## 2015 JSM Session

## Big Data: Modeling, Tools, Analytics, & Training

**Organizer:** Ivo D. Dinov, Statistics Online Computational Resource (SOCR), Michigan
**Moderator/Chair**: Robin Jeffries, CSU Chico
**Logistics:**
- **Date/Time**: Mon, 8/10/2015, 10:30 AM - 12:20 PM
- **Venue**: Washington State Convention Center, CC-606

**Session**: #211347, Sponsor: Statistical Computing

### Speakers and Topics

- **Ivo Dinov (Michigan)**, Management, Modeling & Analytic Challenges of Big Biomedical Data
- **Max Robinson** & **Gustavo Glusman**, ISB ESPALIERS: A visualization method for Big Data
- **Moo K. Chung (Wisconsin)**, The computational challenges of constructing and visualizing large-scale brain networks
- **Ravi Madduri**, Argonne/Chicago, Big Data Services: Globus Online, Galaxy, GridFTP
- **Barzan Mozafari**, Large-scale Data Intensive Systems, Big Data, Interactive Data Processing

http://wiki.socr.umich.edu/index.php/SOCR_Events_JSS_2015

# Management, Modeling & Analytic Challenges of Big Biomedical Data

**Ivo D. Dinov**

Statistics Online Computational Resource
University of Michigan

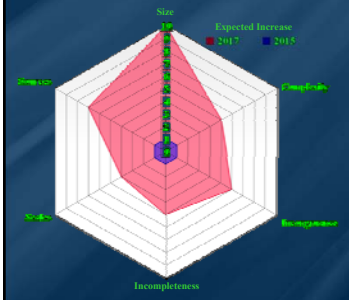
www.SOCR.umich.edu

STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)

## Outline

- Availability, Sharing, Aggregation and Services
- Classical Data Science vs. Innovative Big Data Analytics
  - Amateur Scientists vs. "Experts"
  - Data Scientists vs. Practitioners
  - Domain-specific vs. Trans-disciplinary knowledge
- Commercial vs. Open-source Resourceome
- Rapid Big Data Evolution
- Big Data IT proliferation
- Data democratization
- Big Data is incredibly time, space, protocol, context sensitive
- Big Data Science Training (opportunities and challenges)

## Characteristics of Big 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.

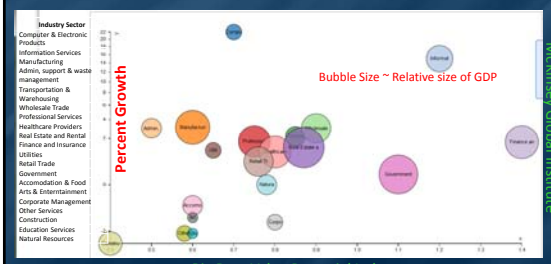
Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers

Mixture of quantitative & qualitative estimates

*Dinov, et al. (2014)*

## Availability, Sharing, Aggregation & Services

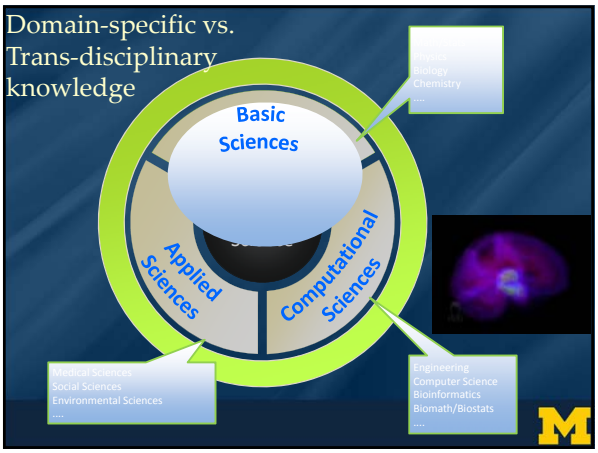
- There will be over 10 billion mobile-connected devices in 2016; i.e., there will be 1.3 mobile devices per capita




## Democratization of Big Data Science

- Doctorate training is not mandatory nor does it guarantee appropriate Big Data expertise
- Lower barriers of entry
- Demand for constant "Continuing Education" and self-training
- Dichotomy between theoretical and empirical sciences
- Differences between fundamental knowledge and experimental skills (big data properties closely approximate core scientific principles)

