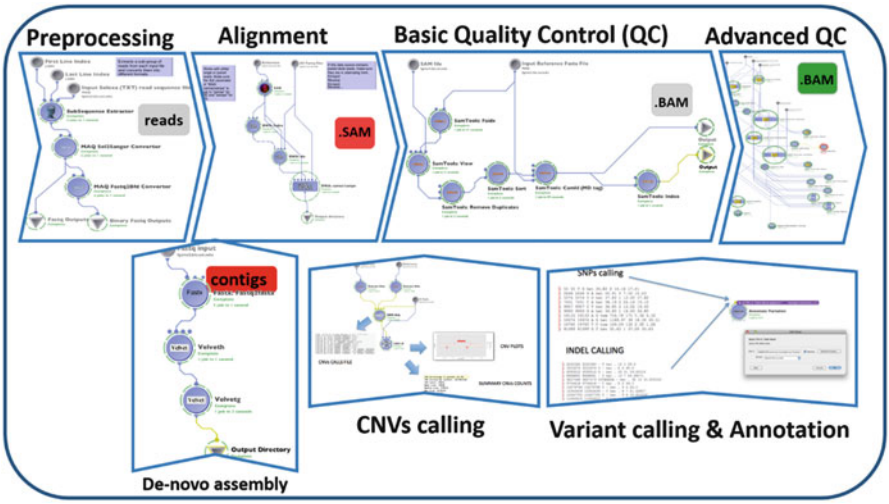


Fig. 1.8 A collage of modules and pipeline workflows from genomic sequence analyses

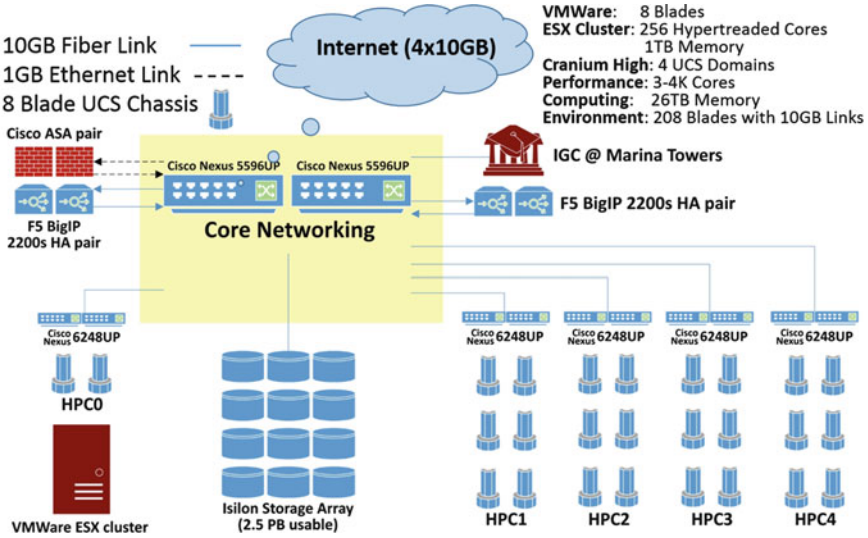


Fig. 1.9 A schematic of a distributed high-throughput computational environment for managing, processing, and visualization of large, complex, and heterogeneous biomedical data

1.3 Common Characteristics of Big (Biomedical and Health) Data

Software developments, student training, utilization of Cloud or IoT (Internet of Things) service platforms, and methodological advances associated with Big Data Discovery Science all present existing opportunities for learners, educators,

Table 1.1 The characteristic six dimensions of Big biomedical and healthcare data

BD dimensions	Necessary techniques, tools, services, and support infrastructure
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multisource	Transfer and joint modeling of disparate elements
Multiscale	Macro to meso- to microscale observations
Incomplete	Reliable management of missing data

researchers, practitioners, and policy makers alike. A review of many biomedical, health informatics, and clinical studies suggests that there are indeed common characteristics of complex big data challenges. For instance, imagine analyzing the observational data of thousands of Parkinson's disease patients, based on tens of thousands of signature biomarkers derived from multisource imaging, genetics, and clinical, physiologic, phenomics, and demographic data elements. IBM had defined the qualitative characteristics of Big Data as 4 Vs: **Volume, Variety, Velocity,** and **Veracity** (there are additional V-qualifiers that can be added).

