Open Science & Data Sharing

Ivo D. Dinov

Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science

University of Michigan

http://SOCR.umich.edu

Slides Online: "SOCR News"

Outline

- ❑ Pillars of Open-Science
- ❑ Rationale (Pros & Cons)
- ❑ Big Data Sharing
- ❑ *DataSifter: Statistical obfuscation*
- ❑ Case-studies
  - ❑ Applications to Neurodegenerative Disease (Udall/MADC)
  - ❑ Parkinson's Disease (PD)
  - ❑ Population Census-like Neuroscience (UKBB)

Pillars of Open-Science

Sources: Characteristics of Big Biomed Data

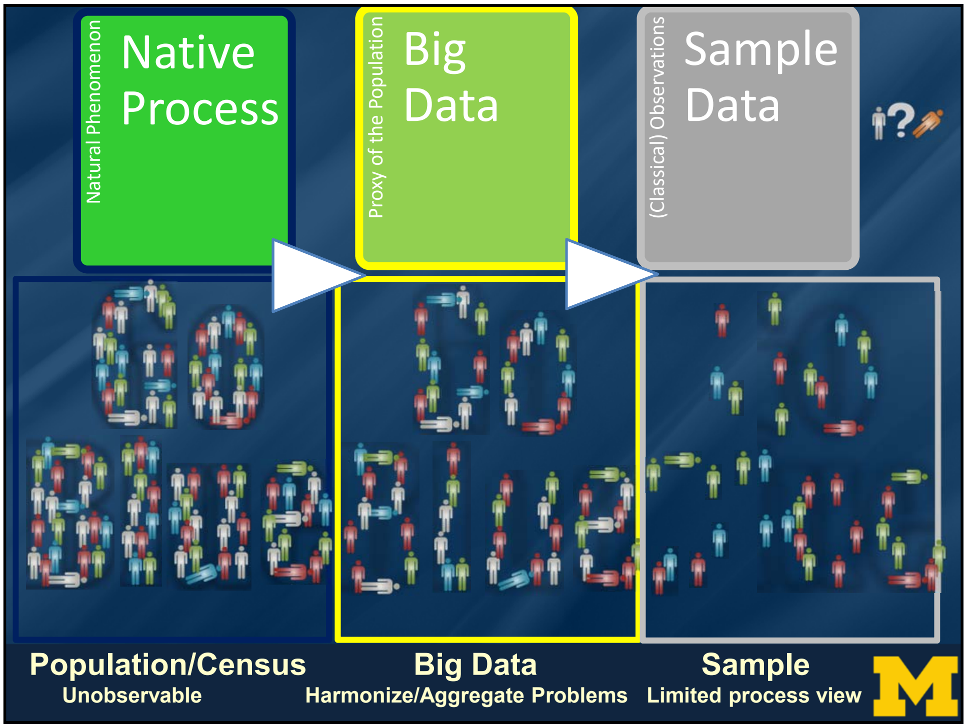
IBM Big Data 4V's: Volume, Variety, Velocity & Veracity

Big Bio Data Dimensions	Tools
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro to meso to micro scale observations
Time	Techniques accounting for longitudinal patterns in the data
Incomplete	Reliable management of missing data

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 (2016) GigaScience   Dinov (2018) Springer



From 23 … to … 2^{23}

- Data Science: 1798 vs. 2019
- In the 18th century, Henry Cavendish used just 23 observations to answer a fundamental question – "What is the Mass of the Earth?" He estimated very accurately the mean density of the Earth/H_2O (5.483 ± 0.1904 g/cm^3)
- In the 21st century to achieve the same scientific impact, matching the reliability and the precision of the Cavendish's 18th century prediction, requires a monumental community effort using massive and complex information perhaps on the order of 2^{23} bytes
- Scalability and Compression
  (per Gerald Friedland/Berkeley): 23 ➔ 10M

Cavendish (1798) Philosophical Transactions of the Royal Society of London | Dinov (2016) JSMI

Big Data	Information	Knowledge	Action
Raw Observations	Processed Data	Maps, Models	Actionable Decisions
Data Aggregation	Data Fusion	Causal Inference	Treatment Regimens
Data Scrubbing	Summary Stats	Networks, Analytics	Forecasts, Predictions
Semantic-Mapping	Derived Biomarkers	Linkages, Associations	Healthcare Outcomes

Dinov, et al. (2016) PMID:26918190

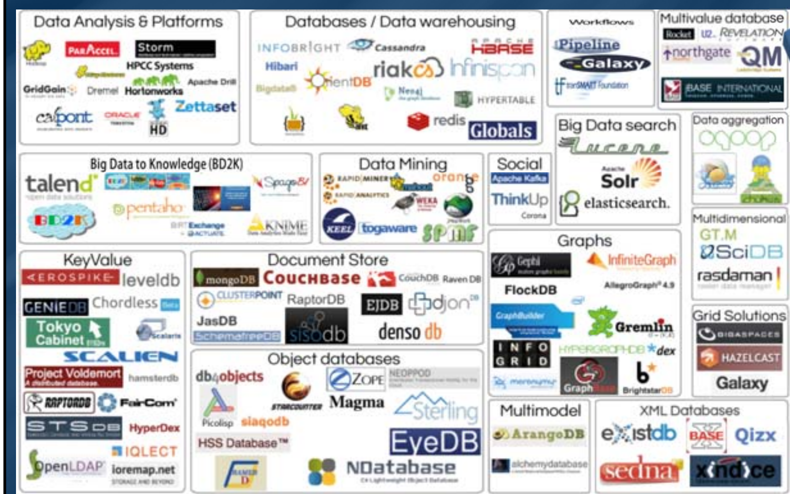
Why is FAIR Data Sharing Important?

- Optimum resource utilization (low cost, high efficiency / policy, security, processing complexity)
- Democratization of the scientific discovery process
- Enhanced inference (e.g., coverage of rare events, increase of stat power)
- Increase of Kryder's Law (Data volume) ≫ Moore's Law (Compute power)
- Exponential decay of data-value
- Incents innovation, transdisciplinary collaborations, and knowledge dissemination
- …

FAIR = Findable + Accessible + Interoperable + Reusable

Infrastructure: Cloud Ecosystem

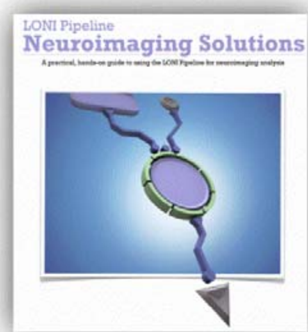
Cloud Services



http://socr.umich.edu/docs/BD2K/BigDataResourceome.html

Infrastructure: Cranium/Pipeline

Tools



http://Pipeline.loni.usc.edu

Dinov, et al. (2013) Brain Imaging and Behavior | Dinov, et al. (2014) Front. NeuroInfo.