## Domain-specific vs. Trans-disciplinary knowledge




## Big Data Resourceome

- There is an explosion of open-data-science resources
  - www.data.gov
  - www.ncbi.nlm.nih.gov/gap
- Spawning of a number of industries and enterprises blending proprietary and open-source data, code, documentation, expert-support, infrastructure and services
- Big Data to Knowledge Initiative: www.BD2K.org
- Google Cloud Platform (GCP)
- Amazon Web Services (AWS)

## Big Data Analytics Resourceome


http://bd2k.org/BigDataResourceome.html

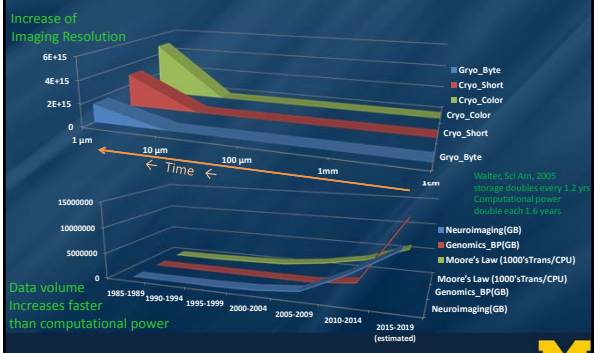

| Big Data | Information | Knowledge | Action |
|---|---|---|---|
| Raw Observations | Processed Data | Maps, Models | Actionable Decisions |
| Data Aggregation | Data Fusion | Causal Inference | Treatment Regimens |
| Data Scrubbing | Summary Stats | Networks, Analytics | Forecasts, Predictions |
| Semantic-Mapping | Derived Biomarkers | Linkages, Associations | Healthcare Outcomes |

## National Big Data to Knowledge (BD2K) Centers

1. Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data (U Pitt, Cooper, Bahar, Berg)
2. Center for Predictive Computational Phenotyping (U Wisconsin, Craven
3. National Center for Mobility Data Integration to Insight (The Mobilize Center) (Stanford, Delp)
4. KnowEng, a Scalable Knowledge Engine for Large-Scale Genomic Data The (U Illinois Urbana-Champaign, Han, Sinha, Sorg, Weinshilboum)
5. Center for Big Data in Translational Genomics (UC Santa Cruz, Haussler, Patterson)
6. Patient-Centered Information Commons (Harvard, Kohane)
7. Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K) (U Memphis, Kumar)
8. Center for Expanded Data Annotation and Retrieval (CEDAR) (Stanford, Musen)
9. A Community Effort to Translate Protein Data to Knowledge: An Integrated Platform (UCLA, Ping, Lindsey, Su, Watson)
10. ENIGMA Center for Worldwide Medicine, Imaging, and Genomics (USC, Thompson)
11. Big Data for Discovery Science (USC, Toga)

www.BD2K.org

## Kryder's law: Exponential Growth of Data



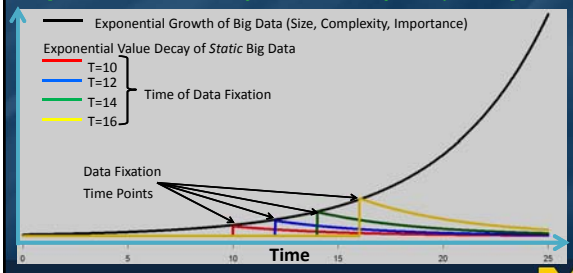

Dinov, et al., 2014

## Big Data Challenges


## Energy & Life-Span of Big Data

**Energy** encapsulates the holistic information content included in the data (exponentially increasing with time), which may often represent a significant portion of the joint distribution of the underlying process

**Life-span** refers to the relevance and importance of the Data (exponentially decreasing with time)

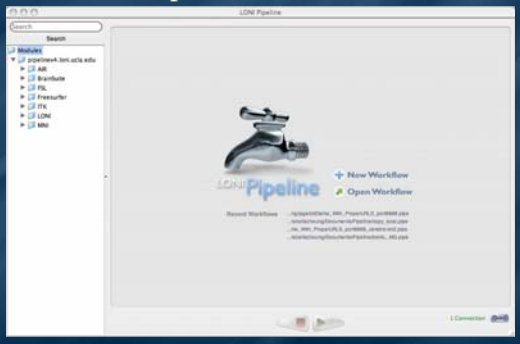



http://www.aaas.org/news/big-data-blog-part-v-interview-dr-ivo-dinov

## End-to-end Pipeline Workflow Solutions


Dinov, *et al.*, 2014, Front. Neuroinform.; Dinov, *et al.*, Brain Imaging & Behavior, 2013

