DataSifter: Sharing of Sensitive Data via Statistical Obfuscation

Ivo D. Dinov

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

University of Michigan

[redacted]

http://SOCR.umich.edu

Slides Online:
"SOCR News"

SCHOOL OF NURSING
STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

My Perspective on Stats & Math @ FSU

- 1947 – FSU
  - FL State College(1905)←FL University(1883)←West FL Seminary (1857)
- 1906 – Math (Chair: Elmer Smith; til' 1942, cf. Smith Hall)
- 1950's → Topology (Morton Curtis, Thomas Wade, Orville Harrold)
- 1959 – Stats spins off of Math
- 1967 – De Witt Sumners (RO Lawton Prof.) joins
- 1982 – Fred Huffer joins Stats (from Stanford)
- 1993-1998 – ID dually enrolled in Math/Stats
- …
- 2019 – FSU-Stats 60th anniversary

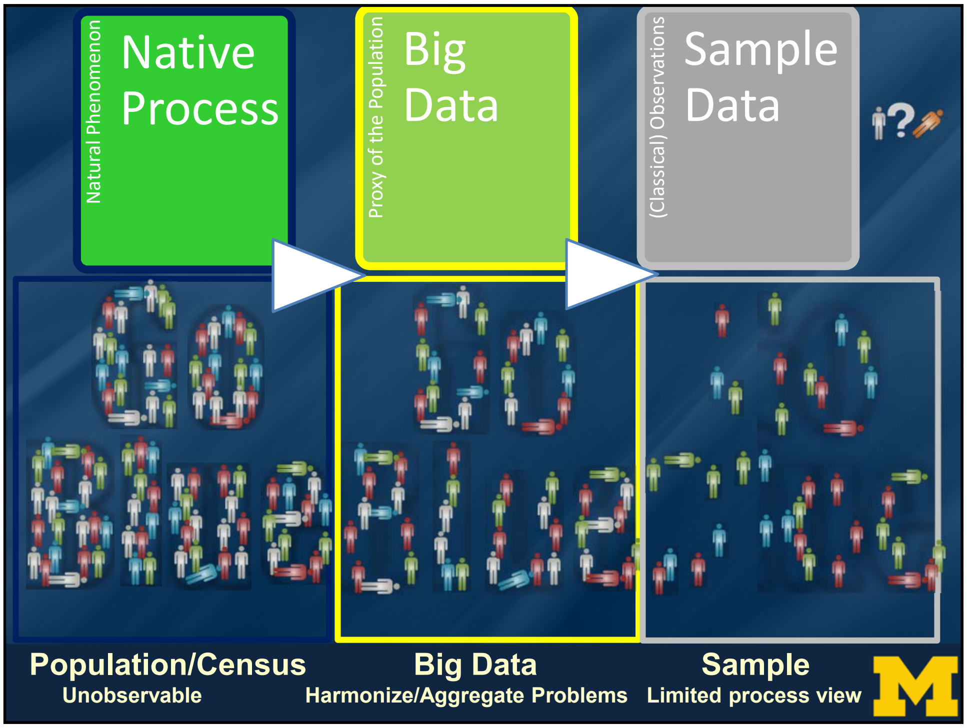
https://www.math.fsu.edu/News/Math_Newsletter_Fal2001.pdf
https://www.math.fsu.edu/News/Math_Newsletter_Spr2006.pdf

FSU Stats Alumni (Academic) Pedigree

http://www.socr.umich.edu/people/dinov/bio.html
https://www.genealogy.math.ndsu.nodak.edu

Outline

- ❑ Driving biomedical & health challenges
- ❑ Common characteristics of Big Neuroscience Data
- ❑ ε-Differential Privacy & Homomorphic Encryption
- ❑ *DataSifter: Statistical obfuscation*
- ❑ Case-studies
  - ❑ Applications to Neurodegenerative Disease (Udall/MADC)
  - ❑ Autism Brain Imaging Data Exchange (ABIDE)
  - ❑ Population Census-like Neuroscience



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

ε-Differential Privacy (εDP) vs. fully Homomorphic Encryption (fHE)

Category	εDP	fHE
Goal	Mine information in a DB without compromising privacy; no access to inspect individual DB entries	Provide a secure encryption allowing program execution on encrypted data; encrypt results, interpretation requires ability to decrypt the data
Pros	Theoretical limits on the balance between utility and risk of sharing data	Elegant and powerful math framework for bijective (encode/decode) encryption. Fast
Cons	Difficult for unstructured, skewed, and categorical data	There are limitations on deriving $\boldsymbol{f}'$ – commutative analytic evaluators

ε-Differential privacy (εDP)

