

2014 JMM/AMS Special Session

Big-Data: Mathematical and Statistical Modeling, Tools, Services, and Training

<http://ucla.in/16foQ4P>



Big Data Challenges in Neuroimaging, Informatics and Genomics Computing

Ivo D. Dinov

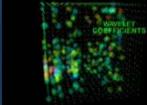
Statistics Online Computational Resource
University of Michigan

www.SOCR.umich.edu

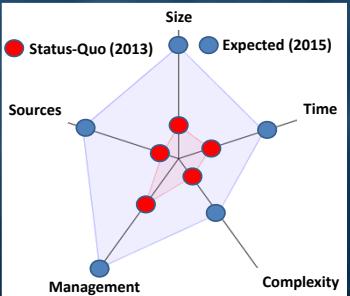
M SCHOOL OF NURSING STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR) UNIVERSITY OF MICHIGAN **M**

Outline

- Big Data
 - Volume/Size: Petabytes of Data (10^{15} Bytes)
 - Heterogeneity: (un)formatted, ASCII/Binary
 - Velocity: change, transfer, discovery
- Computational Challenges
 - 1,000's of Diverse Software Tools
 - 1,000,000's of Dispersed Hardware Devices
- Applications
 - NGS Analysis
 - Neuroimaging-genetics (ADNI MCI conversion)




Big Data Dimensions



M

Kryder's law: Exponential Growth of Data

VOLUME OF DATA MB = MEGABYTE = 10^6 , GB = GIGABYTE = 10^9 TB = TERABYTE = 10^{12} , PB = PETABYTE = 10^{15}				COMPUT. POWER	YEARS
SINGLE CRYO BRAIN VOLUME 1600 CM²				NEUROIMAGING (ANNUALLY)	GENOMICS (BP/Yr)
Voxel Resolution	Gray Scale	Color		200 GB	10 MB
Size	Count	8bits	16bits	1 TB	100 MB
1cm	12x15x9	1620	3000	4860	50 TB
1mm	120x 150x90	1.62	3.24 MB	4.86 MB	250 TB
100 µm	1200x 1500x900	1.62 GB	3.24 GB	4.86 GB	1 PB
10 µm	12000x 15000x 9000	1.62 TB	3.24 TB	4.86 TB	5 PB
1 µm	120000x 150000x 90000	1.62 PB	3.24 PB	4.86 PB	10+ PB
					20+ PB
					1×10^{11}
					(estimated)

Dinov, et al., 2013



Many 1,000's of Software Tools

- Acquisition, processing, storage/DB, service, migration, mining, analysis, visualization, annotation, ... (*data-driven process understanding*)
- Biomedical Imaging
 - There are 100's of different *types* of image processing algorithms and filters
 - For each type of process there may be dozens of concrete software products (instance implementations)
- (Example) Neuroimaging
 - Only NITRC lists > 500 openly shared software tools
 - For each openly shared tool there may be dozens proprietary or less commonly used analogues
- Genomics/Informatics
 - Over 200 Data and Cloud Computing Service Providers
 - Over 200 Public/Private/Non-Profit orgs that provide 1,000's of stand-alone tools

(Eliceiri, et al., *NMeth*, 2012)






Software Tools Discovery

- Acquisition, processing, storage/DB, service, migration, mining, analysis, visualization, annotation, ... ("data-driven) process understanding"

Dinov, et al., 2008

Millions of Dispersed Hardware Devices

- Cisco: "By the end of 2012, the number of mobile-connected devices will exceed the number of people on Earth"
- There will be over 10 billion mobile-connected devices in 2016; i.e., there will be 1.3 mobile devices per capita
 - These include phones, tablets, laptops, handheld gaming consoles, e-readers, in-car entertainment systems, digital cameras, and "machine-to-machine modules"
- DBs, Clients, Servers, Compute-Nodes, Web-Services, Interfaces, ...
- Solution ...

Dinov et al., BMC 2011

What is the Pipeline Environment?

