Data Sharing + Open, Rigorous & Reproducible Science

Ivo D. Dinov

Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics

University of Michigan

https://SOCR.umich.edu

Slides Online: "SOCR News"

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

1

Outline

- Pillars of Open-Science
- Rationale (Pros & Cons)
- Big Data Sharing
- *DataSifter: Statistical obfuscation*
- Case-studies
  - ALS Study
  - Population Census-like Neuroscience (UKBB)
  - Spacekime Analytics

2

Pillars of Open Data Science (HS650 / Bioinfo501)

3

Sources: Characteristics of Big Biomed Data

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

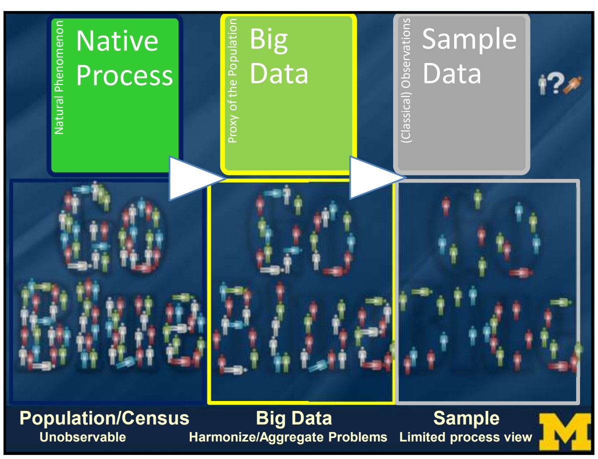
Big Bio Data Dimensions	Tools
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro to meso to micro scale 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 (2023) Springer

4



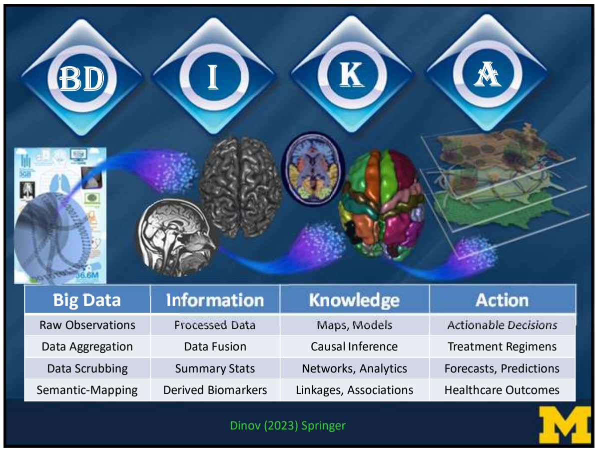
5

From 23 … to … 2^{23}

- Data Science: 1798 vs. 2024
- 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
- Data & Information Science ≅ Scalability & Compression (per Gerald Friedland/Berkeley): 23 → $2^{23} \cong$ 10M

