Big Brain Data & Predictive Analytics

Ivo D. Dinov

Statistics Online Computational Resource
Health Behavior and Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science
University of Michigan
www.SOCR.umich.edu

Outline

- AMIA Biomedical Imaging Working Group Driving Challenge:
  * Does Integrative Data Analytics on Biomedical Imaging Bring Us Closer to Precision Medicine?*
- Common characteristics of Big Brain Data
- Data science & predictive analytics
- Case-studies
- Applications to Neurodegenerative Disease
- Data Dashboarding
- Compressive Big Data Analytics (CBDA)
Integrative Data Analytics ↔ Precision Medicine

- **Neurodegeneration:** Structural Neuroimaging in Alzheimer’s Disease illustrates the Big Data challenges in modeling complex neuroscientific data. 808 ADNI subjects, 3 groups: 200 subjects with Alzheimer’s disease (AD), 383 subjects with mild cognitive impairment (MCI), and 225 asymptomatic normal controls (NC). The 80 neuroimaging biomarkers and 80 highly-associated SNPs.

- **Information Complexity** – large, incongruent, time-varying data

- **Precision Medicine** – customized medical decisions, clinical practice, treatments, or healthcare products to individual patients

- **Individual vs. Population Studies** – inductive (discriminative) vs. deductive (generative) models for clinical decision support

- **Tools** – molecular diagnostics, imaging, clinical, wearables, analytics, …
Characteristics of Big Biomed Data

IBM Big Data 4V’s: Volume, Variety, Velocity & Veracity

<table>
<thead>
<tr>
<th>Big Bio Data Dimensions</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Harvesting and management of vast amounts of data</td>
</tr>
<tr>
<td>Complexity</td>
<td>Wranglers for dealing with heterogeneous data</td>
</tr>
<tr>
<td>Incongruency</td>
<td>Tools for data harmonization and aggregation</td>
</tr>
<tr>
<td>Multi-source</td>
<td>Transfer and joint modeling of disparate elements</td>
</tr>
<tr>
<td>Multi-scale</td>
<td>Macro to meso to micro scale observations</td>
</tr>
<tr>
<td>Incomplete</td>
<td>Reliable management of missing data</td>
</tr>
</tbody>
</table>

Example: analyzing observational data of 1,000’s Parkinson’s disease patients based on 10,000’s signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements.

Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers.

Dinov, et al. (2016) PMID:26918190
Data science & predictive analytics

- **Data science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and biosocial areas. Deals with enormous amounts of complex, incongruent and dynamic data from multiple sources. Aims to develop algorithms, methods, tools and services capable of ingesting such datasets and supplying semi-automated decision support systems.

- **Predictive analytics**: utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools and web-services to represent, interrogate and interpret complex data. Aims to forecast trends, cluster 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).

http://DSPA.predictive.space

---

<table>
<thead>
<tr>
<th>Big Data</th>
<th>Information</th>
<th>Knowledge</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Observations</td>
<td>Processed Data</td>
<td>Maps, Models</td>
<td>Actionable Decisions</td>
</tr>
<tr>
<td>Data Aggregation</td>
<td>Data Fusion</td>
<td>Causal Inference</td>
<td>Treatment Regimens</td>
</tr>
<tr>
<td>Data Scrubbing</td>
<td>Summary Stats</td>
<td>Networks, Analytics</td>
<td>Forecasts, Predictions</td>
</tr>
<tr>
<td>Semantic-Mapping</td>
<td>Derived Biomarkers</td>
<td>Linkages, Associations</td>
<td>Healthcare Outcomes</td>
</tr>
</tbody>
</table>

Dinov, et al. (2016) PMID:26918190
Case-Studies – ALS

- Identify predictive classifiers to detect, track and prognosticate 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

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

Case-Studies – ALS

- Detect, track and prognosticate the progression of ALS
- Predict of adverse events based on subject phenotype and 0-3 month clinical assessment changes

**Methods**

- Linear Regression
- Random Forest
- BART
- SuperLearner

<table>
<thead>
<tr>
<th>Methods</th>
<th>Linear Regression</th>
<th>Random Forest</th>
<th>BART</th>
<th>SuperLearner</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.081</td>
<td>0.174</td>
<td><strong>0.225</strong></td>
<td>0.178</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.619</td>
<td>0.587</td>
<td><strong>0.568</strong></td>
<td>0.585</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.298</td>
<td>0.434</td>
<td><strong>0.485</strong></td>
<td>0.447</td>
</tr>
</tbody>
</table>
Case-Studies – Parkinson’s Disease

- Predict the clinical diagnosis of patients using all available data (with and without the 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