- Pipeline.loni.usc.edu
- Graphical Workflow Interface to computational libraries and informatics resources
- Design, validation, execution, monitoring and dissemination of heterogeneous workflows
- Tool discovery
- Tool interoperability
- Distributed computing
- User-friendly access to
 - Data/Services
 - Hardware infrastructure
 - Computational expertise
 - Cloud Computing

Torri et al., Genes, 2012

Genomics Pipeline Solutions

- Integrated Bioinformatics (MAQ, SAMTools, Bowtie)
- mrFAST Indexing Mapping
- GWAS Impute
- EMBOSS (e.g., Matcher)
- BLAST
- BATWING
- GENEPOL
- PLINK Association
- Migrate
- Many others

<http://pipeline.loni.usc.edu/explore/library-navigator>

Pipeline Server Library Navigator

Workflow Detail Information

Workflow Processing

Workflow Inputs

Incomplete Module

Workflow Outputs

Complete Module

Workflow

- Validate
- Play
- Pause
- Stop
- Status

Pipeline

- Servers
- Tool Library
- Data
- Protocols

Server

- Login
- Load

Workflow Processing

Workflow Inputs

Incomplete Module

Workflow Outputs

Complete Module

Workflow

- Validate
- Play
- Pause
- Stop
- Status

Pipeline

- Servers
- Tool Library
- Data
- Protocols

Server

- Login
- Load

Grid & Cloud Computing

- LONI Grids

Cerebro	Cranium
<ul style="list-style-type: none"> • 1,200 cores • 1.4TB RAM • 12,000 jobs/day • 700 users 	<ul style="list-style-type: none"> • 4,300 cores • 9.6 TB RAM • (new)
- Amazon Cloud
 - EC2 (Elastic Cloud Computing)
 - S3 (Simple Storage Service)
- UC Grid
- Globus GridFTP
- UMICH ACR/Flux
- UMSN SOCR Pipeline Server (1.5TB RAM) available now ...

M

Examples of Validated NGS Workflows

Process	Software	Website
(0) Preprocessing	homemade scripts	many
	MAQ	http://maq.sourceforge.net
	BWA	http://bio-bwa.sourceforge.net/bwa.html
	BWA-SW (SE only)	http://bio-bwa.sourceforge.net/bwa.html
	PERM	http://code.google.com/p/perm/
	BOWTIE	http://bowtie-bio.sourceforge.net
	SOPAR-Z	http://www.sopar.org
	MOSAIK	http://bioinformatics.bc.edu/marthlab/MosaiK
	NOVOALIGN	http://www.novocraft.com/
(1a) Alignment		
	VELVET	http://www.ebi.ac.uk/~doherty/velvet
	SOPAgenovo	http://sopagenomics.org.cn
	ABYSS	http://www.bcgsc.ca/platform/bioinfo/software/abyss
(2) Basic QC	SAMTOOLS	http://sourceforge.net/projects/samtools/files/
	PICARD	http://picard.sourceforge.net/CommandLineOverview.shtml
(3) Advanced QC	GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
	PICARD	http://picard.sourceforge.net/
	SAMTOOLS	http://sourceforge.net/projects/samtools/files/
	IGVtools	http://www.broadinstitute.org/gpu/igvtools

(Torri, et al., Genes, 2012)