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

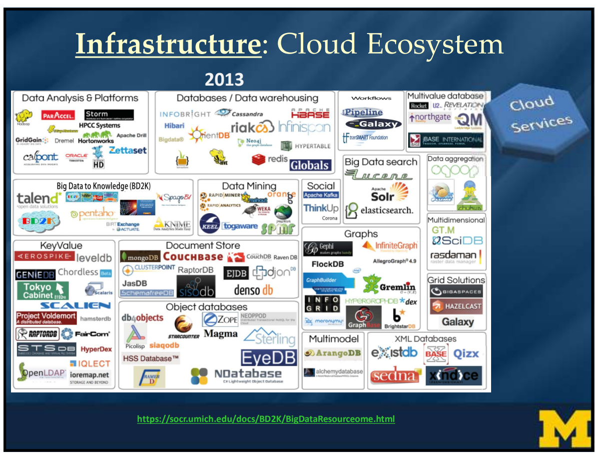
6


BD
I
K
A
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 (2023) Springer

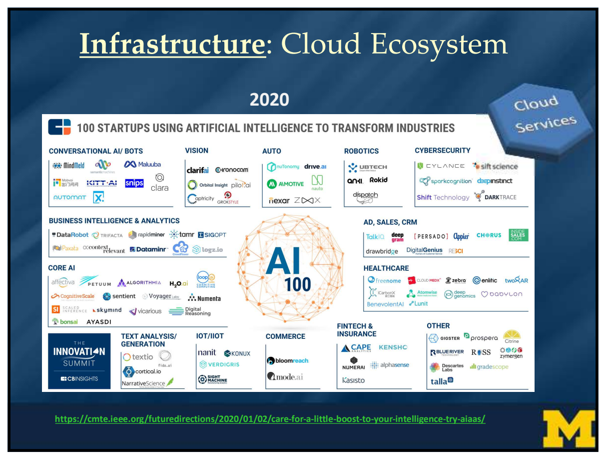
7


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
…
Data
FAIR = Findable + Accessible + Interoperable + Reusable

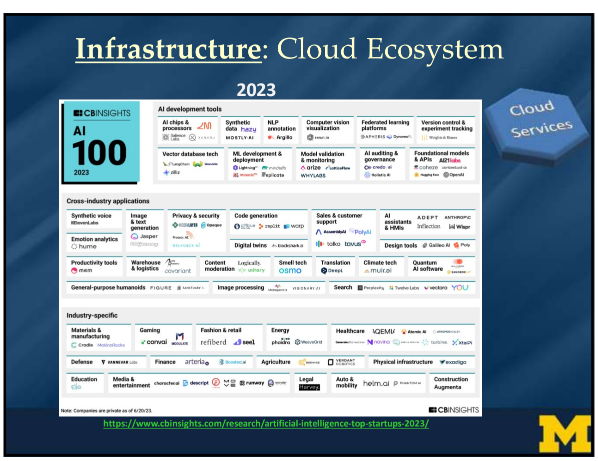
8


Infrastructure: Cloud Ecosystem
2013
Data Analysis & Platforms
Databases / Data warehousing
Big Data to Knowledge (BD2K)
Data Mining
Social
Big Data search
Data aggregation
KeyValue
Document Store
Graphs
Multidimensional
Grid Solutions
Object databases
Multimodel
XML Databases
Cloud Services
https://socr.umich.edu/docs/BD2K/BigDataResourceome.html

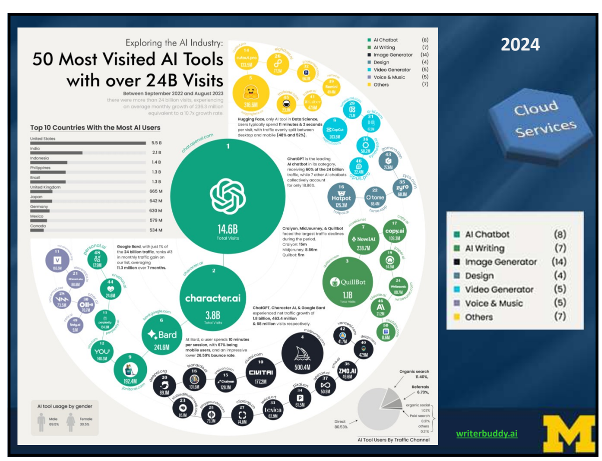
9


Infrastructure: Cloud Ecosystem
2020
100 STARTUPS USING ARTIFICIAL INTELLIGENCE TO TRANSFORM INDUSTRIES
CONVERSATIONAL AI/ BOTS
VISION
AUTO
ROBOTICS
CYBERSECURITY
BUSINESS INTELLIGENCE & ANALYTICS
AD, SALES, CRM
CORE AI
AI 100
HEALTHCARE
TEXT ANALYSIS/ GENERATION
IOT/IIOT
COMMERCE
FINTECH & INSURANCE
OTHER
Cloud Services
https://cmte.ieee.org/futuredirections/2020/01/02/care-for-a-little-boost-to-your-intelligence-try-aiaas/