More recently (PMID:26998309) we defined a constructive characterization of Big Data that clearly identifies the methodological gaps and necessary tools to handle such archives, Table 1.1.

1.4 Data Science

Data science is an emerging new field that (1) is extremely transdisciplinary – bridging between the theoretical, computational, experimental, and biosocial areas; (2) deals with enormous amounts of complex, incongruent, and dynamic data from multiple sources; and (3) aims to develop algorithms, methods, tools, and services capable of ingesting such datasets and generating semiautomated decision support systems. The latter can mine the data for patterns or motifs, predict expected outcomes, suggest clustering or labeling of retrospective or prospective observations, compute data signatures or fingerprints, extract valuable information, and offer evidence-based actionable knowledge. Data science techniques often involve data manipulation (wrangling), data harmonization and aggregation, exploratory or confirmatory data analyses, predictive analytics, validation, and fine-tuning.

1.5 Predictive Analytics

Predictive analytics is the process of utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools and services to represent, interrogate, and interpret complex data. As its name suggests, a core aim of predictive analytics is to forecast trends, predict patterns in the data, or

prognosticate the process behavior either within the range or outside the range of the observed data (e.g., in the future, or at locations where data may not be available). In this context, *process* refers to a natural phenomenon that is being investigated by examining proxy data. Presumably, by collecting and exploring the intrinsic data characteristics, we can track the behavior and unravel the underlying mechanism of the system.

The fundamental goal of predictive analytics is to identify relationships, associations, arrangements, or motifs in the dataset, in terms of space, time, and features (variables) that may prune the dimensionality of the data, i.e., reduce its complexity. Using these process characteristics, predictive analytics may predict unknown outcomes, produce estimations of likelihoods or parameters, generate classification labels, or contribute other aggregate or individualized forecasts. We will discuss how the outcomes of these predictive analytics may be refined, assessed, and compared, e.g., between alternative methods. The underlying assumptions of the specific predictive analytics technique determine its usability, affect the expected accuracy, and guide the (human) actions resulting from the (machine) forecasts. In this textbook, we will discuss supervised and unsupervised, model-based and model-free, classification and regression, as well as deterministic, stochastic, classical, and machine learning-based techniques for predictive analytics. The type of the expected outcome (e.g., binary, polytomous, probability, scalar, vector, tensor, etc.) determines if the predictive analytics strategy provides prediction, forecasting, labeling, likelihoods, grouping, or motifs.

1.6 High-Throughput Big Data Analytics

The pipeline environment provides a large tool chest of software and services that can be integrated, merged, and processed. The Pipeline workflow library and the workflow miner illustrate much of the functionality that is available. Java-based and HTML5 webapp graphical user interfaces (GUIs) provide access to a powerful 4,000 core grid compute server (Fig. 1.10).

1.7 Examples of Data Repositories, Archives, and Services

There are many sources of data available on the Internet. A number of them provide open access to the data based on FAIR (Findable, Accessible, Interoperable, Reusable) principles. Below are examples of open-access data sources that can be used to test the techniques presented in this textbook. We demonstrate the tasks of retrieval, manipulation, processing, analytics, and visualization using example datasets from these archives.

- SOCR Wiki Data, http://wiki.socr.umich.edu/index.php/SOCR_Data
- SOCR Canvas datasets, https://umich.instructure.com/courses/38100/files/folder/data