Examples of Validated NGS Workflows

Process	Software	Website
(4) Variant Calling and Annotation		
Sequence Variant Analyzer v1.0, for hg18 annotations	SVA	http://www.svaproject.org/
	SAMTOOLS	http://sourceforge.net/projects/samtools/files/
Sequence Variant Analyzer v1.1, for hg19 annotations	SVA	http://www.svaproject.org/
	SAMTOOLS	http://sourceforge.net/projects/samtools/files/
LGRS	LGRS	http://www.duke.edu/~m33/eris.htm
SAMTOOLS and ANNOVAR for annotation	SAMTOOLS	http://sourceforge.net/projects/samtools/files/
	ANNOVAR	http://www.openbioinformatics.org/annovar/
UnifiedGenotyper and ANNOVAR for annotation	GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
	ANNOVAR	http://www.openbioinformatics.org/annovar/
(5) Copy Number Variation		
CNVseq	CNVseq	http://iger.dbs.msu.edu.sg/cnv-seq/
	R	http://www.r-project.org/
SAMTOOLS/ERDS/Sequence variant analyzer v1.0 ERDS	SVA	http://www.svaproject.org/
	SAMTOOLS	http://sourceforge.net/projects/samtools/files/
	ERDS	http://www.duke.edu/~m33/erds.htm
SAMTOOLS/ERDS/Sequence variant analyzer v1.1 ERDS	SVA	http://www.svaproject.org/
	SAMTOOLS	http://sourceforge.net/projects/samtools/files/
	ERDS*	http://www.duke.edu/~m33/erds.htm

M

Preprocessing Example

- Hierarchical workflow approach for analyzing NGS data
- Several pipelines can be run independently or logically connected
- Once the reads have been pre-processed, they can be aligned, undergo basic/advanced QC, SNP/Indel and CNVs calling & annotation

M

Perfect Neuroimaging-Genetics-Computation Storm?

- Single Subject Studies (N=1)
 - Genetics:
 - Depending on Coverage(X)
 - Whole Genome Seq Data > 320GB (>80X)
 - Require 2+ TB RAM, and 100+ hrs CPU
 - Imaging:
 - Depending on protocols
 - 40-512 gradient directions Diffusion imaging data
 - Raw (multimodal) Neuroimaging Data > 10 GB
 - Derived Data > 100 GB
 - Require 100GB RAM and 70+ hrs CPU
- Large Subject Studies
 - Cohort studies (N>10, Typically N~100's)
 - Multi-institutional Population-wide Studies (N>1,000)
 - Longitudinal (neuroimaging) studies ...

M

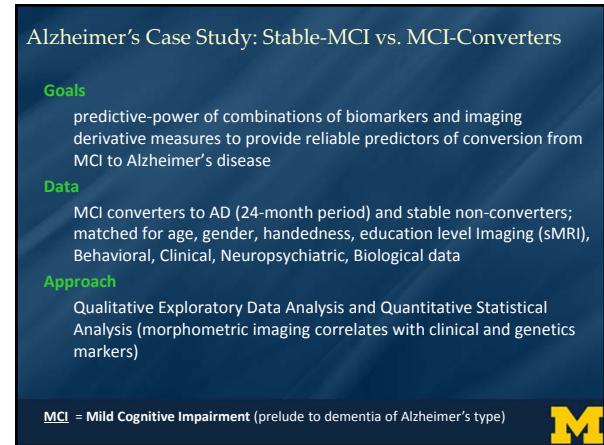
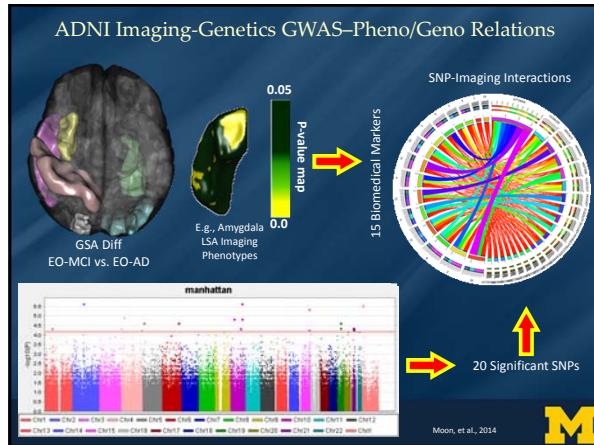
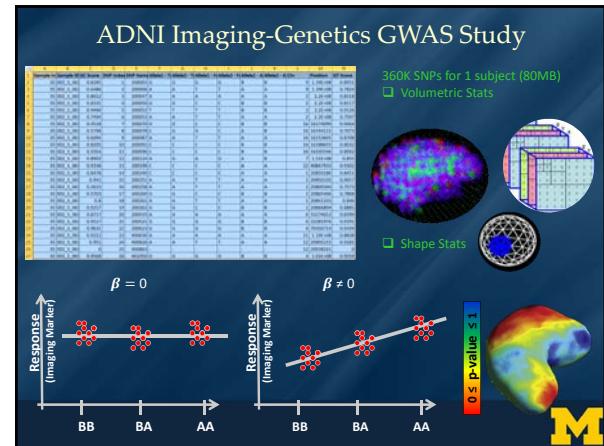
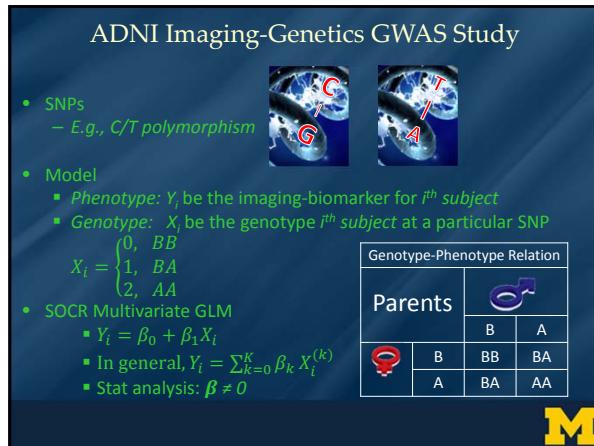
ADNI Imaging-Genetics GWAS Study