10

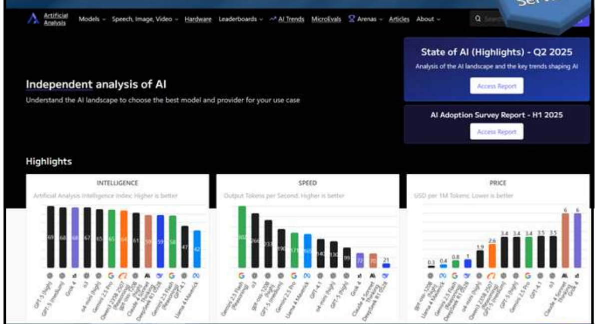

Infrastructure: Cloud Ecosystem
2023
CBINSIGHTS AI 100 2023
AI development tools
Cross-industry applications
Industry-specific
Cloud Services
https://www.cbinsights.com/research/artificial-intelligence-top-startups-2023/

11


Exploring the AI Industry:
50 Most Visited AI Tools with over 24B Visits
Top 10 Countries With the Most AI Users
14.6B
character.ai
3.8B
2024
Cloud Services
AI Chatbot (8)
AI Writing (7)
Image Generator (14)
Design (4)
Video Generator (5)
Voice & Music (5)
Others (7)
writerbuddy.ai

12

2025



Cloud Services

artificialanalysis.ai

13

Scholarly Research: OA Pubs/Sharing

Pubs

- 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., https://DSPA2.predictive.space
- Licensing
  - https://www.gnu.org/licenses
  - https://socr.umich.edu/html/SOCR_CitingLicense.html

15

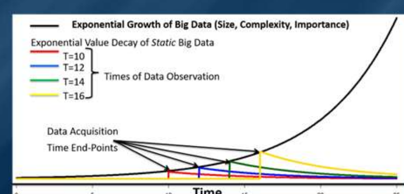
Rationale for Open Science (Cons)

- 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?

18

Rationale for Open Science (Pros)

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



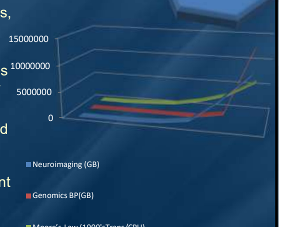
https://www.aaas.org/news/big-data-blog-part-v-interview-dr-ivo-dinov

19

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

20

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), …
- Digital Twins: https://EHR-Sim.StatisticalComputing.org/clinical-phenotype

21

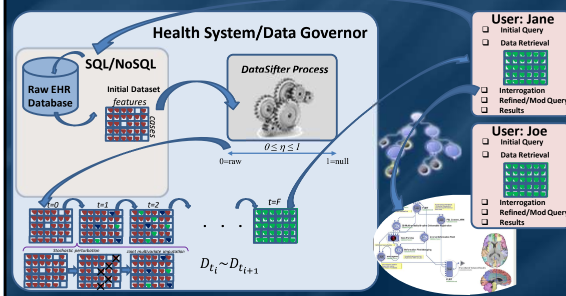
DataSifter

- 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.

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

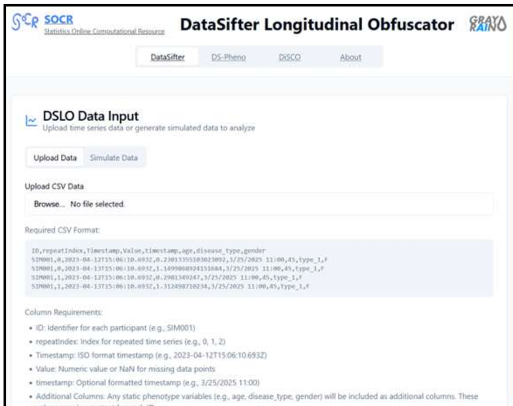
22

DataSifter



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

23



DataSifter

Try-It: https://SOCR-DSLO.StatisticalComputing.org

24

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

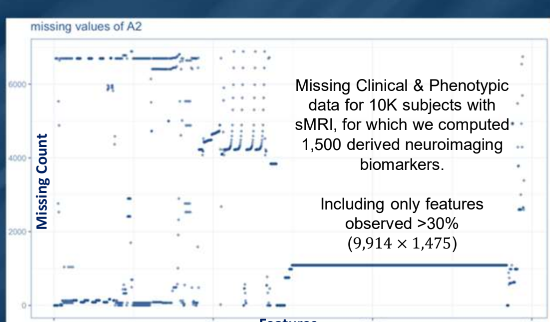
- UK Biobank – discriminate between HC, single and multiple comorbid conditions
- Predict likelihoods of various developmental or aging disorders
- Forecast cancer

