Integration of Data Science into the STEM Curricula

Ivo D. Dinov

Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science
Michigan Center for Applied & Interdisciplinary Mathematics
University of Michigan

https://SOCR.umich.edu

Joint work with Magdalena Ivanova (Michigan)

STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

Slides Online:
"SOCR News"

1

Outline

- ❑ Motivation & Rationale
- ❑ Data Science Foundations

- ❑ Physics ↔ STEM ↔ Data Science R&D ↔ Education & Training Curricula
- ❑ Learning Resources & Instructional Materials
- ❑ Some Relevant Hands-on Demos

2

Motivation & Rationale

3

From 23 … to … 2^{23}

- Data Science: 1798 vs. 2022
- In the 18^{th} 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 21^{st} century to achieve the same scientific impact, matching the reliability and the precision of the Cavendish's 18^{th} 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 ➔ $2^{23} \cong 10$M

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

5

Data Science Foundations

7

Pillars of Open-Science

8

Characteristics of Big 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

9

Physics ↔ STEM ↔ Data Science R&D ↔ Education & Training Curricula

- ❑ Transdisciplinary training integrating *theoretical models, experimental science, computational algorithms, data science applications & domain-specific practice*
- ❑ Curriculum Models (*quant STEM-based* vs. *qual EDA-based*)
  - ❑ Lightweight (MOOCs, <12 semester credits),
  - ❑ Intermediate (13-29 credits)
  - ❑ Heavyweight (30-56 credits, UG/Grad) curricula
- ❑ Physics, Data Science and *X* Training Programs
- ❑ Some (Michigan) data science and biophysics course examples

11

Spacekime Analytics: Example of Translating Mathematical-Physics ⟹ Data Science & AI

Physics	Data/Neuro Sciences
A **particle** is a small localized object that permits observations and characterization of its physical or chemical properties	An **object** is something that exists by itself, actually or potentially, concretely or abstractly, physically or incorporeal (e.g., person, subject, etc.)
An **observable** a dynamic variable about particles that can be measured	A **feature** is a dynamic variable or an attribute about an object that can be measured
Particle **state** is an observable particle characteristic (e.g., position, momentum)	**Datum** is an observed quantitative or qualitative value, an instantiation, of a feature
Particle **system** is a collection of independent particles and observable characteristics, in a closed system	**Problem**, aka Data System, is a collection of independent objects and features, without necessarily being associated with a priori hypotheses
Wave-function	**Inference-function**
Reference-Frame **transforms** (e.g., Lorentz)	Data **transformations** (e.g., wrangling, log-transform)
State of a system is an observed measurement of all particles ~ wavefunction	**Dataset (data)** is an observed instance of a set of datum elements about the problem system, $O = \{X, Y\}$
A **particle system is computable** if (1) the entire system is logical, consistent, complete and (2) the unknown internal states of the system don't influence the computation (wavefunction, intervals, probabilities, etc.)	**Computable data object** is a very special representation of a dataset which allows direct application of computational processing, modeling, analytics, or inference based on the observed dataset
…	…

Dinov & Velev (2021)

12

A Transdisciplinary Approach – Biomedical Informatics & Data Science Training Program (BIDS-TP)

- **Fellows & Trainees**
  - BIDS Grads
  - BIDS Fellows (Seniors Yr 2)
  - New BIDS Fellows (Juniors Yr1)
  - BIDS Trainees (Junior and Senior)

- **Faculty Mentors** (~40)
- **Curriculum**: 18 credits: 4 core & 2 elective courses + other activities (seminars, workshops)
- **Outcomes Tracking**: Time to Degree, Completion Rate, Graduate Career Pathways, Trainees Awards & Fellowships, Publications (GoogleScholar & ORCID profiles), Soft Metrics
- **BIDS-TP Program Leadership**: Maureen Sartor, Margit Burmeister, Brian Athey, Ivo Dinov

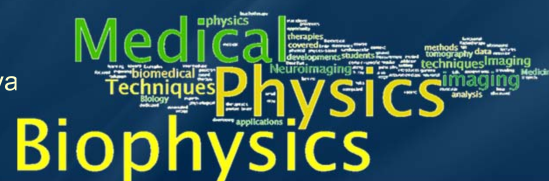
https://bids-tp.umich.edu

Modernizing the Methods and Analytics Curricula for Health Science Doctoral Programs
DOI: 10.3389/fpubh.2020.00022