- **Data-features**: $\{C_1, C_2, \ldots, C_k\}$, categorical or numerical.
- **DB** = list of cases $\{x_1, x_2, \ldots, x_n\}$, $x_i \in C_1 \times C_2 \times \cdots, \times C_k, \quad 1 \leq i \leq n.$

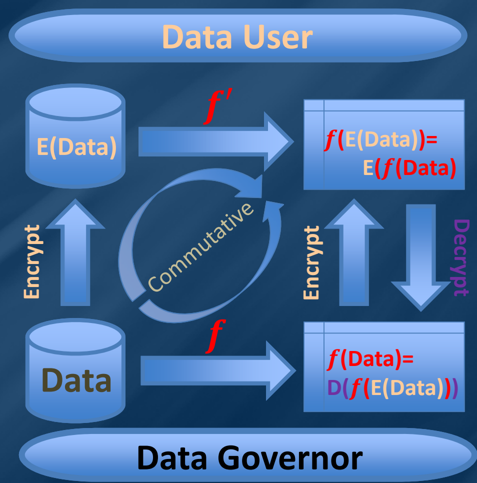
- ε-Differential privacy relies on adding noise to data to protect the identities of individual records. An **algorithm f** is ε-differentially private if for all possible inputs (datasets/DBs) D_1, D_2 that differ on a single record, and all possible f outputs, y, the probability of correctly guessing D_1 knowing y is not significantly different from that of D_2:

$$\frac{P(f(D_1) = y)}{P(f(D_2) = y)} \leq e^{\varepsilon}, \qquad \forall y \in Range(f).$$

- The global sensitivity of f is the smallest number $S(f)$, such that $\forall D_1, D_2$ that differ on at most one element $\|f(D_1) - f(D_2)\|_1 \leq S(f)$
- There are many differentially private algorithms, e.g., random forests, decision trees, k-means clustering, etc.
- E.g., $f: D = DB \to R^m$, the algorithm outputting $f(D) + (y_1, y_2, \ldots, y_m)$, with $y_i \in Laplace\left(\mu = 0, \sigma = \sqrt{2}\frac{S(f)}{\varepsilon}\right), \forall i$ is ε-differentially private

Dwork, LNCS, 2008

Homomorphic Encryption (HE)



Legend

Data – dataset/DB
f – data analytic process
f' – commutative analytic evaluator
E – encryption protocol
D – decryption protocol

Rivest & Adleman, Academic Press, 1978

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	$0 \le \eta = \eta(k_0 + k_1 + k_2 + k_3 + k_4) \le 1$				
	k_0	k_1	k_2	k_3	$\boldsymbol{k_4}$
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				

$\boldsymbol{k_0}$: A Boolean; obfuscate the unstructured features?

$\boldsymbol{k_1}$: proportion of artificial missing data values that should be introduced

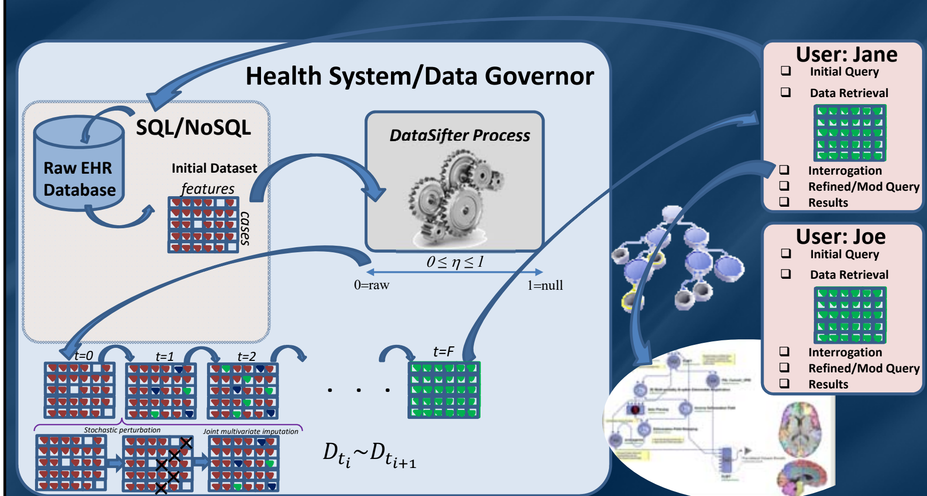
$\boldsymbol{k_2}$: The number of times to iterate

$\boldsymbol{k_3}$: The fraction of structured features to be obfuscated in all the cases