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)

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

30

Case-Studies – UK Biobank (Complexities)



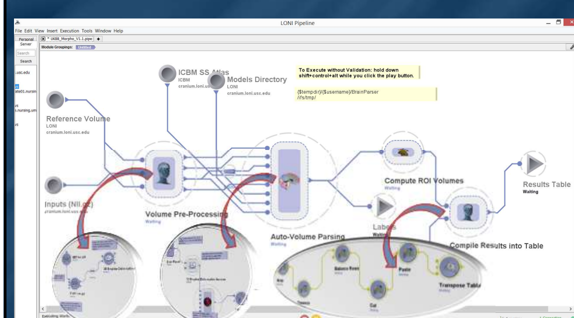
Missing Clinical & Phenotypic data for 10K subjects with sMRI, for which we computed 1,500 derived neuroimaging biomarkers.

Including only features observed >30% (9,914 × 1,475)

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

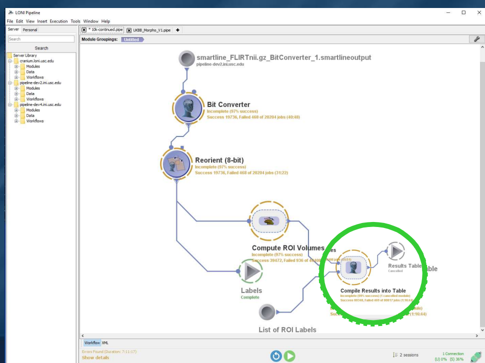
31

Case-Studies – UK Biobank – NI Biomarkers



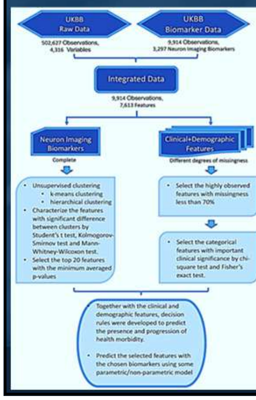
32

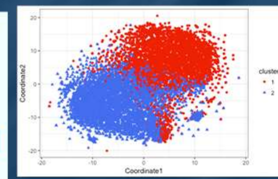
Case-Studies – UK Biobank – Successes/Failures



33

Case-Studies – UK Biobank – Results





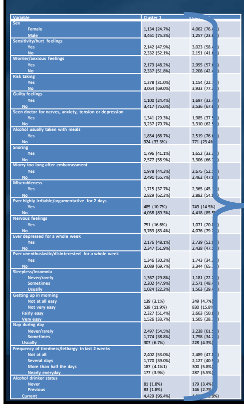
t-SNE plot of the brain neuroimaging biomarkers

		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

34

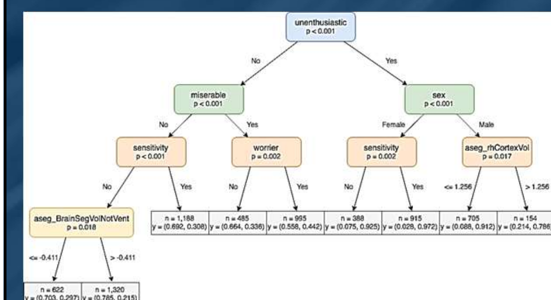
Case-Studies – UK Biobank – Results



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%)

35

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.

