Data De-Identification & Clinical Decision Support

Ivo D. Dinov

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

University of Michigan

https://SOCR.umich.edu

Slides Online: "SOCR News"

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

Pillars of Open-Science

- ❑ **Data Resources & Analytical Tools**
- ❑ **Data De-Identification**
- ❑ **Clinical Decision Support Systems**

Characteristics of Big Biomed Data

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

Big Bio Data Dimensions	Challenges
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

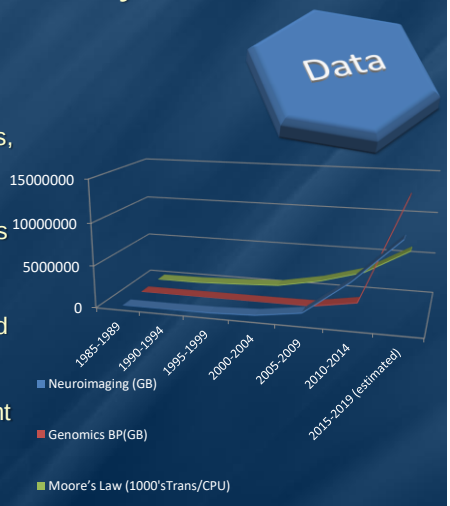
Example: analyzing observational data of 1,000's Parkinson's disease patients based on 10,000's signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements

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

Rationale for Open Science: Kryder vs. Moore

- 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

Data Sources

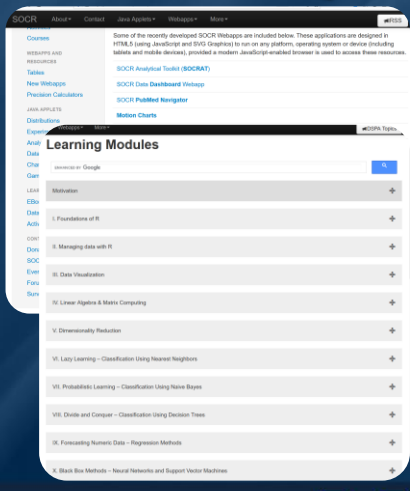
- UKBB https://www.ukbiobank.ac.uk
- MIMIC-III https://mimic.physionet.org
- SOCR Data Archives
  https://wiki.socr.umich.edu/index.php/SOCR_Data
- NIH Databases https://eresources.nlm.nih.gov/nlm_eresources

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** **SARS-CoV-2 virus tests**	The longitudinal archive of the UK population (NHS)
MIMIC-III	ADMISSIONS, DIAGNOSES, ICUSTAYS,MICROBIOLOGY, PRESCRIPTIONS, PROCEDURES_ICD,SERVICES	ICU Data for over 40K patients
NIH Databases	100's of open-access DBs	

FAIR = Findable + Accessible + Interoperable + Reusable

Analytical Tools

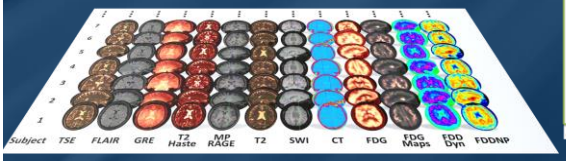
- SOCR Webapps
  (HTML-based) https://socr.umich.edu/HTML5
- Data Science and Predictive Analytics (DSPA)
  (R-based) https://dspa.predictive.space
- Spacekime analytics: https://spacekime.org

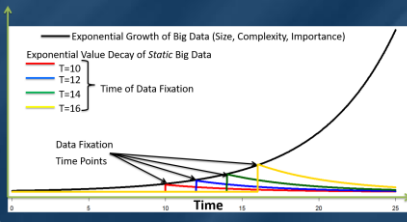


- Data Resources & Analytical Tools
- **Data De-Identification**
- Clinical Decision Support Systems

Data Size, Privacy, Usage & Impact

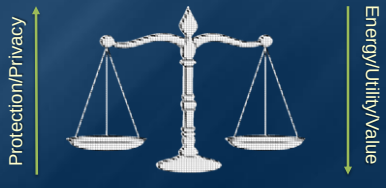
- Volume vs. Value of Data

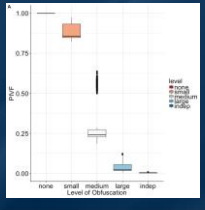




Dinov (2016) JMSI

- Security vs. Utility





Zhou *et al.* (2020), pending

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 derived info
Pros	Theoretical limits on the balance between utility and risk of sharing data	Fast, elegant, and powerful math framework for bijective (encode/decode) encryption
Cons	Difficult for unstructured, skewed, and categorical data	There are limitations on deriving f' – commutative analytic evaluators

