Pillars of Open-Science

Sources: Characteristics of Big Biomed Data

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

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

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/cm³)

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
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 $\gamma$) power
- Increase of Kryder’s Law (Data volume) $\gg$ Moore’s Law (Compute power)
- Exponential decay of data-value
- Incent innovation, transdisciplinary collaborations, and knowledge dissemination
- ...

FAIR = Findable + Accessible + Interoperable + Reusable

Infrastructure: Cloud Ecosystem

Infrastructure: Cranium/Pipeline

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/DCICR)
    - https://cran.r-project.org
  - Licensing
    - https://www.gnu.org/licenses (e.g., https://en.wikipedia.org/wiki/Open_source)

Findings: Open Science Career Assessment Matrix

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 to address regional administrative changes and provide services to the organizational research community in an efficient and effective manner.
  - Ensuring the implementation of appropriate service agreements, policies, and practices to achieve organizational goals.
- Academic standing: Engaging in public outreach activities to promote the organization’s vision and mission.
  - Participating in public debates, forums, and interviews.
- Professional responsibility: Contributing to the organization’s sustainability planning and implementation.
  - Ensuring the organization’s sustainability planning and implementation.
- Interwoven: Contributions to the organization’s sustainability planning and implementation.
  - Ensuring the organization’s sustainability planning and implementation.

**Communication and dissemination**
- Stakeholders in Digital: Integrating new digital requirements into the organization.
  - Ensuring the integration of new digital requirements into the organization.
- IP strategy (big data): Developing a data governance strategy.
  - Ensuring the development of a data governance strategy.
- Knowledge management: Developing new knowledge management strategies.
  - Ensuring the development of new knowledge management strategies.
- Teaching: Integrating new teaching requirements into the organization.
  - Ensuring the integration of new teaching requirements into the organization.
- Reporting: Establishing new reporting requirements.
  - Ensuring the establishment of new reporting requirements.
- Educating professional short-term: Developing new professional short-term training programs.
  - Ensuring the development of new professional short-term training programs.
- Project management: Integrating our new project requirements into the organization.
  - Ensuring the integration of our new project requirements into the organization.
- Service provision: Developing new service provision requirements.
  - Ensuring the development of new service provision requirements.

**Rationale for Open Science (Pros)**
- We are always stronger together.
- Long-term sustainability prefers diversity.
- Optimized investments: career advancement, impact & cost-efficiency.
- Expediency: discovery, innovation, production & impact.
- Rapid devaluation: data-handling, clandescence science, knowledge obliteration.


**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.

![Image](https://datafinder.org)

**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/quality of the science or tools may be difficult to evaluate on mass.
- 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?

![Image](https://datafinder.org)

**DataSifter**
- A detailed description and [DataSifter](https://datafinder.org) R method implementation are available on our GitHub repository ([Available here](https://github.com/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.

<table>
<thead>
<tr>
<th>Table 1: Oblfuscation of multivariate data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>$P$</td>
</tr>
<tr>
<td>$K$</td>
</tr>
<tr>
<td>$\alpha$</td>
</tr>
<tr>
<td>$\theta$</td>
</tr>
<tr>
<td>$\phi$</td>
</tr>
</tbody>
</table>

**Table:**

<table>
<thead>
<tr>
<th>Oblfuscation</th>
<th>$P$</th>
<th>$K$</th>
<th>$\alpha$</th>
<th>$\theta$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>Small</td>
<td>0.05</td>
<td>2</td>
<td>0.6</td>
<td>0.05</td>
<td>0.3</td>
</tr>
<tr>
<td>Medium</td>
<td>0.25</td>
<td>2</td>
<td>0.6</td>
<td>0.05</td>
<td>0.3</td>
</tr>
<tr>
<td>Large</td>
<td>0.5</td>
<td>2</td>
<td>0.6</td>
<td>0.05</td>
<td>0.3</td>
</tr>
<tr>
<td>Indep</td>
<td>0.5</td>
<td>2</td>
<td>0.6</td>
<td>0.05</td>
<td>0.3</td>
</tr>
</tbody>
</table>
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.

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).