36

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

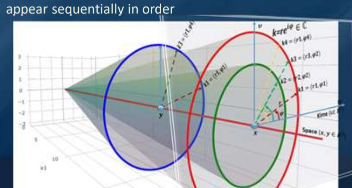
Cross-validated (random forest) prediction results for four types of mental disorders

Zhou, et al. (2019), SREP

37

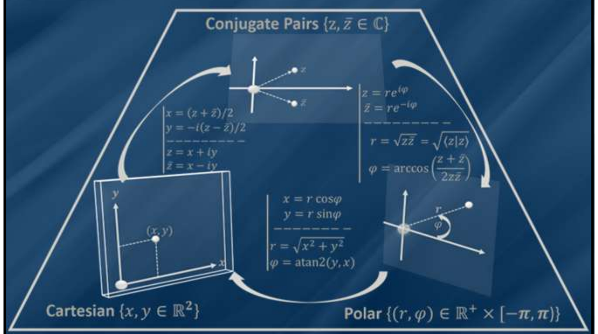
Complex-Time (*Kime*)

- At a given spatial location, $\boldsymbol{x}$, complex time (*kime*) is defined by $\kappa = re^{i\varphi} \in \mathbb{C}$, where:
  - the magnitude represents the longitudinal events order ($r > 0$) and characterizes the longitudinal displacement in time, and
  - event phase ($-\pi \le \varphi < \pi$) is an angular displacement, or event direction
- There are multiple alternative parametrizations of kime in the complex plane
- Space-kime manifold is $R^3 \times \mathbb{C}$:
  - $(\boldsymbol{x}, k1)$ and $(\boldsymbol{x}, k4)$ have the same spacetime representation, but different spacekime coordinates,
  - $(\boldsymbol{x}, k1)$ and $(\boldsymbol{y}, k1)$ share the same kime, but represent different spatial locations,
  - $(\boldsymbol{x}, k2)$ and $(\boldsymbol{x}, k3)$ have the same spatial-locations and kime-directions, but appear sequentially in order



38

Kime Parameterizations



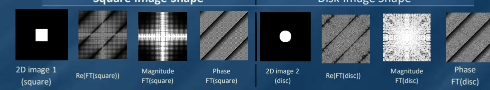
39

The Importance of Kime-Magnitude (*time*) and Kime-Phase (*direction*)

Fourier Analysis

(real part of the Forward Fourier Transform)

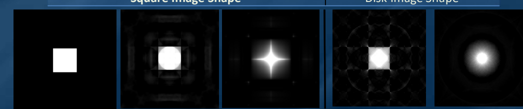
Square Image Shape | Disk Image Shape



Fourier Synthesis

(real part of the Inverse Fourier Transform)

Square Image Shape | Disk Image Shape

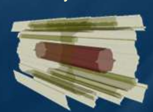


IFT(FT(square)) ≡ square	IFT using square-magnitude & disc-phase	IFT using square-magnitude & nil-phase	IFT using disc-magnitude & square-phase	IFT using disc-magnitude & nil-phase

40

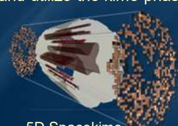
Longitudinal Data Analytics

- **Neuroimaging**:
  - *4D fMRI:* time-series, represents measurements of hydrogen atom densities over a 3D lattice of spatial locations ($1 \le x, y, z \le 64$ pixels), about 3×3 millimeters^2 resolution. Data is recorded longitudinally over time ($1 \le t \le 180$) in increments of about 3 seconds, then post-processed
  - *State-of-the-art Approaches*: Time-series modeling or Network analysis
  - *Spacekime Analytics*: 5D fMRI kime-series, represent the hydrogen atom densities over the same 3D lattice of spatial locations, longitudinally over the 2D kime space, $\kappa = re^{i\varphi} \in \mathbb{C}$
  - *Differences*: Spacekime analytics estimate and utilize the kime-phases



4D Spacetime

4D/5D Reconstructions



5D Spacekime

Dinov & Velev (2021)

41

Spacekime Calculus

- Kime **Wirtinger derivative** (first order kime-derivative at $\boldsymbol{k} = (r, \varphi)$):

In Cartesian coordinates:

$$f'(z) = \frac{\partial f(z)}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right) \quad \text{and } f'(\bar{z}) = \frac{\partial f(\bar{z})}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right).$$

In Conjugate-pair basis: $df = \partial f + \bar{\partial} f = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \bar{z}}d\bar{z}$.