- Background
 - ADNI: *Alzheimer's Disease Neuroimaging Initiative*
 - Goal: Understand the early-onset (EO) cognitive impairment using neuroimaging and genetics biomarkers (55-65 y/o MCI and AD cohorts)
- Approach
 - Subjects
 - TBM
 - GWAS: 630-360K SNPs
 - SOCR Stats (MLR)
 - Pipeline workflows
- Results
 - Detected significant correlations between SNPs and various neuroimaging biomarkers in 36 EO subjects
 - Observed differences between EO-AD and EO-MCI

Cohort	Demographics	AD	MCI	P
Early Onset (EO)	N	9	27	-
	Age	60.4/3.34	61.2/2.87	0.0810
	Gender(m/f)	4/5	15/12	0.5630
	Education	16.142 ± 2.304	16.226 ± 2.764	0.8834
	MMSE	21.571±2.795	26.745 ± 2.342	0.0001
Handedness (R/L)	5/4	24/3	0.0286	
ApoE(+/-)	5/4	14/13	0.8471	

M

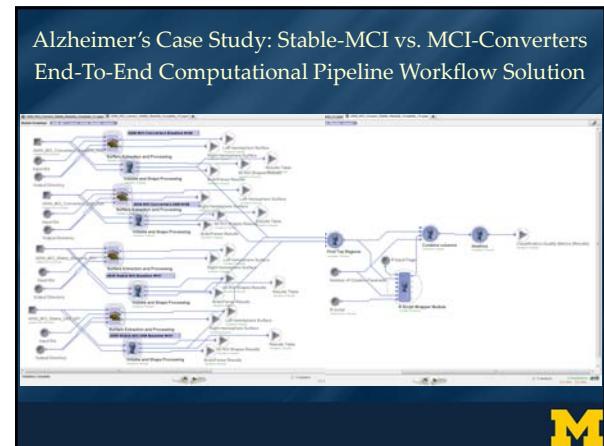
Dinov, et al., 2013



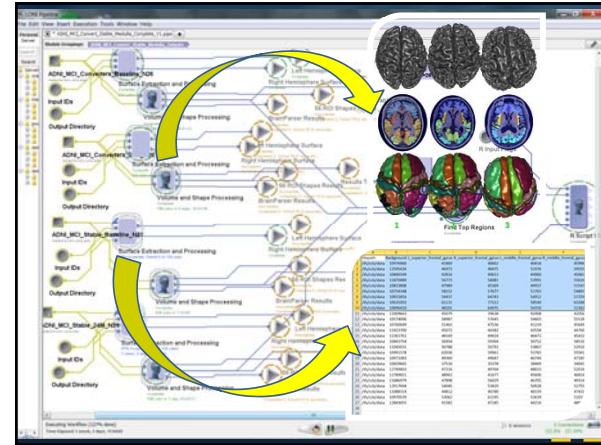
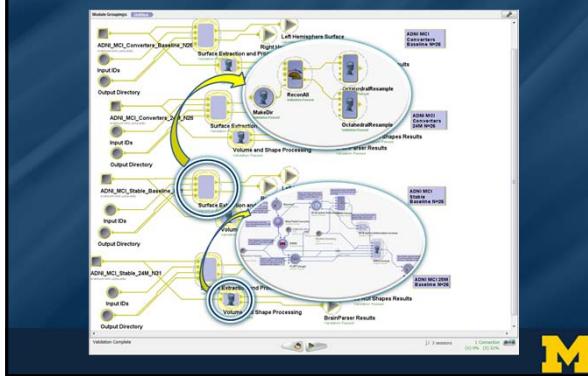
Prior Studies: MCI Conversion to Alzheimer's

Authors	Approach
Gaser, Franke et al. 2013	<u>BrainAGE</u> - age-related brain tissue loss
Conti, Vicini-Chilovi et al. 2013	<u>CA-SIT Smell Identification Test</u>
Mangialasche, Westman et al. 2013	Plasma levels of various natural <u>vitamin E</u> compounds
Dukart, Mueller et al. 2013	<u>Glucose Hypometabolism</u>
Zola, Manzanares et al. 2013	<u>Behavioral Assays</u> like visual paired comparison task
Albin, Giordani , et al, 2013	¹¹ C-PIB PET cerebral amyloids
Paulson & Igo, 2011	<u>ApoE4</u> status, imaging and CSF biomarkers
Apostolova, Dinov et al. 2006	<u>Structural Morphometry</u>

