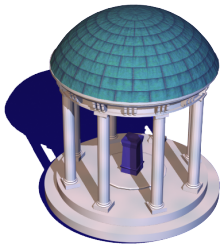
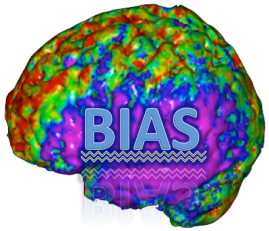
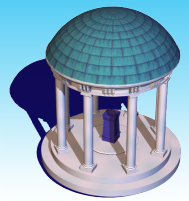
UNC SCHOOL OF PUBLIC HEALTH



Big Data Challenges in Neuroscience and Neuroimaging Studies

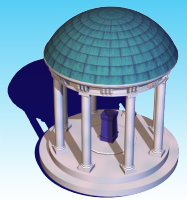
Hongtu Zhu, Ph.D
Department of Biostatistics† and Biomedical Research Imaging Center‡
The University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599, USA



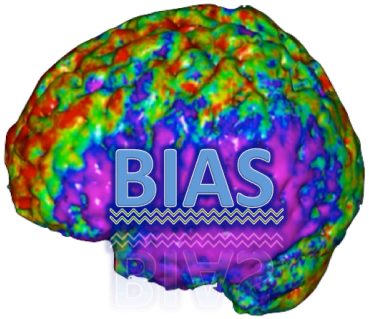


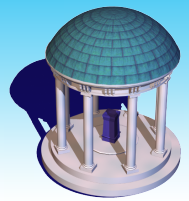
Outline

- Big Data Challenges (BDC)
- BDC in Neuroscience and Neuroimaging
- Big Data Integration



Big Data Challenges





Big Data

What? ***Wikipedia for Big data***

Big data refers data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

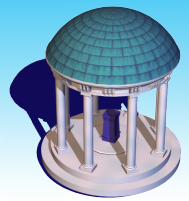
Big data is a set of techniques and techologies that require new forms of integration to uncover large hidden values from large daatsets that are diverse, complex, and of a massive scale

Size?

A few dozen terabytes to many petabytes of data.

Characteristics?

Volume, Variety, Velocity, Variability, Veracity, Complexity, ….



Big Data or Pig Data

Why?

Answer questions of commercial or scientific interest.

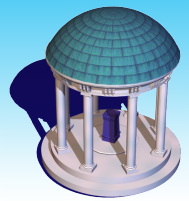
What matters?

Ensuring accurate and appropriate data collection.
Correct variables, Collection methods (techniques and sampling),

Quality assurance and Quality control

Does it work?

Big data does not work in many cases, since we do not know
(i) which variables (information at which scale) are critical;
(ii) whether we are able to collect such information.



Big Data Integration

Big data integration is to integrate multiple sources of data to improve knowledge discovery.

Data Sources Discovery: all related information

Data Exploration (e.g., meta analysis):

(i) the use of prior knowledge,- and its efficient storage;
(ii) the development of statistical methods to analyze heterogeneous data sets;
(iii) the creation of data explorative tools that incorporate both useful summary statistics and new visualization tools.

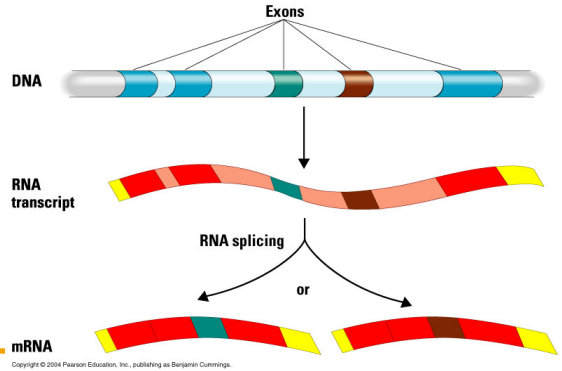
Human Genome Project

The HGP aims to determine the sequence of chemical base pairs which make up human DNA and identify and map all of the genes of the human genome.

1000 Genomes Project

Encyclopedia of DNA Elements Project (ENCODE)

The Cancer Genome Atlas Project (TGCA) is to generate insights into the heterogeneity of different cancer subtypes by creating a map of molecular alternations for every type of cancer at multiple levels.



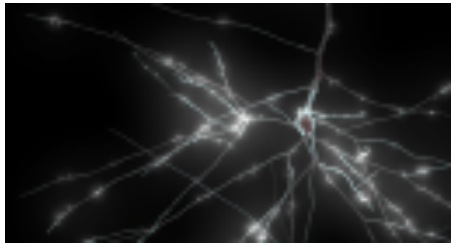
Immunological Genome Project (ImmGen)

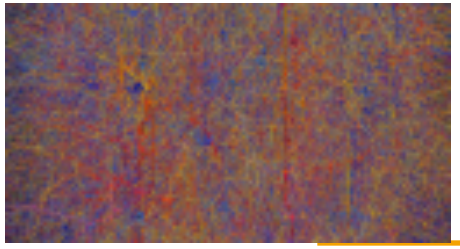
HBP and BRAIN

Human Brain Project

aims to simulate the complete human brain on Supercomputers to better understand how it functions.

The Brain Research through
Advancing Innovative Neurotechnologies or BRAIN,
aims to reconstruct the activity of every single neuron as they fire simultaneously in different brain circuits, or perhaps even whole brains.