In Polar kime coordinates:

$$f'(k) = \frac{\partial f(k)}{\partial k} = \frac{1}{2}\left(\cos\varphi\frac{\partial f}{\partial r} - \frac{1}{r}\sin\varphi\frac{\partial f}{\partial \varphi} - i\left(\sin\varphi\frac{\partial f}{\partial r} + \frac{1}{r}\cos\varphi\frac{\partial f}{\partial \varphi}\right)\right) = \frac{e^{-i\varphi}}{2}\left(\frac{\partial f}{\partial r} - \frac{i}{r}\frac{\partial f}{\partial \varphi}\right)$$

$$f'(\bar{k}) = \frac{\partial f(\bar{k})}{\partial \bar{k}} = \frac{1}{2}\left(\cos\varphi\frac{\partial f}{\partial r} - \frac{1}{r}\sin\varphi\frac{\partial f}{\partial \varphi} + i\left(\sin\varphi\frac{\partial f}{\partial r} + \frac{1}{r}\cos\varphi\frac{\partial f}{\partial \varphi}\right)\right) = \frac{e^{i\varphi}}{2}\left(\frac{\partial f}{\partial r} + \frac{i}{r}\frac{\partial f}{\partial \varphi}\right).$$

- Kime **Wirtinger integration**:

Path-integral $\lim_{|z_{i+1}-z_i|\to 0} \sum_{i=1}^{n-1}\left(f(z_i)(z_{i+1}-z_i)\right) \cong \oint_{z_a}^{z_b} f(z_i)dz$.

Definite area integral: for $\Omega \subseteq \mathbb{C}$, $\int_\Omega f(z)dzd\bar{z}$.

Indefinite integral: $\int f(z)dzd\bar{z}$, $df = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \bar{z}}d\bar{z}$.

The *Laplacian* in terms of conjugate pair coordinates is $\Delta f = d^2 f = 4\frac{\partial f}{dz}\frac{\partial f}{d\bar{z}} = 4\frac{\partial f}{d\bar{z}}\frac{\partial f}{dz}$.

Dinov & Velev (2021)

42

Quantum Mechanics, AI & Data Science

Mathematical-Physics	Data Science
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, $\boldsymbol{O} = \{X, Y\}$
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
…	…

43

Quantum Mechanics, AI & Data Science

Math-Physics	Data Science
Wavefunction Wave equ problem: $\left(\frac{\partial^2}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t}\right)\psi(x,t) = 0$ Complex Solution: $\psi(x,t) = Ae^{i(kx-wt)}$ where $\left\lvert\frac{w}{k}\right\rvert = v$, represents a traveling wave	*Inference function* - describing a solution to a specific data analytic system (a problem). For example, • A linear (GLM) model represents a solution of a prediction inference problem, $\boldsymbol{Y} = \boldsymbol{X}\beta$, where the inference function quantifies the effects of all independent features ($\boldsymbol{X}$) on the dependent outcome ($\boldsymbol{Y}$), data: $\boldsymbol{O} = \{\boldsymbol{X}, \boldsymbol{Y}\}$: $\psi(\boldsymbol{O}) = \psi(\boldsymbol{X}, \boldsymbol{Y}) \Rightarrow \hat{\beta} = \hat{\beta}^{OLS} = \langle X\|X\rangle^{-1}\langle X\|Y\rangle = \left(X^TX\right)^{-1}X^TY$. • A non-parametric, non-linear, alternative inference is SVM classification. If $\boldsymbol{\psi_x} \in \boldsymbol{H}$, is the lifting function $\boldsymbol{\psi}: \boldsymbol{R^\eta} \to \boldsymbol{R^d}$ ($\boldsymbol{\psi}: \boldsymbol{x} \in \boldsymbol{R^\eta} \to \tilde{x} = \boldsymbol{\psi_x} \in \boldsymbol{H}$), where $\boldsymbol{\eta} \ll \boldsymbol{d}$, the kernel $\boldsymbol{\psi_x}(\boldsymbol{y}) = \langle \boldsymbol{x}\|\boldsymbol{y}\rangle: \boldsymbol{O} \times \boldsymbol{O} \to \boldsymbol{R}$ transformes non-linear to linear separation, the observed data $\boldsymbol{O_i} = \{\boldsymbol{x_i}, \boldsymbol{y_i}\} \in \boldsymbol{R^\eta}$ are lifted to $\boldsymbol{\psi_{O_i}} \in \boldsymbol{H}$. Then, the SVM prediction operator is the weighted sum of the kernel functions at $\boldsymbol{\psi_{O_i}}$, where $\boldsymbol{\beta^*}$ is a solution to the SVM regularized optimization: $\langle \psi_O \| \beta^* \rangle_H = \sum_{i=1}^{n} p_i^* \langle \psi_O \| \psi_{O_i} \rangle_H$ The linear coefficients, $\boldsymbol{p_i^*}$, are the dual weights that are multiplied by the label corresponding to each training instance, $\{\boldsymbol{y_i}\}$. Inference always depends on the (input) data; however, it does not have 1-1 and onto bijective correspondence with the data, since the inference function quantifies predictions in a probabilistic sense.