Findings: OA Pubs/Sharing

- OA Pubs
  - https://en.wikipedia.org/wiki/Open_access
  - https://arxiv.org | https://www.biorxiv.org
  - Blogs (e.g., https://TerryTao.wordpress.com)
- Cloud Services
  - Computing (e.g., Azure, Google, AWS)
  - Storage
  - ICT (information and communication technologies)
- SW
  - https://GitHub.com (e.g., https://github.com/SOCR)
  - http://Cran.r-project.org | Jupyter.org | Rmarkdown.rstudio.com
  - E.g., http://DSPA.predictive.space
- Licensing
  - https://www.gnu.org/licenses (e.g., http://socr.umich.edu/html/SOCR_CitingLicense.html)

Findings: Open Science Career Assessment Matrix

Open Science activities	Metrics: Possible evaluation criteria
RESEARCH OUTPUT	
Research activity	Pushing forward the boundaries of open science as a research topic
Publications	Publishing in open access journals Self-archiving in open access repositories
Datasets and research results	Using the FAIR data principles Adopting quality standards in open data management and open datasets Making use of open data from other researchers
Open source	Using open source software and other open tools Developing new software and tools that are open to other users
Funding	Securing funding for open science activities
RESEARCH PROCESS	
Stakeholder engagement / citizen science	Actively engaging society and research users in the research process Sharing provisional research results with stakeholders through open platforms (e.g. Arxiv, Figshare, OverLeaf) Involving stakeholders in peer review processes
Collaboration and Interdisciplinarity	Widening participation in research through open collaborative projects Engaging in team science through diverse cross-disciplinary teams
Research integrity	Being aware of the ethical and legal issues relating to data sharing, confidentiality, attribution and environmental impact of open science activities Fully recognizing the contribution of others in research projects, including collaborators, co-authors, citizens, open data providers
Risk management	Taking account of the risks involved in open science

Declaration on Research Assessment (DORA) | https://sfdora.org/good-practices/funders

Findings: Open Science Career Assessment Matrix

SERVICE & LEADERSHIP	
Leadership	Developing a vision and strategy on how to integrate OS practices in the normal practice of doing research Driving policy and practice in open science Being a role model in practicing open science
Academic standing	Developing an international or national profile for open science activities Contributing as editor or advisor for open science journals or bodies
Peer review	Contributing to open peer review processes Examining or assessing open research
Networking	Participating in national and international networks relating to open science
RESEARCH IMPACT	
Communication and Dissemination	Participating in public engagement activities Sharing research results through non-academic dissemination channels Translating research into a language suitable for public understanding
IP (patents, licenses)	Being knowledgeable on the legal and ethical issues relating to IPR Transferring IP to the wider economy
Societal impact	Evidence of use of research by societal groups Recognition from societal groups or for societal activities. **h-index, i10-index, sharing-index, other quant metrics of impact**
Knowledge exchange	Engaging in open innovation with partners beyond academia
TEACHING & SUPERVISION	
Teaching	Training other researchers in open science principles and methods Developing curricula and programs in open science methods, including open science data management Raising awareness and understanding in open science in undergraduate and masters' programs
Mentoring	Mentoring and encouraging others in developing their open science capabilities
Supervision	Supporting early stage researchers to adopt an open science approach
PROFESSIONAL EXPERIENCE	
Continuing professional development	Investing in own professional development to build open science capabilities
Project management	Successfully delivering open science projects involving diverse research teams
Personal qualities	Demonstrating the personal qualities to engage society and research users with open science Showing the flexibility and perseverance to respond to the challenges of conducting open science

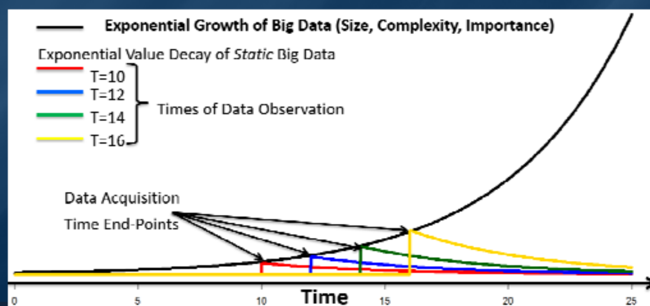
Declaration on Research Assessment (DORA) | **https://sfdora.org/good-practices/funders**

Rationale for Open Science (Cons)

- Journals impact factor (compared to pay-per-view journals, OA are newer)
- *Predatory* science (dubious quality, profit-centric, spam camouflage)
- Discovery is easy, but validity/utility of the science or tools may be difficult to evaluate *en masse*
- Extra work may be required by all scholars to sift through and identify appropriate materials
- Ambiguity of usage-rights/copyrights/licenses
- Democratization and socialization of science may suffer from some of the same downsides as social-networks
- Is science *competitive* or *collaborative*? Is it a *zero-sum* enterprise?

Rationale for Open Science (Pros)

- We are always stronger together
- Long-term sustainability prefers diversity
- Optimized investments, career advancement, impact & cost-efficiency
- Expeditious discovery, innovation, productization & impact
- Rapid devaluation of data-hoarding, clandescent science, knowledge obfuscation
- …

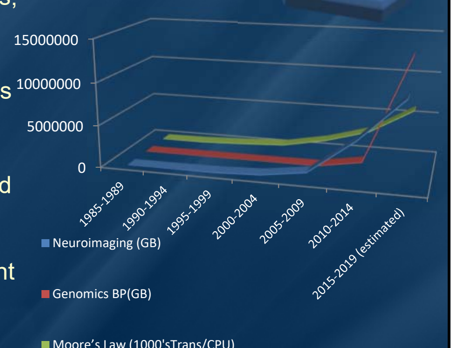


https://www.aaas.org/news/big-data-blog-part-v-interview-dr-ivo-dinov

Rationale for Open Science: Kryder vs. Moore

Collabs

- Moore's law = the expectation that our computational capabilities, specifically the number of transistors on integrated circuits, doubles approximately every 18-24 months.
- Kryder's law = the volume of data, in terms of disk storage capacity, is doubling every 14-18 months.
- **Kryder ≫ Moore**: Although both laws yield exponential growth, data volume is increasing at a faster pace. Thus, there are clear interests and needs for significant private, public and government engagement in opening, managing, processing, interrogating and interpreting the information content of Big Data.



Dinov (2016) SMSI | https://www.aaas.org/news/big-data-blog-part-v-interview-dr-ivo-dinov

DataSifter

- DataSifter is an iterative statistical computing approach that provides the data-governors controlled manipulation of the trade-off between sensitive information obfuscation and preservation of the joint distribution.
- The DataSifter is designed to satisfy data requests from pilot study investigators focused on specific target populations.
- Iteratively, the DataSifter stochastically identifies candidate entries, cases as well as features, and subsequently selects, nullifies, and imputes the chosen elements. This statistical-obfuscation process relies heavily on nonparametric multivariate imputation to preserve the information content of the complex data.

http://DataSifter.org US patent #16/051,881 Marino, *et al.*, *JSCS* (2019)

DataSifter

- A detailed description and *dataSifter()* R method implementation are available on our GitHub repository (https://github.com/SOCR/DataSifter).
- Data-sifting different data archives requires customized parameter management. Five specific parameters mediate the balance between protection of sensitive information and signal energy preservation.