## End-to-end Pipeline Workflow Solutions




Dinov, *et al.*, 2014, Front. Neuroinform.; Dinov, *et al.*, Brain Imaging & Behavior, 2013

## Neurodegenerative Disease Applications:

### Imaging-Genetic Biomarker Interactions in Alzheimer's Disease

**Goals**

Investigate AD subjects (age 55 – 65) using Neuroimaging Initiative (ADNI) database to understand early-onset (EO) cognitive impairment using neuroimaging and genetics biomarkers

**Data**

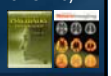
9 EO-AD and 27 EO-MCI

Derived 15 most impactful neuroimaging markers (out of 336 morphometry measures)

Obtained 20 most significant single nucleotide polymorphisms (SNPs) associated with specific neuroimaging biomarkers (out of 620K SNPs)

**Approach**

Global Shape Analysis (GSA) Pipeline workflow



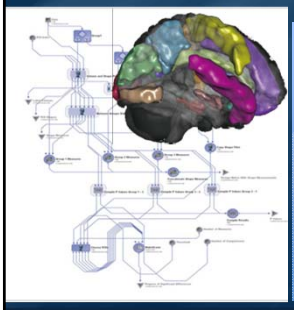
*Moon, Dinov et al., 2015, Psych. Investig.*
*Moon, Dinov et al., 2015, J Neuroimaging*

## Results: AD Imaging-Genetic Biomarker Interactions

- Identified associations between neuroimaging phenotypes and genotypes for the cohort of 36 EO subjects
- Overall most significant associations:
  - rs7718456 (Chr 15) and L_hippocampus (volume)
  - rs7718456 and R_hippocampus (volume)
- For the 27 EO-MCI's, most significant associations
  - rs6446443 (Chr 4, JAKMIP1 janus kinase and microtubule interacting protein 1 gene) and R_superior_frontal_gyrus (volume)
  - rs17029131 (Chr 2) and L_Precuneus (volume)

*Moon, Dinov et al., 2015, Psych. Investig.*

## Results: AD Imaging-Genetic Biomarker Interactions

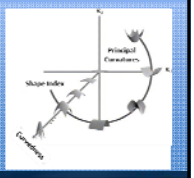


Shape Index

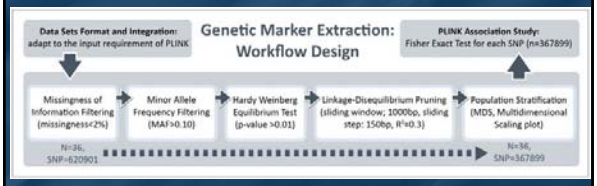
$$SI = \frac{2}{\pi}\arctan\left(\frac{\kappa_2+\kappa_1}{\kappa_2-\kappa_1}\right) \qquad -1 \le SI \le 1$$

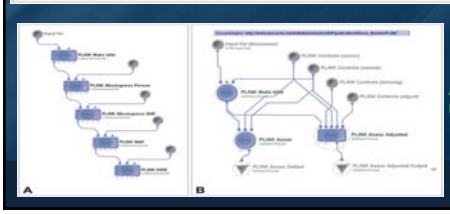
Curvedness

$$CV = \sqrt{\frac{\kappa_1^2+\kappa_2^2}{2}} \qquad 0 \le CV$$




## Results: AD Imaging-Genetic Biomarker Interactions

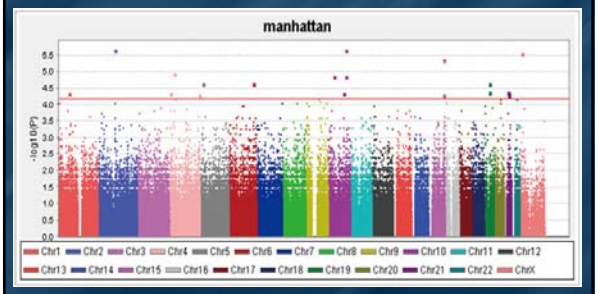



**QC Protocol**

A. QC Plink workflow
B. Genetic association study

## Results: AD Imaging-Genetic Biomarker Interactions

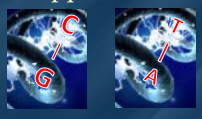



Manhattan plot for all the single nucleotide polymorphisms (SNPs)