GLM/SVM: http://DSPA.predictive.space | Dinov, Springer (2018)

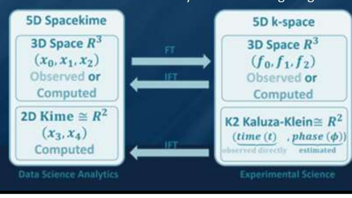
44

Spacekime Analytics

- Let's assume that we have:
  (1) Kime extension of Time, and
  (2) Parallels between wavefunctions ↔ inference functions
- Often, we can't directly observe (record) data natively in 5D spacekime.
- Yet, we can measure quite accurately the kime-magnitudes (r) as event orders, "times".
- To reconstruct the 2D spatial structure of kime, borrow tricks used by crystallographers [1] to resolve the structure of atomic particles by only observing the magnitudes of the diffraction pattern in k-space. This approach heavily relies on (1) prior information about the kime directional orientation (that may be obtained from using similar datasets and phase-aggregator analytical strategies), or (2) experimental reproducibility by repeated confirmations of the data analytic results using longitudinal datasets.

Spacetime → Spacekime Transforms

(1) Phase-estimation
(2) Phase-modeling
(3) Laplace Transform

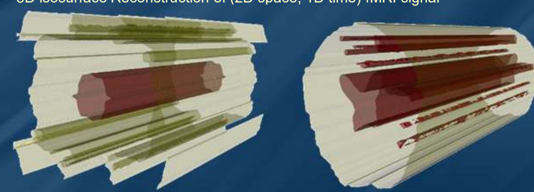


[1] Rodriguez, Ivanova, Nature 2015

45

Spacekime Analytics: fMRI Example

- 3D isosurface Reconstruction of (2D space, 1D time) fMRI signal



4D spacetime: Reconstruction using trivial phase-angle; kime=time=(magnitude, 0)

5D Spacekime: Reconstruction using correct kime=(magnitude, phase)

3D pseudo-spacetime reconstruction:

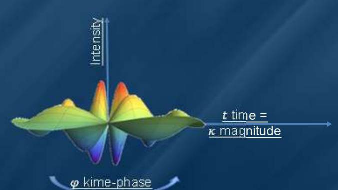
$$f = \hat{h}\,(\underbrace{x_1, x_2}_{space},\ \underbrace{t}_{time})$$

46

Spacekime Analytics: Kime-series = Surfaces (not curves)

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.



47

Bayesian Inference Representation

- We can formulate spacekime inference as a Bayesian parameter estimation problem:

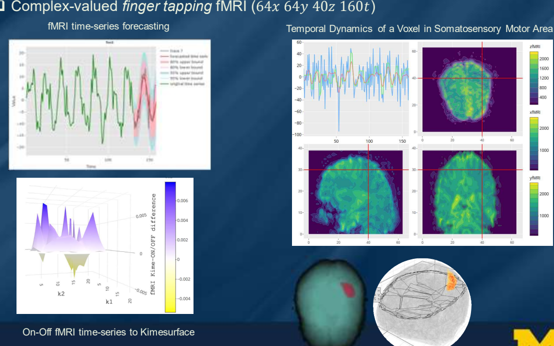
$$\underbrace{p(\gamma|X,\varphi')}_{\text{posterior distribution}} = \frac{p(\gamma,X,\varphi')}{p(X,\varphi')} = \frac{p(X|\gamma,\varphi')\times p(\gamma,\varphi')}{p(X,\varphi')} = \frac{p(X|\gamma,\varphi')\times p(\gamma,\varphi')}{p(X|\varphi')\times p(\varphi')} =$$
$$= \frac{p(X|\gamma,\varphi')}{p(X|\varphi')}\times\frac{p(\gamma,\varphi')}{p(\varphi')} = \frac{p(X|\gamma,\varphi')\times p(\gamma|\varphi')}{\underbrace{p(X|\varphi')}_{\text{observed evidence}}} \propto \underbrace{p(X|\gamma,\varphi')}_{\text{likelihood}}\times\underbrace{p(\gamma|\varphi')}_{\text{prior}}.$$

- In Bayesian terms, the posterior probability distribution of the unknown parameter γ is proportional to the product of the likelihood and the prior.
- In probability terms, the posterior = likelihood times prior, divided by the observed evidence, in this case, a single spacetime data point, x_{i_o}.

48

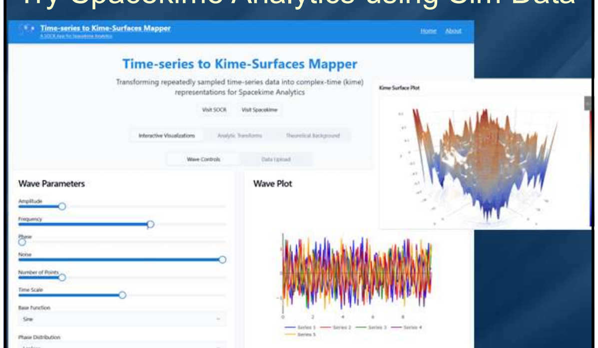
Spacekime Analytics using fMRI

- Complex-valued *finger tapping* fMRI ($64x\ 64y\ 40z\ 160t$)



49

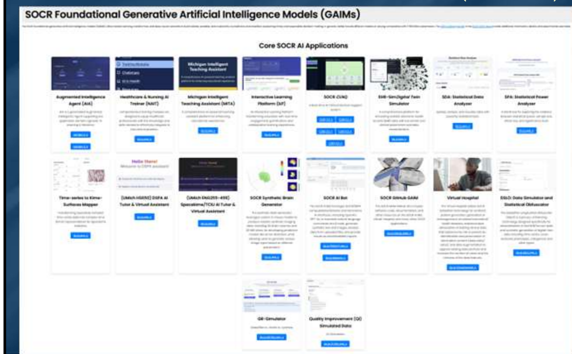
Try Spacekime Analytics using Sim Data



https://Kime.StatisticalComputing.org

50

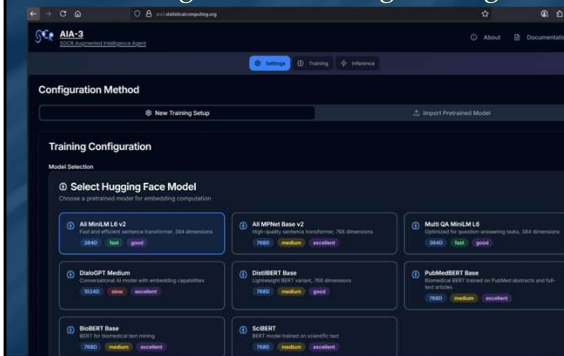
Foundational Generative AI Models (GAIMs)



https://socr.umich.edu/GAIM/

52

SOCR Augmented Intelligence Agent



https://AIA3.statisticalcomputing.org/

53

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* ≫ *Moore Law* trend
- **Web**: www.SOCR.umich.edu
- **Git**: https://github.com/SOCR
- **Projects**: www.socr.umich.edu/html/SOCR_Research.html
- **Apps**: https://socr.umich.edu/HTML5/

Share

55

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, Achu Shankar, Ryan Kwong, 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

56