Obfuscation level	$0 \leq \eta = \eta(k_0 + k_1 + k_2 + k_3 + k_4) \leq 1$				
	k_0	k_1	k_2	k_3	k_4
None	0	0	0	0	**0**
Small	0	0.05	1	0.1	**0.01**
Medium	1	0.25	2	0.6	**0.05**
Large	1	0.4	5	0.8	**0.2**
Indep	Output synthetic data with independent features				

k_0: A Boolean; obfuscate the unstructured features?

k_1: proportion of artificial missing data values that should be introduced

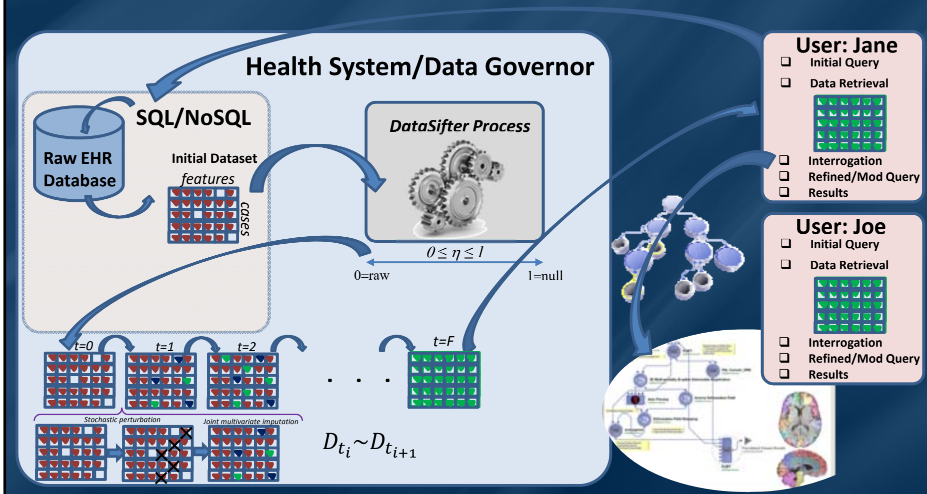
k_2: The number of times to iterate

k_3: The fraction of structured features to be obfuscated in all the cases

k_4: The fraction of closest subjects to be considered as neighbours of a given subject

http://DataSifter.org US patent #16/051,881 Marino, *et al.*, *JSCS* (2019)

DataSifter



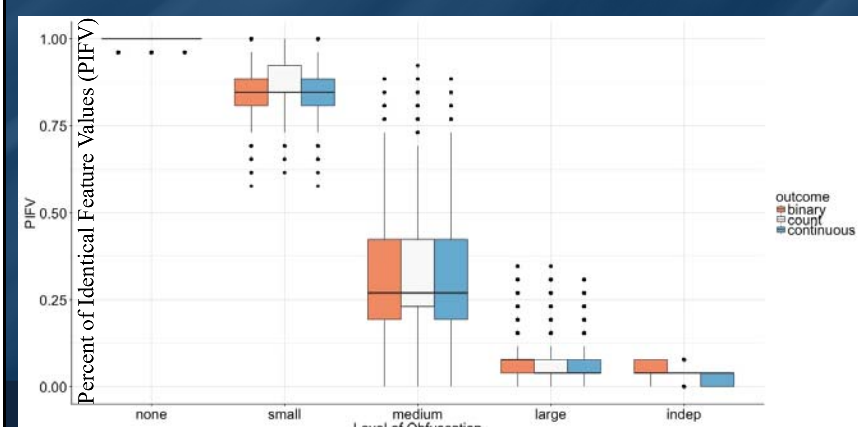
http://*DataSifter*.org US patent #16/051,881 Marino, *et al.*, *JSCS* (2019)

DataSifter Validation

I. Protection of sensitive information (privacy)

PIFV under Different Privacy Levels. Binary outcome refers to the first experiment; Count refers to the second experiment; Continuous refers to the third experiment.

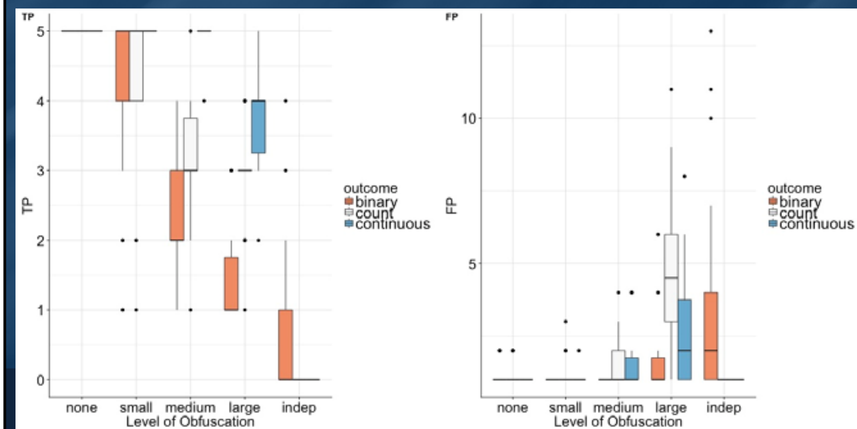
Each box represents 30 different "sifted" data or 30,000 "sifted" cases.



DataSifter Validation

II. Preserving utility information of the original dataset

Logistic Model with Elastic Net Signal Capturing Ability. TP is the number of true signals (total true predictors = 5) captured by the model. FP is the number of null signals that the model has falsely selected (total null signals=20).



DataSifter Validation

III. Clinical Data Application: Using DataSifter to Obfuscate the ABIDE Data

Comparing the Original and "Sifted" Data for the 22nd ABIDE Subject

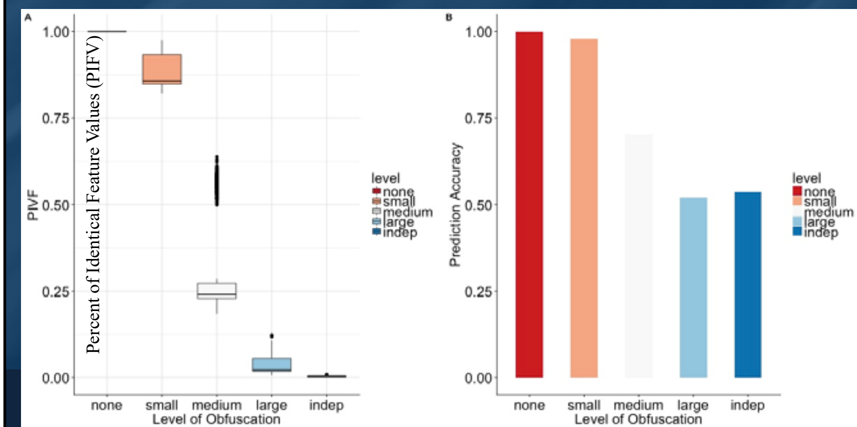
η	Output	Sex	Age	Acquisition Plane	IQ	thick_std_ctx.lh.cuneus	curv_ind_ctx_lh_G_front_inf.Triangul	gaus_curv_ctx.lh.medialorbitofrontal	curv_ind_ctx_lh_S_interm_prim.Jensen
original	Autism	M	31.7	Sagittal	131	0.475	2.1	0.315	NA
none	Autism	M	31.7	Sagittal	131	0.475	2.1	0.315	0.51
small	Autism	M	31.7	Sagittal	131	0.475	2.1	0.315	0.4589
medium	Autism	M	31.7	Sagittal	111	0.548	2.85	0.315	0.463
large	Control	M	18.2	Sagittal	104	0.5347	3.198	0.1625	0.4524
indep	Control	M	15.4	Coronal	104	0.4842	3.383	0.1079	1.002