## Results: AD Imaging-Genetic Biomarker Interactions

| Neuroimaging phenotypes | p-value | Index | SNPs | Chromosome | p-value | Gene |
|---|---|---|---|---|---|---|
| L_cingulate_gyrus (Average mean curvature) | 0.0335 | 1 | rs17029131 | 2 | 3.52E-06 | |
| L_gyrus_rectus (Surface area) | 0.01728 | 2 | rs1822144 | 2 | 2.28E-06 | |
| R_cuneus (Surface area) | 0.04203 | 3 | rs6446443 | 4 | 6.68E-05 | JAKMIP1 (janus kinase & microtubule interacting protein 1) |
| R_superior_frontal_gyrus (Volume) | 0.03706 | 4 | rs12506164 | 4 | 1.75E-05 | |
| L_precentral_gyrus (Volume) | 0.04125 | 5 | rs7718456 | 5 | 3.36E-05 | |
| L_precuneus (Volume) | 0.0508 | 6 | rs9377090 | 6 | 3.36E-05 | |
| L_middle_occipital_gyrus (Volume) | 0.01805 | 7 | rs2776932 | 10 | 2.20E-05 | NRP1 (neuropilin1) |
| R_superior_temporal_gyrus (Volume) | 0.03353 | 8 | rs4933672 | 10 | 6.48E-05 | |
| L_hippocampus (Volume) | 0.00067 | 9 | rs11193270 | 10 | 3.52E-06 | |
| R_hippocampus (Volume) | 0.00539 | 10 | rs11193272 | 10 | 3.52E-06 | |
| R_precentral_gyrus (Shape index) | 0.03411 | 11 | rs11193274 | 10 | 3.52E-06 | |
| R_precuneus (Shape index) | 0.03186 | 12 | rs12218153 | 10 | 3.52E-06 | |
| L_cuneus (Shape index) | 0.04952 | 13 | rs1338956 | 10 | 2.20E-05 | |
| R_inferior_occipital_gyrus (Curviness) | 0.05037 | 14 | rs1338025 | 10 | 2.20E-05 | |
| R_putamen (Curviness) | 0.03504 | 15 | rs12101936 | 15 | 7.08E-06 | |
| | | 16 | rs16964473 | 19 | 3.53E-05 | Intergenic |
| | | 17 | rs12972537 | 19 | 6.14E-05 | |
| | | 18 | rs2212356 | 21 | 6.23E-05 | |
| | | 19 | rs2831165 | 21 | 6.68E-05 | |
| | | 20 | rs1266320 | 23 | 4.46E-06 | |

## GWAS Imaging-Genetics Approach







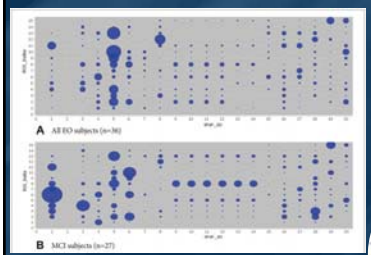
- **SNPs**
  - ***E.g., C/T polymorphism***
- Model
  - ***Phenotype:*** $Y_i$ be the imaging-biomarker for $i^{th}$ *subject*
  - ***Genotype:*** $X_i$ be the genotype $i^{th}$ *subject* at a particular:

$$\text{SNP } X_i = \begin{cases} 0, & BB \\ 1, & BA \text{ or } AB \\ 2, & AA \end{cases}$$

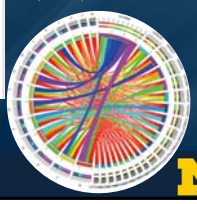
- SOCR Multivariate Regression Models
  - $Y_i = \beta_0 + \beta_1 X_i$
  - $In\ general, Y_i = \sum_{k=0}^{K} \beta_k X_i^{(k)} + \varepsilon$
  - Stat analysis: $\boldsymbol{\beta_k \neq 0}$

| Genotype-Phenotype Relation | | | |
|---|---|---|---|
| Parents | | ♂ B | ♂ A |
| ♀ B | | BB | BA |
| ♀ A | | BA | AA |