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPMI Archive</td>
<td>Demographics: age, medical history, sex. Clinical data: physical, verbal learning and language, neurological and olfactory, UPSIT, UPDRS scores, ADL, GDS-15, ... Imaging data: structural MRI. Genetics data: APOE genotypes e2/e3 Cohorts: Group 1 = (PD Subjects), N₁ = 263; Group 2 = (PD Subjects with Scans without Evidence of a Dopaminergic Deficit (SWEDD)), N₂ = 40; Group 3 = (Control Subjects), N₃ = 127.</td>
<td>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</td>
</tr>
</tbody>
</table>

Predictive Big Data Analytics: Applications to Parkinson’s Disease

![Diagram of Predictive Big Data Analytics](image)

- Data Representation Fusion Harmonization Aggregation
- Cleaning Imputation Wrangling Synthesis
- Rebalancing, Feature Selection
- Model-based, Model-free, Classification
- Prediction, Statistical Validation, Inference

Varplot
- Critical predictive data elements (Y-axis)
- Their impact scores (X-axis)

AdaBoost classifier for Controls vs. Patients prediction

<table>
<thead>
<tr>
<th>ML classifier</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive predictive value</th>
<th>Negative predictive value</th>
<th>Log odds ratio (LOR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaBoost</td>
<td>0.996324</td>
<td>0.992645</td>
<td>0.990822</td>
<td>0.9980392</td>
<td>0.9948097</td>
<td>11.4892058</td>
</tr>
<tr>
<td>SVM</td>
<td>0.985294</td>
<td>0.994140</td>
<td>0.977431</td>
<td>0.9750958</td>
<td>0.9946996</td>
<td>8.902166</td>
</tr>
</tbody>
</table>

Dinov, et al., (2016) PMID:27494614
Case-Studies – General Populations

- UK Biobank – discriminate between HC, single and multiple comorbid conditions
- Predict likelihoods of various developmental or aging disorders
- Forecast cancer

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK Biobank</td>
<td>Demographics: &gt; 500K cases Clinical data: &gt; 4K features Imaging data: T1, resting-state fMRI, task fMRI, T2, FLAIR, dMRI, SWI Genetics data</td>
<td>The longitudinal archive of the UK population (NHS)</td>
</tr>
</tbody>
</table>

| Case-Studies – UK Biobank - Complexities |

Missing Clinical & Phenotypic data for 10K subjects with sMRI, for which we computed 1.5K derived neuroimaging biomarkers.

Including only features observed >30%

$(9,914 \times 1,475)$
Case-Studies – UK Biobank – NI Biomarkers

Case-Studies – UK Biobank – Successes/Failures
Case-Studies – UK Biobank – Derived NI Biomarkers

- Subjects=9,915
- Clinical Features=4,100
- Derived NI Biomarkers=3,297

lh_aparc.a2009s_volume__lh_S_oc.temp_med.Lingual_volume
lh_aparc.a2009s_volume__lh_S_orbital_lateral_volume
lh_aparc.a2009s_volume__lh_S_orbital_med.olfact_volume
lh_aparc.a2009s_volume__lh_S_orbital.H_Shaped_volume
lh_aparc.a2009s_volume__lh_S_parieto_occipital_volume
lh_aparc.a2009s_volume__lh_S_pericallosal_volume
lh_aparc.a2009s_volume__lh_S_postcentral_volume
lh_aparc.a2009s_volume__lh_S_precentral.inf.part_volume
lh_aparc.a2009s_volume__lh_S_precentral.sup.part_volume
lh_aparc.a2009s_volume__lh_S_suborbital_volume
lh_aparc.a2009s_volume__lh_S_subparietal_volume
lh_aparc.a2009s_volume__lh_S_temporal_inf_volume
lh_aparc.a2009s_volume__lh_S_temporal_sup_volume
lh_aparc.a2009s_volume__lh_S_temporal_transverse_volume
lh_aparc.DKTatlas_area__lh_caudalanteriorcingulate_area
lh_aparc.DKTatlas_area__lh_caudalmiddlefrontal_area
lh_aparc.DKTatlas_area__lh_cuneus_area
lh_aparc.DKTatlas_area__lh_entorhinal_area
lh_aparc.DKTatlas_area__lh_fusiform_area
lh_aparc.DKTatlas_area__lh_inferiorparietal_area
lh_aparc.DKTatlas_area__lh_inferiortemporal_area
lh_aparc.DKTatlas_area__lh_inferior temporal_area

Case-Studies – UK Biobank – Unstructured Data

- n=11K cases with clinical diagnosis (notes)
- Computed Corpus Fingerprint – signature tensor
- LASSO regularize modeling (outcome death in 12 months)
  - max AUC = 0.5657 (poor)
- External Validation using 8 patients from the SOCR Hospital Admissions Data

```
# Initial
<<DocumentTermMatrix (documents: 11134, terms: 2566)
## Non-sparse entries: 607,58/309,086
## Sparsity: 100%
## Maximal term length: 25
## Weighting: term frequency (tf)
```

```
# Post-processing
<<DocumentTermMatrix (documents: 364, terms: 64)
## Non-sparse entries: 1,421/23,155
## Sparsity: 94%
## Maximal term length: 16
## Weighting: term frequency (tf)
```

```
# Training + Cross-Validation Testing on Hospital Admissions Data

table(TestY, predTestPatients )
## TestY | 0 | 1
## 0 | 9127 | 45
## 1 | 897 | 16
```