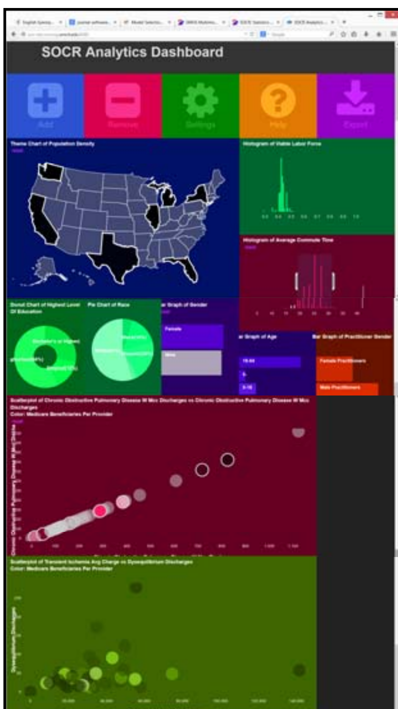
Autism Brain Imaging Data Exchange (ABIDE) case-study

DataSifter Validation

IV. Clinical Data Application: Using DataSifter to Obfuscate the ABIDE Data

PIFVs for ABIDE under different levels of DataSifter obfuscations.
Each box represents 1098 subjects among the ABIDE sub-cohort
Random forest prediction of binary clinical outcome - autism spectrum disorder – (ASD) status (ASD vs. control)





SOCR Big Data Dashboard

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

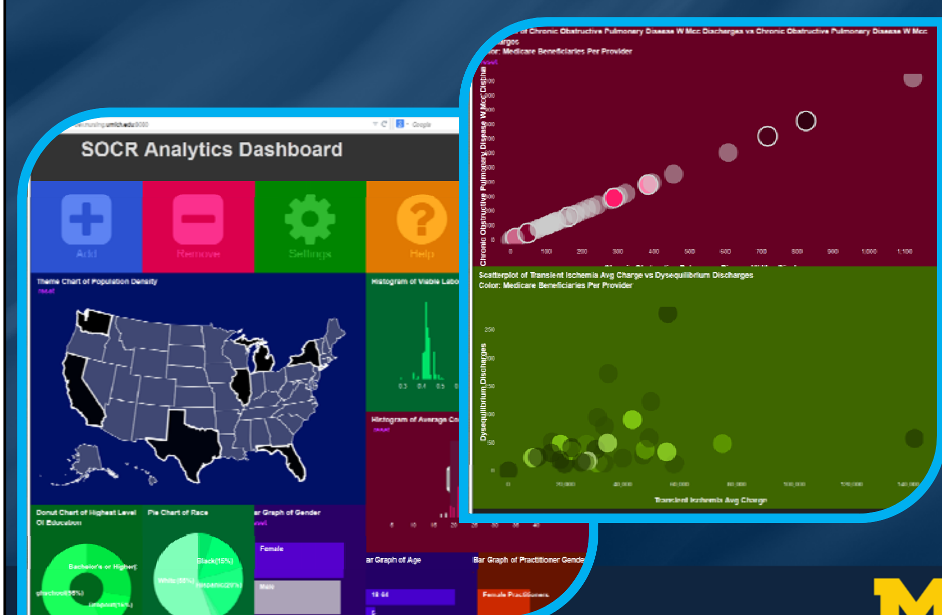
- Web-service combining and integrating multi-source socioeconomic and medical datasets
- Big data analytic processing
- Interface for exploratory navigation, manipulation and visualization
- Adding/removing of visual queries and interactive exploration of multivariate associations
- Powerful HTML5 technology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

Husain, et al., 2015, PMID:26236573

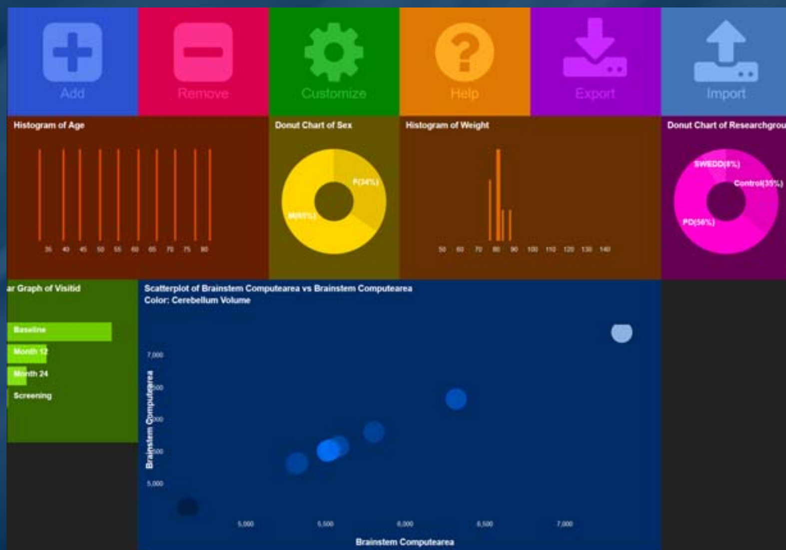
SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion

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

SOCR Dashboard (Exploratory Big Data Analytics): Associations



SOCR Dashboard (Exploratory Big Data Analytics): Udall PD Data



http://wiki.socr.umich.edu/index.php/SOCR_Videos_Dashboard

Data Science & Predictive Analytics

- **Data Science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and applied 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**: process utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools, and distributed 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

Dinov (2018) Springer

Case-Studies – ALS

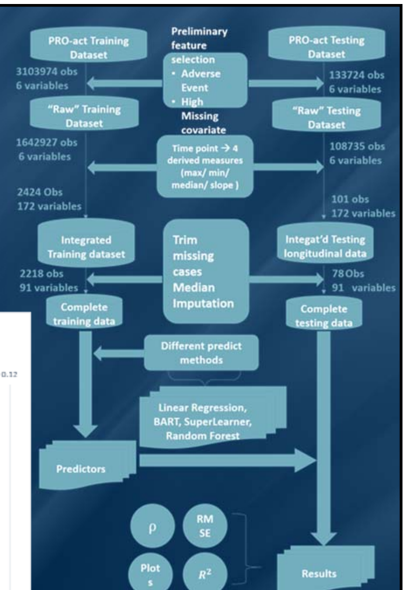
- ❑ Identify predictive classifiers to detect, track and prognosticate the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function)
- ❑ Provide a decision tree prediction of adverse events based on subject phenotype and 0-3 month clinical assessment changes