## SNP-Neuroimaging interactions in Alzheimer's Disease




- Overall Associations
  - rs7718456 and L_hippocampus (volume)
  - rs7718456 and R_hippocampus (volume)
- EO-MCI associations
  - rs6446443 R_superior_frontal_gyrus (volume)
  - rs17029131 and L_Precuneus (volume)



## Neurodegenerative Disease Applications: Imaging-Genetic Biomarker Interactions in Alzheimer's Disease

**Goals**

Investigate AD subjects (age 55 – 65) using Neuroimaging Initiative (ADNI) database to understand early-onset (EO) cognitive impairment using neuroimaging and genetics biomarkers

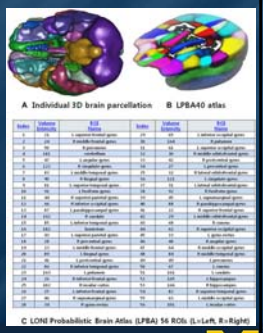
**Results**

Generated mean geometric models of the left and right hippocampi, middle frontal gyri, and the middle temporal gyri.

Linear model statistical maps, at each vertex on the triangulated shapes. **Dependent variable** was the radial distance morphometry measure (deviation of individual shape model from mean shape atlas) and **independent regressors** including diagnosis, age, education years, APOE (ε4), MMSE, visiting times, and logical memory (immediate and delayed recall).

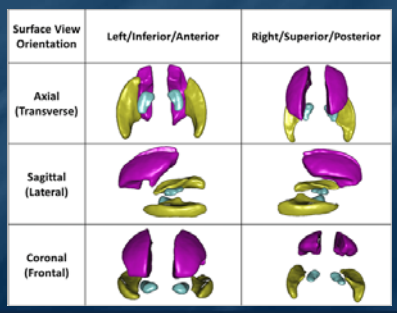
**Approach**

Local Shape Analysis (LSA) Pipeline workflow



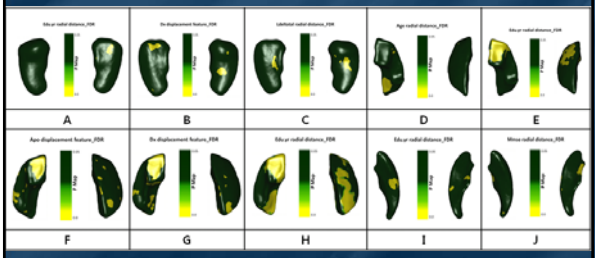

*Moon, Dinov et al., 2015, J Neuroimaging*

## Neurodegenerative Disease Applications: Imaging-Genetic Biomarker Interactions in Alzheimer's Disease




*Moon, Dinov et al., 2015, Neuroimaging.*

## Neurodegenerative Disease Applications: Imaging-Genetic Biomarker Interactions in Alzheimer's Disease




*P-maps (A) for Lt. hippocampus, P-maps (B,C) for Rt. Hippocampus, P-maps (D,E) for Lt. middle frontal gyrus, P-maps (F,G,H) for Rt. middle frontal gyrus, P-maps (I, J) for Lt. middle temporal gyrus.*