…
End-to-end Pipeline Workflow Solutions


SOCR Big Data Dashboard

http://socr.umich.edu/HTML5/Dashboard

- Web-service combining and integrating multi-source socioeconomic and medical datasets
- Big data analytic processing
- Interface for exploratory navigation, manipulation and visualization
- Adding/removing of visual queries and interactive exploration of multivariate associations
- Powerful HTML5 technology enabling mobile on-demand computing

Husain, et al., 2015, PMID:26236573
SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion

http://socr.umich.edu/HTML5/Dashboard

SOCR Dashboard (Exploratory Big Data Analytics): Associations
Compressive Big Data Analytics (CBDA)

- Foundations for Compressive Big Data Analytics (CBDA)
  - Iteratively generate random (sub)samples from the Big Data collection
  - Then, using classical techniques to obtain model-based or non-parametric inference based on the sample
  - Next, compute likelihood estimates (e.g., probability values quantifying effects, relations, sizes)
  - Repeat – the process continues iteratively until a criterion is met – the (re)sampling and inference steps many times (with or without using the results of previous iterations as priors for subsequent steps)

Dinov, 2016, PMID: 26998309
FAIR Data & Open-Science Principles

- Share resources
- Collaborate
- Permissive licenses (e.g., LGPL/CC-BY)
- Project management (e.g., GitHub/Jira)
- Open-access pubs
- Public-private partnerships
- Co-mentoring of trainees
- Effective transdisciplinary methods
- Resource Interoperability
- Result Reproducibility

Clinical Decision Support

- Hospital Admissions
  - Survival Inference and Clinical outcome forecasting using hospital admissions data (N~60K and k=9):
    - Admission_Length: Duration of hospital stay (in days)
    - Death: Indicator of Death (1) or survival (0)
    - ...
    - Demographics & (Human-labeled) Diagnoses

VarImpPlot: RF Survival Prediction

<table>
<thead>
<tr>
<th>VarImPlot</th>
<th>RF Survival Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdmissionLengthDays</td>
<td>hosAdmissions_RFmodel(pred 0 1 0.9937 963 1.722 15)</td>
</tr>
<tr>
<td>Admission_Type</td>
<td>Accuracy: 0.8568759</td>
</tr>
<tr>
<td>Race</td>
<td>95% CI: (0.85, 0.86) Sensitivity: 0.9322638 Specificity: 0.1355476</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.9116514 Neg Pred Value</td>
</tr>
</tbody>
</table>

NeuralNet Deep Learning: Survival Probability (N=11,773)

Test/Validation Cases
Clinical Decision Support

- Population-wide Study of Health and Disease
  National Health Service (NHS)
  Longitudinal Data (N>1M, k>4,300)

Clinical Decision Support

- Personalized medicine – Traumatic Brain Injury (TBI)
Acknowledgments

Funding
NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30 AG053760
NSF: 1734853, 1636840, 1416953, 0716055, 1023115
The Elsie Andresen Fiske Research Fund

Collaborators
- SOCR: Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Matt Leventhal, Ashwini Khare, Rami Elkest, Abhishek Chowdhury, Patrick Tan, Gary Chan, Andy Foglia, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Jingzhu Xu, Nellie Ponarul, Ming Tang, Asiyah Lin, Nicolas Christou, Hanbo Sun, Tuo Wang
- LONI: Arthur Toga, Roger Woods, Jack Van Horn, Zhuowen Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Anand Joshi, Shantanu Joshi, Paul Thompson, Luminita Vese, Stan Osher, Stefano Soatto, Seok Moon, Junming Li, Young Sung, Carl Kesselman, Fabio Macciardi, Federica Torri
- UMich MIDAS/MNORC/AD/PD Centers: Cathie Spino, Chuck Burant, Ben Hampstead, Stephen Goutman, Stephen Strobbe, Hiroko Dodge, Hank Paulson, Bill Dauer, Brian Athey

Slides available online:
Google "SOCR News"

http://SOCR.umich.edu

Demo(s)?
- Complex DB Search, retrieval (IDA)
- Multidimensional data visualization (MotionCharts, BrainViewer, R)
- Distributed high-throughput pipeline workflow computing
- SOCRAT Framework
- Data Dashboard
- Education and Training Resources
  - Probability and Statistics Ebook (EBook)
  - Scientific Methods for Health Sciences (SMHS)
  - Data Science and Predictive Analytics (DSPA) MOOC
  - SOCR Tools (distribution calculators, charts, modeler, analyses, experiments)
- Compressive Big Data Analytics (CBDA)