$\boldsymbol{k_4}$: 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



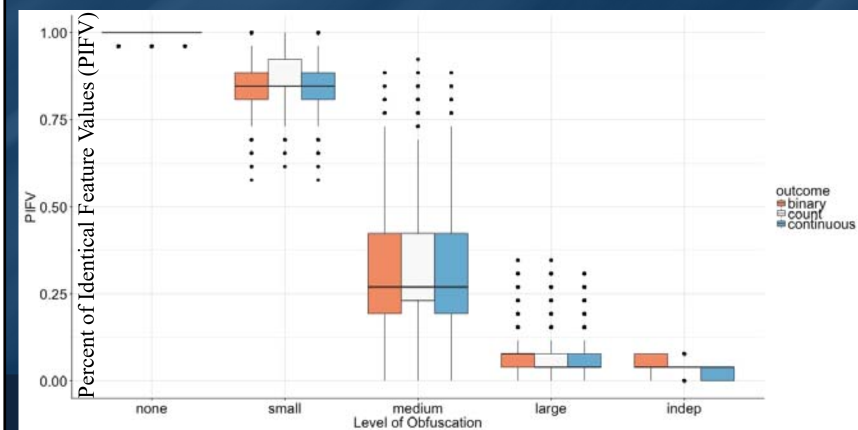
http://DataSifter.org US patent #16/051,881 Marino, *et al.*, *JSCS* (2019)

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.

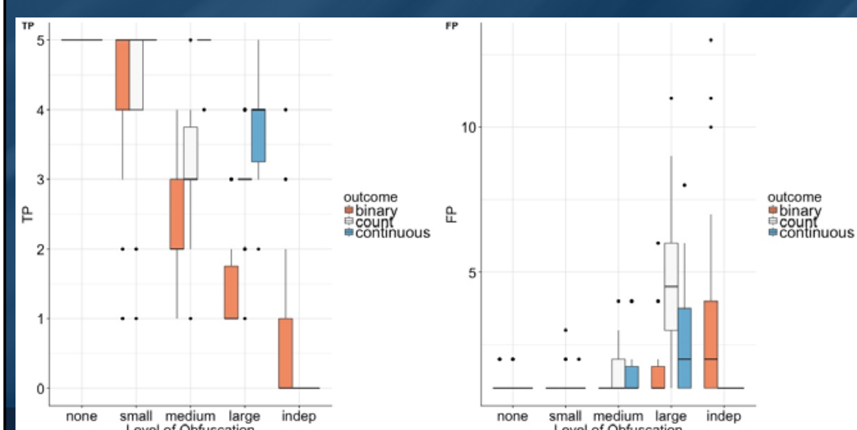
Each box represents 30 different "sifted" data or 30,000 "sifted" cases.



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 a random (22nd) ABIDE Subject

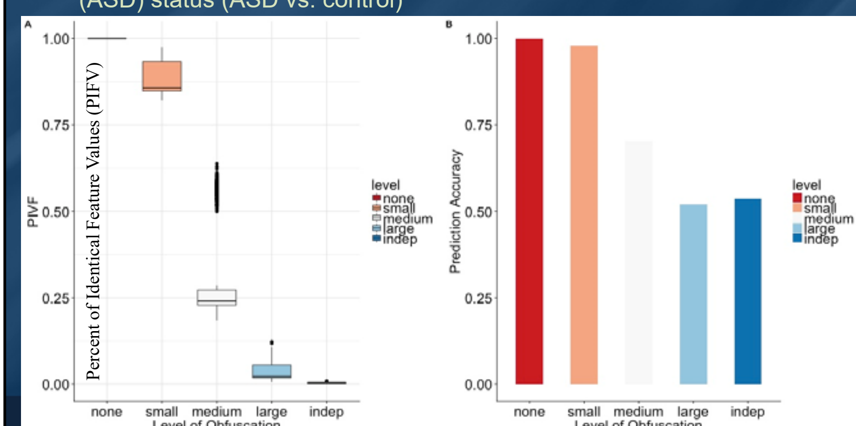
η	Output	Sex	Age	Acquisition Plane	IQ	thick_std_ctx .lh.cuneus	curv_ind_ctx _lh_G_front_ inf.Triangul	gaus_curv_ ctx.lh. medialorbitofront al	curv_ind_ctx _lh_S_interm _prim.Jensen
original	Autism	M	31.7	Sagittal	131	0.475	2.1	0.315	NA
none	Autism	M	31.7	Sagittal	131	0.475	2.1	0.315	0.51
small	Autism	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