*Moon, Dinov et al., 2015, J Neuroimaging*

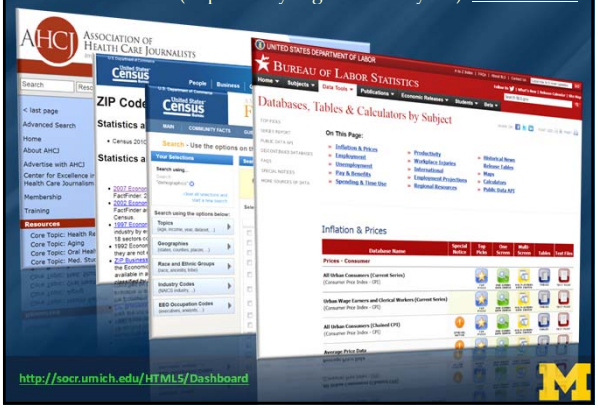
## SOCR Big Data Dashboard


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

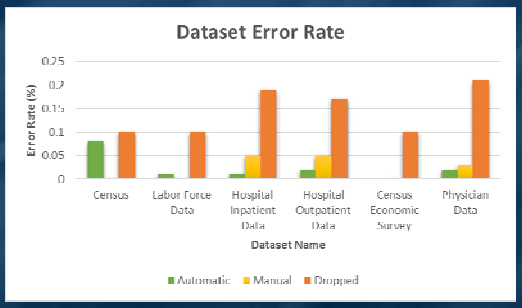
*Husain, et al., 2015, J Big Data*

## SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion




http://socr.umich.edu/HTML5/Dashboard

## SOCR Dashboard (Exploratory Big Data Analytics): Data QC




http://socr.umich.edu/HTML5/Dashboard

## Training: Big Data Skills

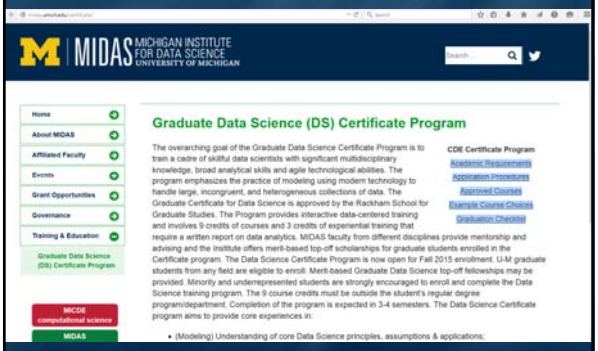
1) **Listening**: streams, information and language, analyzing sentiment, intent and trends;
2) **Looking**: searching, indexing and memory management of heterogeneous datasets; Loading: Raw, derived or indexed data as well as meta-data;
3) **Programming**: Handling Map-Reduce/HDFS, No-SQL DB, protocol provenance, pipeline workflows;
4) **Inferring**: Principals of data analyses, Bayesian modeling, inference, uncertainty and quantification of likelihoods; Connecting: Reasoning, logic and its limits, dealing with uncertainty; Analytics: Regression, feature selection, dimensionality reduction, temporal patterns;
5) **Learning**: Classification, clustering, mining, information extraction, knowledge retrieval, decision making;
6) **Predicting**: Forecasting, neural models, deep learning, and research topics;
7) **Summarizing**: Presentation of data, processing protocol, analytics provenance, visualization

## Training: Core Proficiencies

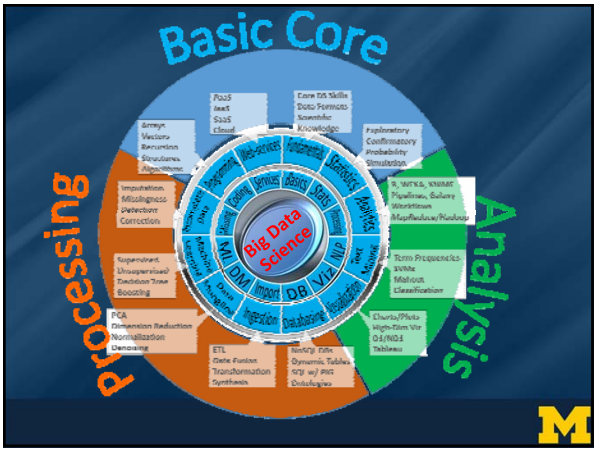
**The Data Science Certificate program aims to ensure that students awarded this certificate would have the following experiences:**

1) (**Modeling**) Understanding of core Data Science principles, assumptions and applications
2) (**Technology**) Knowledge of basic protocols for data management, processing, computation, information extraction & visualization
3) (**Practice**) Hands-on experience with modeling tools and technology resources in a real project setting

## University of Michigan Data Science Training




http://midas.umich.edu/training




## Acknowledgments

**Funding**

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

NSF: DUE 1416953, 0716055, 1023115

**Collaborators**

- **SOCR**: Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Ashwini Khare, Rami Elkest, Abhishek Chowdhury, Patrick Tan, Gary Chan, Andy Foglia, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Nicolas Christou, Rob Gould
- **AtheyLab**: Alex Ade, Ari Allyn-Feuer, Gen Zheng, John Wiley, David Dilworth, Walter Meixner, Jeffrey R. de Wet, Kevin Smith, Amy Creekmore, Indika Rajapakse, Gerry Higgins, Robert W. Veltri, Brian Athey
- **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, Fabio Macciardi, Federica Torri, Carl Kesselman