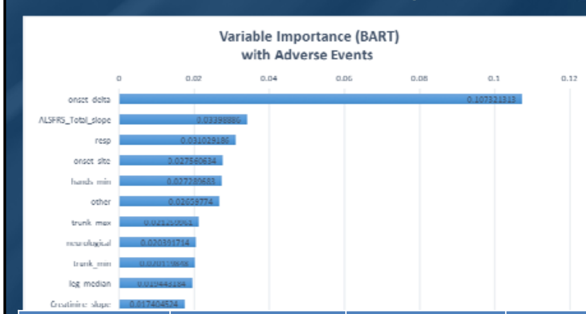
Data Source	Sample Size/Data Type	Summary
ProAct Archive	Over 100 variables are recorded for all subjects including: Demographics: age, race, medical history, sex; Clinical data: **Amyotrophic Lateral Sclerosis** Functional Rating Scale (ALSFRS), adverse events, onset_delta, onset_site, drugs use (riluzole) 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 used for processing, modeling and analysis	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)

Huang et al. (2017) PLoS | *Tang, et al. (2018), Neuroinformatics*

Case-Studies – ALS

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





Methods	Linear Regression	Random Forest	BART	SuperLearner
R-squared	0.081	0.174	**0.225**	0.178
RMSE	0.619	0.587	**0.568**	0.585
Correlation	0.298	0.434	**0.485**	0.447

Case-Studies – ALS

- **Main Finding**: predicting univariate clinical outcomes may be challenging, the (information energy) signal is very weak. We can cluster ALS patients and generate evidence-based ALS hypotheses about the complex interactions of *multivariate factors*
- **Classification vs. Clustering**:
  - Classifying univariate clinical outcomes using the PRO-ACT data yields only marginal accuracy (about 70%).
  - Unsupervised clustering into sub-groups generates stable, reliable and consistent computable phenotypes whose explication requires interpretation of multivariate sets of features

Data Representation Fusion Harmonization Aggregation → Cleaning Imputation Wrangling Synthesis → Model-based, Model-free, Classification, Clustering, Inference

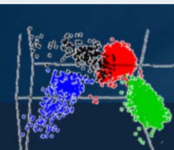
Cluster	Consistency	Variance	Cluster-Size	Silhouette
1	1	0	565	0.58
2	0.986	0.018	427	0.63
3	0.956	0.053	699	0.5
4	0.985	0.018	733	0.5

Tang, et al. (2018), Neuroinformatics

Case-Studies – ALS – Explicating Clustering

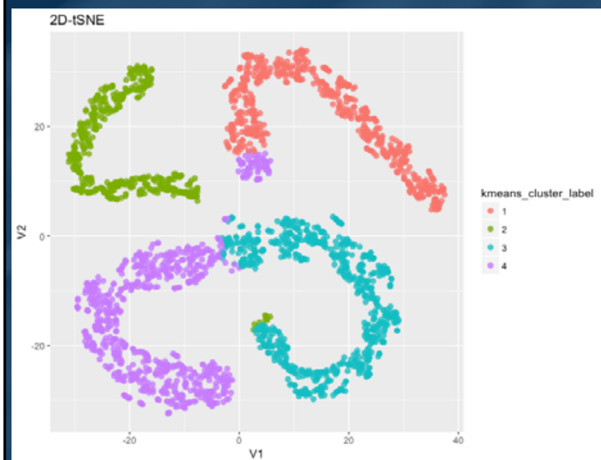
Feature Name	Between Cluster Significant Differences					
	1-2	1-3	1-4	2-3	2-4	3-4
onset_site			1	1	1	
onset_delta.x	1	1	1	1	1	1
onset_delta.y	1	1	1	1		
Red.Blood.Cells..RBC._min	1			1	1	
Red.Blood.Cells..RBC._median	1			1	1	
Red.Blood.Cells..RBC._slope				1	1	
Q4_Handwriting_max		1		1	1	
Q4_Handwriting_min		1		1	1	
Q4_Handwriting_median				1	1	
Q9_Climbing_Stairs_max		1	1	1	1	
Q9_Climbing_Stairs_min		1	1	1	1	
Q9_Climbing_Stairs_median		1	1	1	1	
Q9_Climbing_Stairs_slope	1			1		
Q8_Walking_max		1	1	1	1	
Q8_Walking_min		1	1	1	1	
Q8_Walking_median		1	1	1	1	
trunk_max		1	1	1	1	1
trunk_min		1	1	1	1	
trunk_median		1	1	1	1	
Protein_slope	1			1	1	
Creatinine_max		1	1	1		
Creatinine_min		1	1	1	1	
Creatinine_median		1	1	1	1	
respiratory_rate_max	1			1	1	
hands_max		1		1	1	
hands_min		1		1	1	
hands_median		1		1	1	
Q6_Dressing_and_Hygiene_max		1	1	1	1	
Q6_Dressing_and_Hygiene_min		1		1	1	
Q6_Dressing_and_Hygiene_median		1	1	1	1	
Q7_Turning_in_Bed_max		1	1	1	1	
Q7_Turning_in_Bed_min		1		1	1	
Q7_Turning_in_Bed_median		1	1	1	1	
Sodium_slope	1			1	1	
ALSFRS_Total_max		1		1	1	1
ALSFRS_Total_min		1		1		1
ALSFRS_Total_median		1		1	1	1
ALSFRS_Total_s				1	1	
Hematocrit_max	1			1	1	
Hematocrit_min	1			1	1	
Hematocrit_median	1			1	1	
leg_max		1	1	1	1	
leg_min			1	1	1	
leg_median		1	1	1	1	
mouth_min			1		1	1
Absolute.Basophil.Count_max	1			1	1	
Absolute.Basophil.Count_min	1			1	1	
Absolute.Basophil.Count_median	1			1	1	
Absolute.Basophil.Count_slope	1			1	1	
Absolute.Eosinophil.Count_max	1			1	1	
Absolute.Eosinophil.Count_median	1			1	1	
Absolute.Eosinophil.Count_slope	1			1	1	
Absolute.Lymphocyte.Count_slope	1			1	1	
Absolute.Monocyte.Count_slope	1			1	1	

Feature Name	Between Cluster Significant Differences					
	1-2	1-3	1-4	2-3	2-4	3-4
…	…					
onset_delta.x	1	1	1	1	1	1
…	…					
Q9_Climbing_Stairs_slope	1			1		
…	…					
leg_max		1	1	1	1	
…	…					



Tang, et al. (2018), Neuroinformatics

Case-Studies – ALS – Dimensionality Reduction



2D t-SNE Manifold embedding

Learn a mapping: $f: R^n \xrightarrow{n \gg d} R^d$
$\{x_1, x_2, \ldots, x_n\} \longrightarrow \{y_1, y_2, \ldots, y_d\}$
preserves closely the *original distances*, $p_{i,j}$ and represents the *derived similarities*, $q_{i,j}$ between pairs of embedded points:

$$q_{i,j} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq i} \left(1 + ||y_i - y_k||^2\right)^{-1}}$$

$$\min_f KL(P||Q) = \sum_{i \neq j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$$

$0 = \frac{\partial KL(P||Q)}{\partial y_i} = 2 \sum_j (p_{i,j} - q_{i,j}) f(|x_i - x_j|) u_{i,j}$

$f(z) = \frac{z}{1+z^2}$ and $u_{i,j}$ is a unit vector from y_j to y_i.