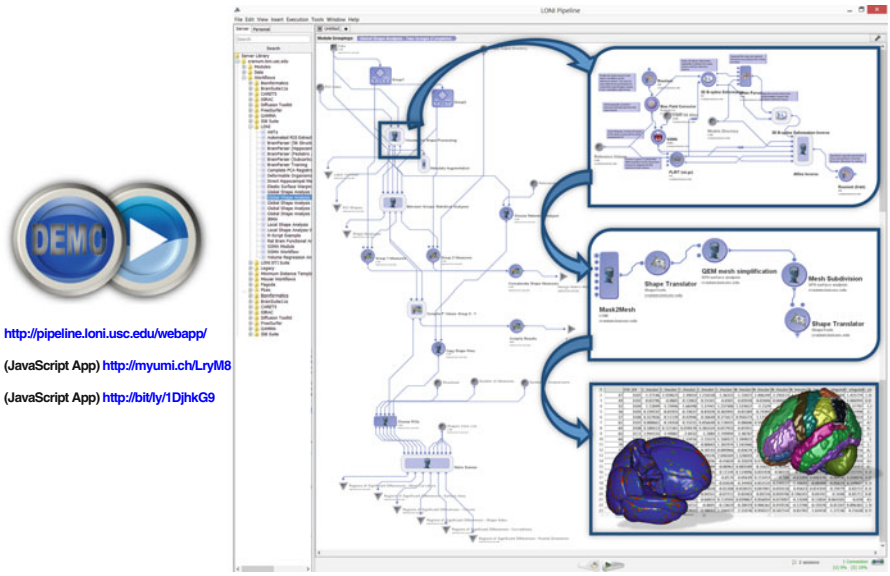


Fig. 1.10 The pipeline environment provides a client-server platform for designing, executing, tracking, sharing, and validating complex data analytic protocols

- SOCR Case-Studies, http://wiki.socr.umich.edu/index.php/SOCR_Data
- XNAT, https://central.xnat.org
- IDA, http://ida.loni.usc.edu
- NIH dbGaP, https://dbgap.ncbi.nlm.nih.gov
- Data.gov (http://data.gov)

1.8 DSPA Expectations

The heterogeneity of data science makes it difficult to identify a precise and complete list of prerequisites guaranteeing deep and lasting understanding of all the presented methods and techniques. However, the reader is strongly encouraged to glance over the preliminary prerequisites, the self-assessment pretest and remediation materials, and the outcome competencies. Throughout this journey, it is useful to *remember the following points*:

- You *don't have to* satisfy all prerequisites, be versed in all mathematical foundations, have substantial statistical analysis expertise, or be an experienced programmer.
- You *don't have to complete all chapters and sections* in the order they appear in the DSPA Topics Flowchart. Completing one, or several, of the suggested pathways may be sufficient for many readers.

- The *DSPA textbook aims* to expand the trainees' horizons, improve understanding, enhance skills, and provide a set of advanced, validated, and practice-oriented code, scripts, and protocols.
- To varying degrees, readers will develop abilities to skillfully utilize the **tool chest** of resources provided in the DSPA textbook. These resources can be revised, improved, customized, expanded, and applied to other biomedicine and biosocial studies, as well as to Big Data predictive analytics challenges in other disciplines.
- The DSPA *materials will challenge most readers*. When *the going gets tough*, seek help, engage with fellow trainees, search for help on the DSPA site and the Internet, communicate via DSPA discussion forum/chat, and review references and supplementary materials. Be proactive! Remember that you will gain, but it will require commitment, prolonged emersion, hard work, and perseverance. If it were easy, its value would be compromised.
- When covering some chapters, some readers may be *underwhelmed or bored*. Feel free to skim over chapters or sections that sound familiar and move forward to the next topic. Still, it is worth trying the corresponding assignments to ensure that you have a firm grasp of the material, and that your technical abilities are sound.
- Although the *return on investment* (e.g., time, effort) may vary between readers, those that complete the DSPA textbook will discover something new, acquire some advanced skills, learn novel data analytic protocols, and may conceive of cutting-edge ideas.
- The complete *R* code (*R* and *Rmd* markdown) for all examples and demonstrations presented in this textbook are available as electronic supplements.
- The author acknowledges that these *materials may be improved*. If you discover problems, typos, errors, inconsistencies, or other problems, please contact us (DSPA.info@umich.edu) to correct, expand, or polish the resources, accordingly. If you have alternative ideas, suggestions for improvements, optimized code, interesting data and case-studies, or any other refinements, please send these along, as well. All suggestions and critiques will be carefully reviewed, and potentially incorporated in revisions or new editions with appropriate credits.