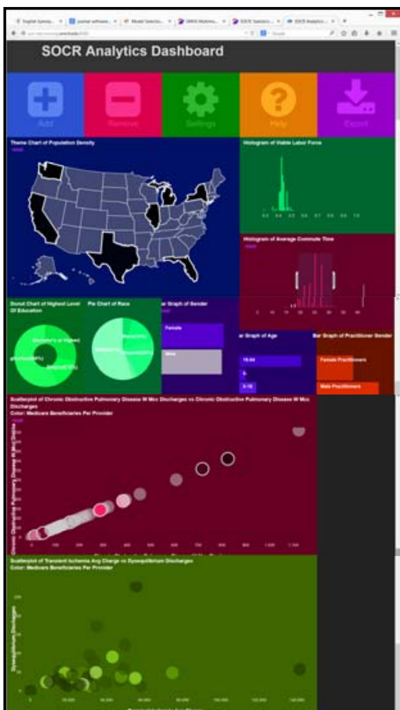
IV. Clinical Data Application: Using DataSifter to Obfuscate ABIDE

PIFVs for ABIDE under different levels of DataSifter obfuscations.
Each box represents 1,098 subjects among the ABIDE sub-cohort
Random forest prediction of binary clinical outcome - autism spectrum disorder (ASD) status (ASD vs. control)



$$\underbrace{\text{Raw Data}}_{Sensitive\ Info} \rightarrow \underbrace{\text{Statistical Obfuscation}}_{DataSifter} \rightarrow \underbrace{\text{Data Analytics}}_{\substack{Model-based \\ Model-Free}}$$

APPLICATIONS



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

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

Case-Studies – ALS

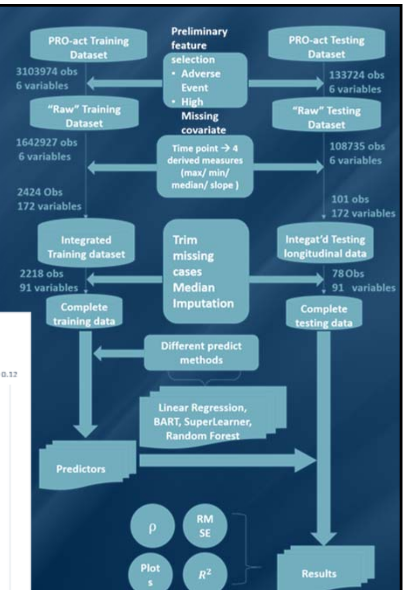
- ❑ Identify predictive classifiers to detect, track and prognosticate the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function)
- ❑ Provide a decision tree prediction of adverse events based on subject phenotype and 0-3 month clinical assessment changes

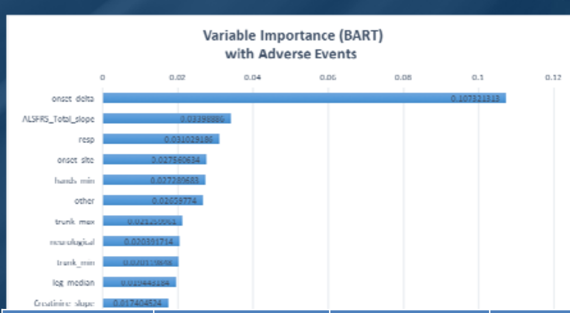
Data Source	Sample Size/Data Type	Summary
ProAct Archive	Over 100 variables are recorded for all subjects including: Demographics: age, race, medical history, sex; Clinical data: **Amyotrophic Lateral Sclerosis** Functional Rating Scale (ALSFRS), adverse events, onset_delta, onset_site, drugs use (riluzole)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

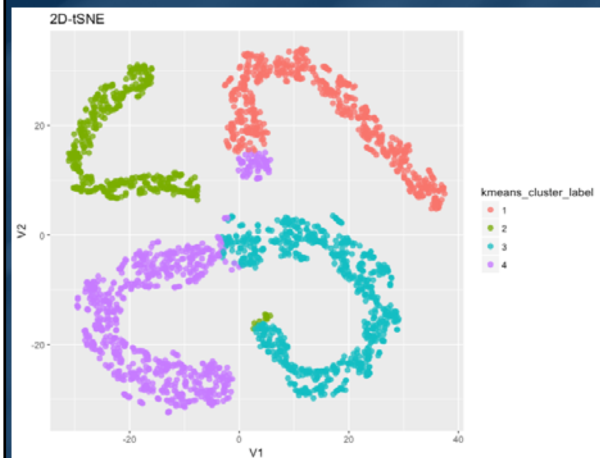
Case-Studies – ALS – Explicating Clustering

Feature Name	Between Cluster Significant Differences					
	1-2	1-3	1-4	2-3	2-4	3-4
onset_site	[illegible]	[illegible]	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	[illegible]	[illegible]	[illegible]	[illegible]	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[illegible]				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	[illegible]	[illegible]	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

Acknowledgments

Slides Online:
"SOCR News"

US patent #16/051,881

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