Tang, et al. (2018), Neuroinformatics

Case-Studies – Parkinson's Disease

- **Investigate falls in PD patients** using clinical, demographic and neuroimaging data from two independent initiatives (UMich & Tel Aviv U)
- Applied **controlled feature selection** to identify the most salient predictors of patient falls (gait speed, Hoehn and Yahr stage, postural instability and gait difficulty-related measurements)
- **Model-based** (e.g., GLM) and **model-free** (RF, SVM, Xgboost) analytical methods used to forecasts clinical outcomes (e.g., falls)
- Internal statistical cross **validation** + external out-of-bag validation
- Four specific **challenges**
  - Challenge 1, harmonize & aggregate complex, multisource, multisite PD data
  - Challenge 2, identify salient predictive features associated with specific clinical traits, e.g., patient falls
  - Challenge 3, forecast patient falls and evaluate the classification performance
  - Challenge 4, predict tremor dominance (TD) vs. posture instability and gait difficulty (PIGD).
- **Results**: model-free machine learning based techniques provide a more reliable clinical outcome forecasting, e.g., falls in Parkinson's patients, with classification accuracy of about 70-80%.

Gao, et al. SREP (2018)

Case-Studies – Parkinson's Disease

Falls in PD are extremely difficult to predict …

PD phenotypes
Tremor-Dominant (TD)
Postural Instability & gait difficulty (PI & GD)

Case-Studies – Parkinson's Disease

Method	acc	sens	spec	ppv	npv	lor	auc
Logistic Regression	0.728	0.537	0.855	0.710	0.736	1.920	0.774
Random Forests	**0.796**	**0.683**	**0.871**	**0.778**	**0.806**	**2.677**	**0.821**
AdaBoost	0.689	0.610	0.742	0.610	0.742	1.502	0.793
XGBoost	0.699	0.707	0.694	0.604	0.782	1.699	0.787
SVM	0.709	0.561	0.806	0.657	0.735	1.672	0.822
Neural Network	0.699	0.610	0.758	0.625	0.746	1.588	
Super Learner	0.738	0.683	0.774	0.667	0.787	1.999	

Results of binary fall/no-fall classification (5-fold CV) using top 10 selected features (gaitSpeed_Off, ABC, BMI, PIGD_score, X2.11, partII_sum, Attention, DGI, FOG_Q, H_and_Y_OFF)

Gao, et al. SREP (2018)

Open-Science & Collaborative Validation

End-to-end Big Data analytic protocol jointly processing complex imaging, genetics, clinical, demo data for assessing PD risk

- o Methods for rebalancing of imbalanced cohorts
- o ML classification methods generating consistent and powerful phenotypic predictions
- o Reproducible protocols for extraction of derived neuroimaging and genomics biomarkers for diagnostic forecasting

Collabs

https://github.com/SOCR/PBDA

Case-Studies – General Populations

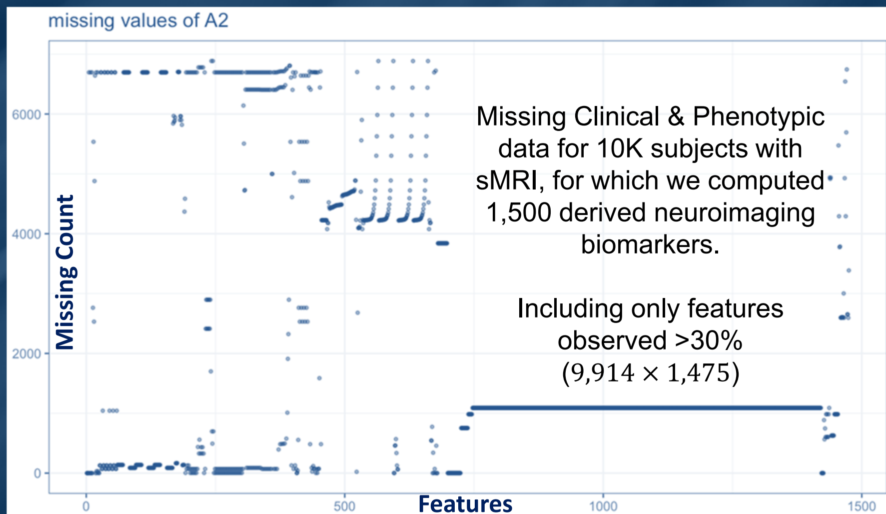
2	20005	Ongoing characteristics	Email access
2	110007	Ongoing characteristics	Newsletter communications, date sent
100	25780	Brain MRI	Acquisition protocol phase.
100	12139	Brain MRI	Believed safe to perform brain MRI scan
100	12188	Brain MRI	Brain MRI measurement completed
100	12187	Brain MRI	Brain MRI measuring method
100	12663	Brain MRI	Reason believed unsafe to perform brain MRI
100	12704	Brain MRI	Reason brain MRI not completed
100	12652	Brain MRI	Reason brain MRI not performed
101	12292	Carotid ultrasound	Carotid ultrasound measurement completed
101	12291	Carotid ultrasound	Carotid ultrasound measuring method
101	20235	Carotid ultrasound	Carotid ultrasound results package
101	22672	Carotid ultrasound	Maximum carotid IMT (intima-medial thickness) at 120 degrees
101	22675	Carotid ultrasound	Maximum carotid IMT (intima-medial thickness) at 150 degrees
101	22678	Carotid ultrasound	Maximum carotid IMT (intima- degrees
101	22681	Carotid ultrasound	Maximum carotid IMT (intima- degrees
101	22671	Carotid ultrasound	Mean carotid IMT (intima-med
101	22674	Carotid ultrasound	Mean carotid IMT (intima-med
101	22677	Carotid ultrasound	Mean carotid IMT (intima-med
101	22680	Carotid ultrasound	Mean carotid IMT (intima-med
101	22670	Carotid ultrasound	Minimum carotid IMT (intima- degrees
101	22673	Carotid ultrasound	Minimum carotid IMT (intima- degrees
101	22676	Carotid ultrasound	Minimum carotid IMT (intima-medial thickness) at 210 degrees
101	22679	Carotid ultrasound	Minimum carotid IMT (intima-medial thickness) at 240 degrees
101	22682	Carotid ultrasound	Quality control indicator for IMT at 120 degrees
101	22683	Carotid ultrasound	Quality control indicator for IMT at 150 degrees