13

Medical Physics

- **BIOPHYS 430 / PHYSICS 430** (Traditional UG/Grad course), 3-credits, students from physics, chemistry, STEM, biosciences
- Introduces the physics of physiological processes (muscular, cardiovascular, neuronal and renal), physics-based therapies, and biomedical imaging. Imaging techniques and physics-based therapies will be elucidated in the context of the underlying physical principles. Ultrasound, computed tomography, magnetic resonance imaging, and positron emission tomography. Radiotherapy methods will be also introduced. Course includes **data acquisition & image analysis using R**
- Instructor: Magdalena Ivanova



14

Biophysics of Disease

- **BIOPHYS 440 / Chem 440** (Traditional UG/Grad course), 3-credits, students from physics, chemistry, STEM, bio sciences
- Introduce the most commonly used biophysical methods for studying complex diseases and the application of these techniques for developing therapies. Emphasis is on protein aggregation diseases (Parkinson's, Alzheimer's and prion), but diseases like cancer, viral (HIV, influenza, and SARS-CoV-2) and bacterial infections will be also discussed. Classical biophysical methods like x-ray crystallography, NMR and cryoEM are covered, along with, recently emerging cutting-edge techniques. Some data science homework projects using real biomedical data.
- Instructor: Magdalena Ivanova

15

Data Science & Predictive Analytics

- **HS650** (Traditional grad-level course + online self-guided MOOC), 4-credits, students from 6 colleges representing STEM, bio, econ, humanities

- Builds computational abilities, inferential thinking, and practical skills for tackling core data scientific challenges. Covers foundational concepts in data management, processing, statistical computing, and dynamic visualization using modern programming tools and agile web-services.
- Blends core math principles ad concepts with computational techniques, tools and services for managing, harmonizing, aggregating, preprocessing, modeling, analyzing and interpreting large, multi-source, incomplete, incongruent, and heterogeneous data (Big Data). Biomedical, healthcare, and social datasets provide context for addressing specific driving challenges.

Dinov, Springer (2018)

16

Learning Resources & Instructional Materials

- EBooks
  - https://DSPA2.predictive.space
  - https://TCIU.predictive.space
  - https://BPAD.predictive.space
  - https://SpaceKime.org

- R Package
  - https://cran.rstudio.com/web/packages/TCIU

- GitHub
  - https://github.com/SOCR

17

Demonstrations

- Distribution (model-based) inference
  https://doi.org/10.1007/s42979-022-01206-w & https://doi.org/10.52041/iase.pdsxt
  https://socr.umich.edu/HTML5/BivariateNormal/BVN2 & http://distributome.org/V3
  https://socr.umich.edu/HTML5/SOCRAT
- Apps
  Fourier/Wavelet: https://socr.umich.edu/HTML5/Fourier_Wavelet_app
  Large Tensors/UMAP/t-SNE: https://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB
- DSPA (Rmarkdown eNotebook, R, Python, C, JS, …)
  https://socr.umich.edu/DSPA2/DSPA2_notes/05_SupervisedClassification.html#16_Case_Study:_Predicting_Galaxy_Spins
  https://socr.umich.edu/DSPA2/DSPA2_notes/10_SpecializedML_FormatsOptimization.html#17_R_Notebook_support_for_other_programming_languages
- Complex-time (Kime) & Spacekime Analytics
  https://www.socr.umich.edu/TCIU/HTMLs/Chapter4_TCIU_Predictive_Analytics.html
- SOCR & GitHub
  - https://socr.umich.edu & https://github.com/SOCR

18

Acknowledgments

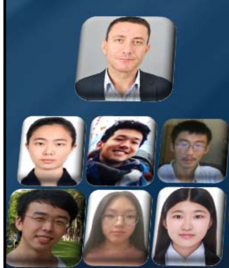
Slides Online: "SOCR News"

Funding

NIH: UL1TR002240, R01CA233487, R01MH121079, R01MH126137, T32GM141746
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, Simeone Marino**
- **UMich MIDAS/MCAIM Centers**: Lydia Bieri, Kayvan Najarian, Chris Monk, Issam El Naqa, HV Jagadish, Brian Athey, Magdalena Ivanova



https://SOCR.umich.edu

19

20