M



Alzheimer's Case Study: Stable-MCI vs. MCI-Converters



Alzheimer's Case Study: Stable-MCI vs. MCI-Converters

Subject	Demographics		Genetics		Clinical		Neuroimaging										
	Age	Kg	Sex	APOE A1	APOE A2	NPI SCORE	MMSE	GD TOTAL	CDR	FAO TOTAL	L Gyrus Rectus BL	R Fusiform Gyrus BL	L Superior Occipital Gyrus BL	L Caudate BL	R Caudate BL	L Putamen BL	R Putamen BL
1	65	59	F	3	4	0	23	1	0.5	7	1695	3976	8363	1296	1992	1749	2776
2	73	93	M	3	3	7	19	1	1	8	1333	6016	13290	835	2137	2290	4327
...
N	64	63	F	3	3	3	29	6	0.5	2	2237	6887	16109	1223	2222	2525	4110



Alzheimer's Case Study: Stable-MCI vs. MCI-Converters

Classification Results Using Baseline Data		True State (Dx at 24 month follow up)		
		Converter	Stable	Total
Hierarchical Clustering Prediction Ana (7 Regions)	Converter	TP=21	FP=12	33
	Stable	FN=5	TN=19	24
	Total	26	31	57

Metric	Value	
	Top 7 Regions	Top 20 Regions
Sensitivity	0.81	1.0
Specificity	0.61	0.87
Power to detect Converters	0.91	1.0
Accuracy	0.70	0.93



Acknowledgments

• Collaborators

Arthur Toga, Roger Woods, Jack Van Horn, Zhiowen Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Anand Joshi, Shantanu Joshi, Paul Thompson, Luminita Vese, Nicolas Christou, Stan Osher, Stefano Soatto, Seok Moon, Junning Li, Dennis Pearl, Kyle Siegrist, Nicolas Christou, Rob Gould, Young Sung, Fabio Macchiaridi, Federica Torri, Carl Kesselman



- Funding
 - NIH: BIRN 1U24-RR025736, LONI P41-EB015922
 - NSF: SOCR 0716055, Distributome 1023115

Live Pipeline Demo?