III. Clinical Data Application: Using DataSifter to Obfuscate the ABIDE Data
Comparing the Original and “Sifted” Data for the 22nd ABIDE Subject

IV. Clinical Data Application: Using DataSifter to Obfuscate the ABIDE Data
PIFV for ABIDE under different levels of DataSifter obfuscations. Each box represents 1098 subjects among the ABIDE sub-cohort. Percent of Identical Feature Values (PIFV) under Different Privacy Levels. Binary outcome refers to the first experiment; Count refers to the second experiment; Continuous refers to the third experiment.

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

Data Source
- UK Biobank
- Demographics: > 5000 cases
- Clinical data: > 80 features
- Imaging data: T1, resting-state fMRI, task fMRI, T2, FLAIR, dwMRI, T2, SWI
- Genetics data

Summary
- The longitudinal archive of the UK population

http://www.ukbiobank.ac.uk
Zhou, et al. (2019), SREP

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>3,461 (75.3%)</td>
<td>1,257 (24.7%)</td>
</tr>
<tr>
<td>Depression feelings</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>2,327 (51.4%)</td>
<td>2,332 (52.1%)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>2,491 (54.5%)</td>
<td>2,497 (54.5%)</td>
</tr>
<tr>
<td>Ever depressed for a whole week</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>3,344 (65.7%)</td>
<td>1,367 (29.8%)</td>
</tr>
<tr>
<td>Ever unenthusiastic/disinterested for a whole week</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>4,038 (89.3%)</td>
<td>1,024 (22.3%)</td>
</tr>
</tbody>
</table>

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 p-value includes the percentage of subjects being labeled as “yes” and “no”, 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

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

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

---

### Complex-Time (Kime)

- **At a given spatial location, \( x \), complex time (kime) is defined by \( x = r e^{i\phi} \in \mathbb{C} \), where:**
  - the magnitude represents the longitudinal events order \( r \geq 0 \) and characterizes the longitudinal displacement in time, and
  - event phase \( (-\pi \leq \phi < \pi) \) is an angular displacement, or event direction.
- **Space-kime manifold is \( \mathbb{R}^2 \times \mathbb{C} \):**
  - \( (x, k_1) \) and \( (x, k_4) \) have the same spacetime representation, but different spacekime coordinates,
  - \( (x, k_1) \) and \( (y, k_1) \) share the same kime, but represent different spatial locations,
  - \( (x, k_2) \) and \( (x, k_3) \) have the same spatial locations and kime directions, but appear sequentially in order.

---

### Kime Parameterizations

#### Conjugate Pairs (\( (r, \phi) \in \mathbb{C} \))

- **Cartesian (\( x, y \in \mathbb{R}^2 \))**
- **Polar (\( (r, \phi) \in \mathbb{R}^+ \times (-\pi; \pi] \))**

---

### Longitudinal Data Analytics

- **Neuroimaging:**
  - 4D fMRI: time-series, represents measurements of hydrogen atom densities over a 3D lattice of spatial locations (1 ≤ \( x, y, z \leq 64 \) pixels), about 3 x 3 millimeters\(^2\) resolution. Data is recorded longitudinally over time (1 ≤ \( t \leq 180 \)) 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, \( x = r e^{i\phi} \in \mathbb{C} \).
- **Differences:** Spacekime analytics estimate and utilize the kime-phases.

---

### Spacekime Calculus

- **Kime Wirtinger derivative** (first order kime-derivative at \( k = (r, \phi) \)):
  \[ f'(k) = \frac{\partial f}{\partial r} \frac{1}{2} \left( \frac{\partial f}{\partial r} + i \frac{\partial f}{\partial \phi} \right) \]
  In Conjugate-pair basis:
  \[ df = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi \]
  In Polar kime coordinates:
  \[ f'(k) = \frac{\partial f}{\partial r} \frac{1}{2} \left( \cos \phi \frac{\partial f}{\partial r} + \sin \phi \frac{\partial f}{\partial \phi} \right) \]