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

Data Source	Sample Size/Data Type	Summary
UK Biobank	**Demographics**: > 500K cases **Clinical data:** > 4K features **Imaging data**: T1, resting-state fMRI, task fMRI, T2_FLAIR, dMRI, SWI **Genetics data**	The longitudinal archive of the UK population (NHS)

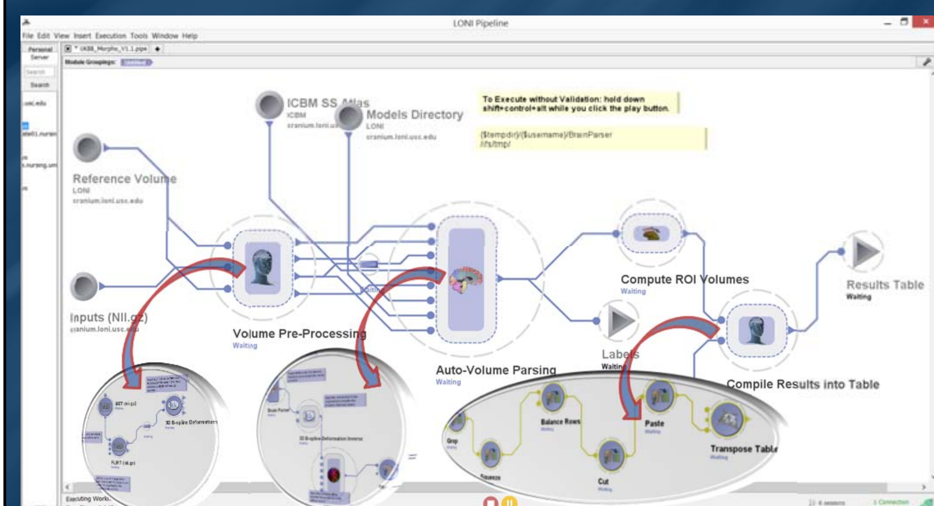
http://www.ukbiobank.ac.uk
http://bd2k.org

Case-Studies – UK Biobank (Complexities)

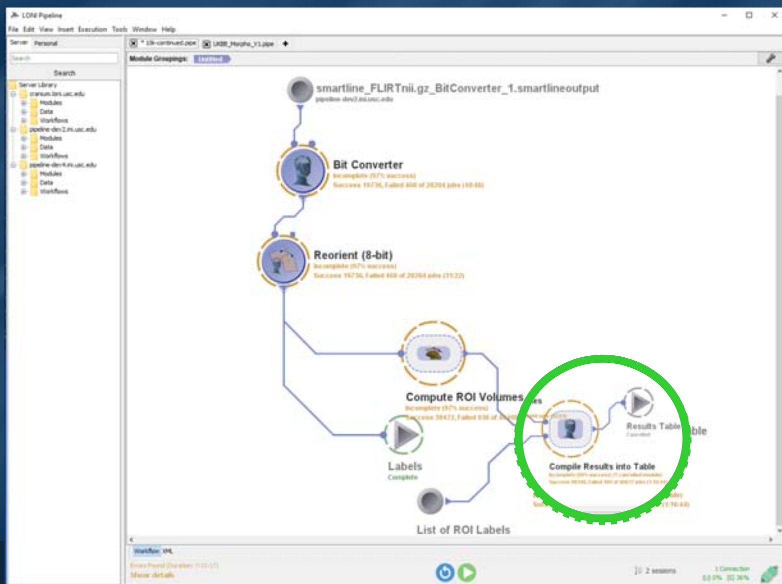


Zhou, et al. (2019), SREP | **https://github.com/SOCR/UKBB_Analytics**

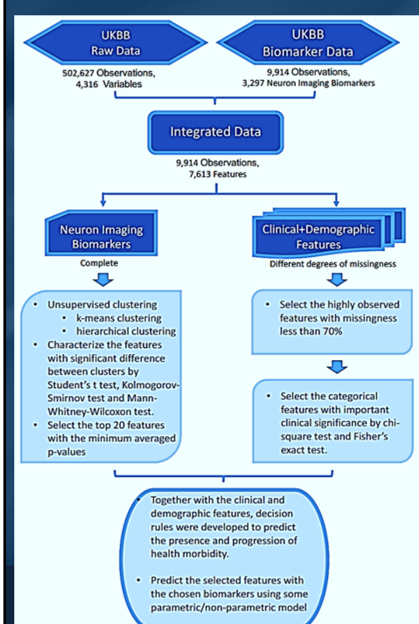
Case-Studies – UK Biobank – NI Biomarkers

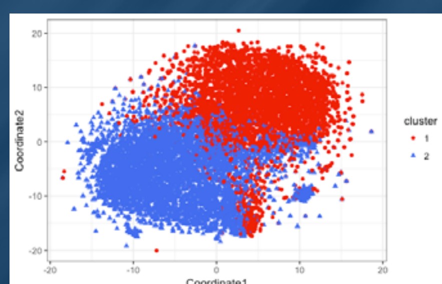


Case-Studies – UK Biobank – Successes/Failures



Case-Studies – UK Biobank – Results





t-SNE plot of the brain neuroimaging biomarkers

k-means clustering			
		Cluster 1	Cluster 2
Hierarchical clustering	Cluster 1	3768 (38.0%)	528 (5.3%)
	Cluster 2	827 (8.3%)	4791 (48.3%)

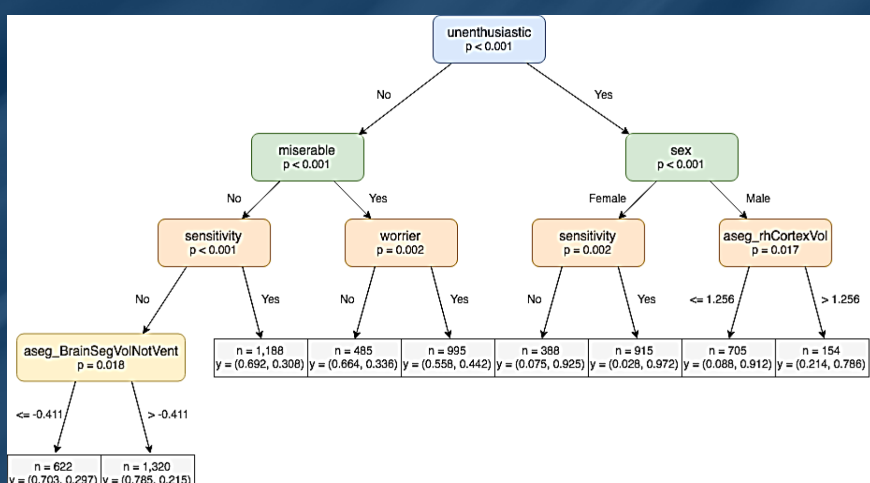
Cluster	Consistency	Variance	Cluster-size	Silhouette
1	0.997	0.001	5344	0.09
2	0.934	0.001	4570	0.05