DataSifter

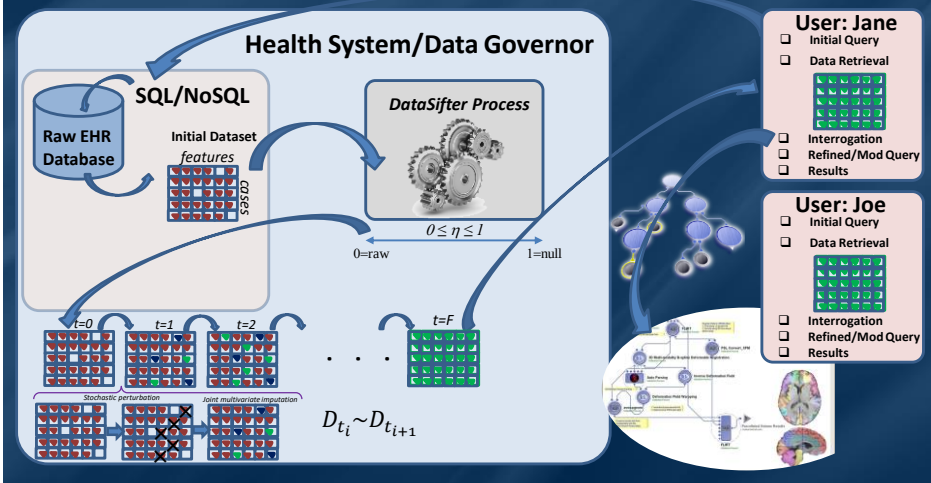
- DataSifter is an iterative statistical computing approach that provides the data-governors controlled manipulation of the trade-off between sensitive information obfuscation and preservation of the joint distribution.
- The DataSifter is designed to satisfy data requests from pilot study investigators 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 #10,776,516 Marino, Zhou, *et al.*, *JSCS* (2019)

Reliable, Effective & Secure Data Sharing

- Why is data-sharing difficult?
  monopoly, preservation of *status-quo,* competitive edge, personally identifiable information, IP protection, security (on multiple levels), red tape, …
- FAIR (Findable, Accessible, Interoperable & Reusable) Data are powerful
- Current Data Sharing Landscape?
  Differential Privacy, fully-homomorphic encryption, statistical obfuscation (DataSifter), …

DataSifter: Reliable, Effective & Secure Data Sharing

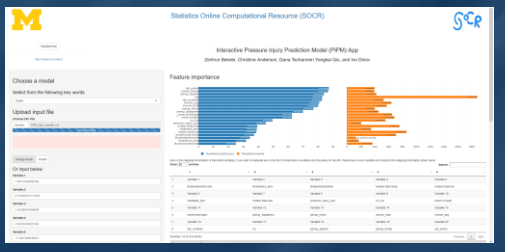


http://*DataSifter*.org US patent #10,776,516 Marino, *et al.*, *JSCS* (2019)

- ❑ Data Resources & Analytical Tools
- ❑ Data De-Identification
- ❑ **Clinical Decision Support Systems**

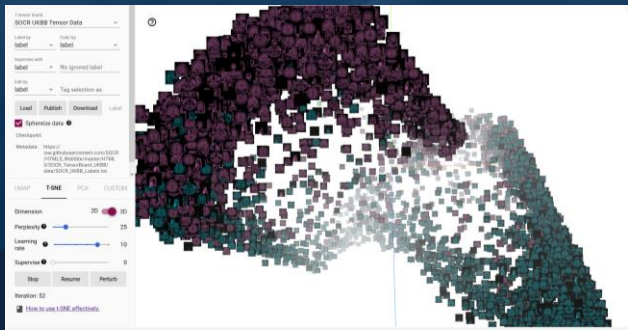
Clinical Decision Support

- Pressure Injury Prediction Model (PIPM) https://myumi.ch/O49zG



- SOCR TensorBoard (10K*200 tensor)

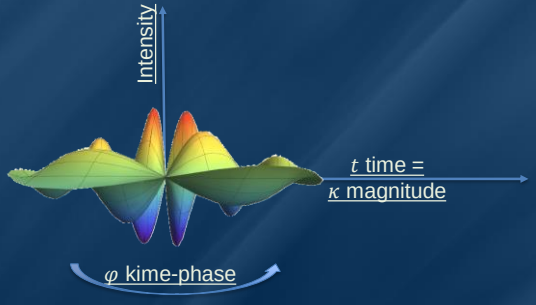
https://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB



Spacekime Analytics: Longitudinal Time-series ➔ Kime Surfaces

In the 5D spacekime manifold, time-series curves extend to kime-series, i.e., surfaces parameterized by kime-magnitude (t) and the kime-phase (φ).

Kime-phase aggregating operators that can be used to transform standard time-series curves to spacekime kime-surfaces, which can be modeled, interpreted, and predicted using advanced spacekime analytics.



Dinov & Velev (2021) De Gruyter | https://spacekime.org

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 open data science methods, enhance the computing infrastructure, train/support students/fellows, and tackle the $Kryder\ Law \gg Moore\ Law$ trend
- **Web**: https://SOCR.umich.edu
- **Git**: https://github.com/SOCR
- **Projects:** https://socr.umich.edu/html/SOCR_Research.html
- **Apps:** https://socr.umich.edu/HTML5/

Acknowledgments

Slides Online: "SOCR News"

Funding

NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, UL1TR002240, R01CA233487

NSF: 1916425, 1734853, 1636840, 1416953, 0716055, 1023115

Collaborators

- **SOCR**: **Milen Velev, Yueyang Shen, Daxuan Deng, Zijing Li, Yongkai Qiu, Zhe Yin, Yufei Yang, Yuxin Wang, Rongqian Zhang, Yuyao Liu, Yupeng Zhang, Yunjie Guo**
- **UMich MIDAS/MNORC/AD/PD Centers**: Chuck Burant, Kayvan Najarian, Stephen Goutman, Stephen Strobbe, Hiroko Dodge, Chris Monk, Issam El Naqa, HV Jagadish, Brian Athey

https://SOCR.umich.edu

Thank you for Participating

- You will receive a follow-up email including
  - A link to the slides and video recordings
  - A post-session evaluation survey
- Presenter contact information
  - Alexandre Dasilva (adasilva@umich.edu)
  - Ivo Dinov (dinov@umich.edu)