- **Kime Wirtinger integration**:
  - Path integral:
    \[ \text{Path integral} = \int_{z_1}^{z_2} f(z) dz = \int_{z_1}^{z_2} f(z) dz + \frac{1}{2} \cdot \text{Conjugate-pair integral} \]
  - Path integral for \( \Omega \subseteq \mathbb{C} \):
    \[ \int_{\Omega} f(z) dz \]
  - Inffinite integral:
    \[ \int f(z) dz = \frac{1}{2} \int f(z) \left( \cos \phi \frac{\partial f}{\partial r} + \sin \phi \frac{\partial f}{\partial \phi} \right) dz \]

The Laplacian in terms of conjugate pair coordinates is \( \Delta f = \frac{\partial^2 f}{\partial r^2} + \frac{\partial^2 f}{\partial \phi^2} \).
### Quantum Mechanics, AI & Data Science

#### Mathematical-Physics

<table>
<thead>
<tr>
<th>A particle</th>
<th>is a small object that can participate in observations and characterization of observable phenomena.</th>
</tr>
</thead>
<tbody>
<tr>
<td>An observable</td>
<td>is a dynamic variable about a particle that can be measured.</td>
</tr>
<tr>
<td>Particles exist</td>
<td>in collections of independent particles and observable characteristics, in a closed system.</td>
</tr>
<tr>
<td>Wave-function</td>
<td>represents a probability amplitude, which can be measured</td>
</tr>
<tr>
<td>Inference-function</td>
<td>is a solution to a specific data analytic system (a prediction inference problem).</td>
</tr>
<tr>
<td>Variables</td>
<td>are observed quantities or qualitative values, e.g., mass, energy, time, etc.</td>
</tr>
</tbody>
</table>

#### Data Science

<table>
<thead>
<tr>
<th>Math-Physics</th>
<th>Inference-function: describing a solution to a specific data analytic system (prediction inference problem).</th>
</tr>
</thead>
<tbody>
<tr>
<td>A traveling wave</td>
<td>represents a traveling wave</td>
</tr>
<tr>
<td>3D space</td>
<td>is a solution to the SVM regularized optimization</td>
</tr>
<tr>
<td>5D spacekime</td>
<td>is a solution to the SVM regularized optimization</td>
</tr>
</tbody>
</table>

### Spacekime Analytics

- Let’s assume that we have:
  - (1) kime-extension of Time, and
  - (2) the unknown internal states of the entire system don’t influence the computation of inference.

- Often, we can’t directly observe (record) data natively in 5D spacekime.

- To reconstruct the 2D spatial structure of kime, we can use tricks used by crystallographers.

- In the 5D spacekime manifold, we can use modeling functions to transform standard time-series curves to spacekime kime-surfaces, which can be modeled, interpreted, and predicted using advanced spacekime analytics.

### Spacekime Analytics: fMRI Example

- 3D and 4D spacekime Reconstruction using 3D time-angle (magnitudes, 0).
- 5D spacekime: Reconstruction using (kime-magnitudes, phase).

### Bayesian Inference Representation

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

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

- In Bayesian terms, the posterior probability distribution of the unknown parameter $$\theta$$ depends on the likelihood of the data and the prior distribution of the parameter.

In probability terms, the posterior distribution is proportional to the product of the likelihood and the prior.
Spacekime Analytics using fMRI

- Complex-valued finger tapping fMRI (64x64x40x160)
  - fMRI time-series forecasting
  - Temporal Dynamics of a Voxel in Somatosensory Motor Area

On-Off fMRI time-series to Kimesurface

Temporal Dynamics of a Voxel in Somatosensory Motor Area

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} \( \gg \) {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/

Acknowledgments

Funding
- NIH: P20 NR015331, U54 EB022465, P50 GM068588, P30 DK059543, UL1 TR002242, R21CA215487, NSF: 1904242, 1739883, 1628040, 1416920, 0714035, 1021115

Collaborators
- SOCR: Milen Velev, Yuezhang Shen, Daxuan Deng, Zijing Li, Yanghuai Gou, Zhe Ye, Yufei Yang, Yuxin Wang, Rongqian Zhang, Yuyao Liu, Yupeng Zhang, Yunjie Guo

https://SOCR.umich.edu

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