Case-Studies – UK Biobank – Results

Variable	Cluster 1	Cluster 2
Sex		
Female	1,134 (24.7%)	4,062 (76.[illegible])
Male	3,461 (75.3%)	1,257 (23.[illegible])
Sensitivity/hurt feelings		
Yes	2,142 (47.9%)	3,023 (58.[illegible])
No	2,332 (52.1%)	2,151 (41.[illegible])
Worrier/anxious feelings		
Yes	2,173 (48.2%)	2,995 (57.[illegible])
No	2,337 (51.8%)	2,208 (42.[illegible])
Risk taking		
Yes	1,378 (31.0%)	1,154 (22.[illegible])
No	3,064 (69.0%)	3,933 (77.[illegible])
Guilty feelings		
Yes	1,100 (24.4%)	1,697 (32.[illegible])
No	3,417 (75.6%)	3,536 (67.[illegible])
Seen doctor for nerves, anxiety, tension or depression		
Yes	1,341 (29.3%)	1,985 (37.[illegible])
No	3,237 (70.7%)	3,310 (62.[illegible])
Alcohol usually taken with meals		
Yes	1,854 (66.7%)	2,519 (76.[illegible])
No	924 (33.3%)	771 (23.4[illegible])
Snoring		
Yes	1,796 (41.1%)	1,652 (33.[illegible])
No	2,577 (58.9%)	3,306 (66.[illegible])
Worry too long after embarrassment		
Yes	1,978 (44.3%)	2,675 (52.[illegible])
No	2,491 (55.7%)	2,462 (47.[illegible])
Miserableness		
Yes	1,715 (37.7%)	2,365 (45.[illegible])
No	2,829 (62.3%)	2,882 (54.[illegible])
Ever highly irritable/argumentative for 2 days		
Yes	485 (10.7%)	749 (14.5%)
No	4,038 (89.3%)	4,418 (85.[illegible])
Nervous feelings		
Yes	751 (16.6%)	1,071 (20.[illegible])
No	3,763 (83.4%)	4,076 (79.[illegible])
Ever depressed for a whole week		
Yes	2,176 (48.1%)	2,739 (52.[illegible])
No	2,347 (51.9%)	2,438 (47.[illegible])
Ever unenthusiastic/disinterested for a whole week		
Yes	1,346 (30.3%)	1,743 (34.[illegible])
No	3,089 (69.7%)	3,344 (65.[illegible])
Sleepless/insomnia		
Never/rarely	1,367 (29.8%)	1,181 (22.[illegible])
Sometimes	2,202 (47.9%)	2,571 (48.[illegible])
Usually	1,024 (22.3%)	1,563 (29.[illegible])
Getting up in morning		
Not at all easy	139 (3.1%)	249 (4.7%[illegible]
Not very easy	538 (11.9%)	830 (15.8[illegible]
Fairly easy	2,327 (51.4%)	2,663 (50.[illegible])
Very easy	1,526 (33.7%)	1,505 (28.[illegible])
Nap during day		
Never/rarely	2,497 (54.5%)	3,238 (61.[illegible])
Sometimes	1,774 (38.8%)	1,798 (34.[illegible])
Usually	307 (6.7%)	228 (4.3%[illegible]
Frequency of tiredness/lethargy in last 2 weeks		
Not at all	2,402 (53.0%)	2,489 (47.[illegible])
Several days	1,770 (39.0%)	2,127 (40.[illegible])
More than half the days	187 (4.1%1)	300 (5.8%[illegible]
Nearly everyday	177 (3.9%)	287 (5.5%[illegible]
Alcohol drinker status		
Never	81 (1.8%)	179 (3.4%[illegible]
Previous	83 (1.8%)	146 (2.7%[illegible]
Current	4,429 (96.4%)	[illegible].9%)

Variable	Cluster 1	Cluster 2
Sex		
Female	1,134 (24.7%)	4,062 (76.4%)
Male	3,461 (75.3%)	1,257 (23.6%)
• • •	• • •	
Nervous feelings		
Yes	751 (16.6%)	1,071 (20.8%)
No	3,763 (83.4%)	4,076 (79.2%)
• • •	• • •	
Frequency of tiredness/lethargy in last 2 weeks		
Not at all	2,402 (53.0%)	2,489 (47.8%)
Several days	1,770 (39.0%)	2,127 (40.9%)
More than half the days	187 (4.1%1)	300 (5.8%)
Nearly everyday	177 (3.9%)	287 (5.5%)
Alcohol drinker status		
Never	81 (1.8%)	179 (3.4%)
Previous	83 (1.8%)	146 (2.7%)
Current	4,429 (96.4%)	4,992 (93.9%)

Case-Studies – UK Biobank – Results



Decision tree illustrating a simple clinical decision support system providing machine guidance for identifying **depression feelings** based on categorical variables and neuroimaging biomarkers. In each terminal node, the y vector includes the percentage of subjects being labeled as "no" and "yes", in this case, answering the question "Ever depressed for a whole week." The p-values listed at branching nodes indicate the significance of the corresponding splitting criterion.

Case-Studies – UK Biobank – Results

	Accuracy	95% CI (Accuracy)	Sensitivity	Specificity
Sensitivity/hurt feelings	0.700	(0.676, 0.724)	0.657	0.740
Ever depressed for a whole week	0.782	(0.760, 0.803)	0.938	0.618
Worrier/anxious feelings	0.730	(0.706, 0.753)	0.721	0.739
Miserableness	0.739	(0.715, 0.762)	0.863	0.548

Cross-validated (random forest) prediction results for four types of mental disorders

Zhou, et al. (2019), SREP

What's Next?

- Lots of "open problems" in data-science, e.g., fundamentals of data representation & analytics
- The SOCR team is developing:
  - Compressive Big Data Analytics (CBDA) technique – an ensemble learning meta-algorithm
  - DS Time-Complexity and Inferential-Uncertainty
- Need lots of community, institutional, state, federal, and philanthropic support to advance data science methods, enhance the computing infrastructure, train/support students/fellows, and tackle the *Kryder Law* ≫ *Moore Law* trend

Acknowledgments

Slides Online:
"SOCR News"

Funding

NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760, UL1TR002240
NSF: 1734853, 1636840, 1416953, 0716055, 1023115
The Elsie Andresen Fiske Research Fund

http://SOCR.umich.edu

Collaborators

- **SOCR**: Milen Velev, Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Jingshu Xu, Nellie Ponarul, Ming Tang, Asiyah Lin, Nicolas Christou, Hanbo Sun, Tuo Wang. Simeone Marino, Nina Zhou, Yi Zhao, Lu Wang, Qiucheng Wu
- **LONI/INI**: 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, Junning 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