Big Data to Knowledge (BD2K)

The four aims of BD2K are

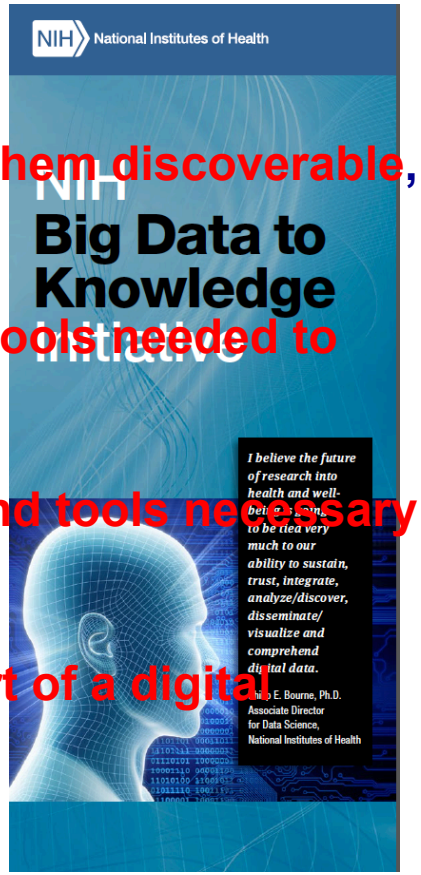
NIH Big Data to Knowledge (BD2K)

To facilitate broad use of biomedical digital assets by making them discoverable, accessible, and citable

To conduct research and develop the methods, software, and tools needed to analyze biomedical data.

To enhance training in the development and use of methods and tools necessary for biomedical Big Data science

To support a data ecosystem that accelerates discovery as part of a digital enterprise.



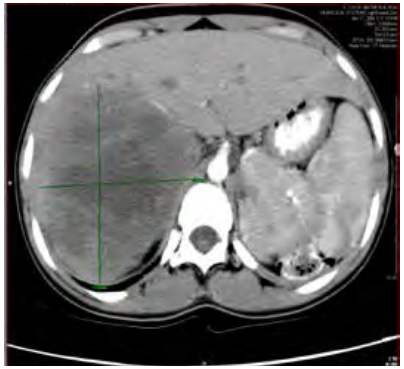
Precision Medicine

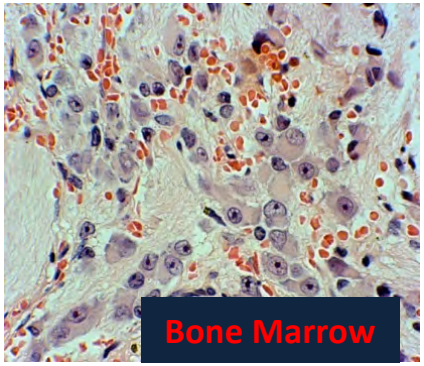
Precision medicine **(PM) is a medical model that proposes the customization of healthcare—with medical decisions, practices, and/or products being tailored to the individual patient.**

Precision Medicine refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment.

PM (wiki)

Cover Art: Nicolle Rager Fuller, Sayo-Art LLC
Photo: © Graham Bell/Corbis





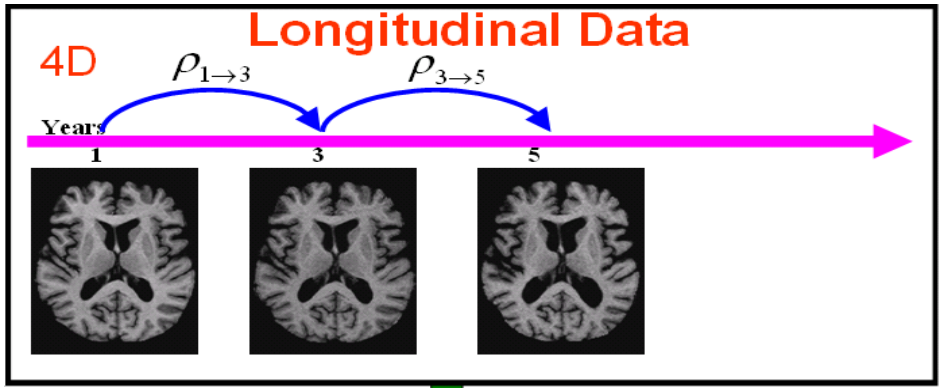
Dream Challenges

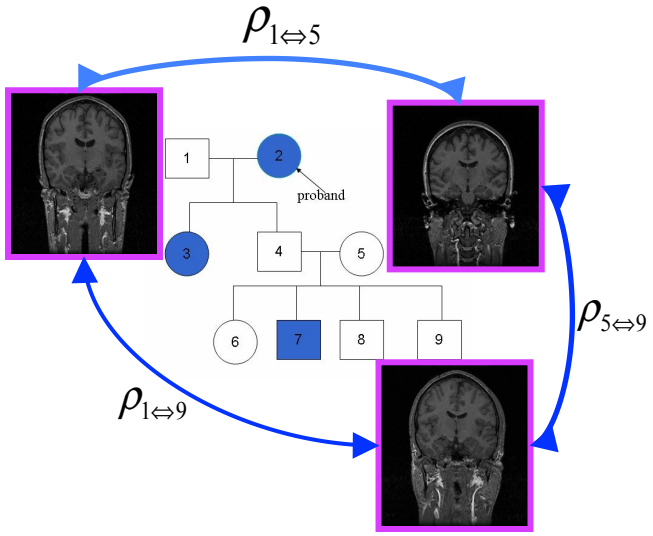
http://dreamchallenges.org

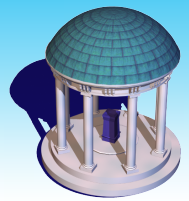
Study Design

Scientific Questions

Design: cross-sectional studies;
clustered studies including
longitudinal and twin/familial studies;







Imaging Data

Structural MRI

Diffusion MRI

- Variety of acquisitions
- Measurement basics
- Limitations & artefacts
- Analysis principles
- Acquisition tips

Functional MRI (task)

Functional MRI (resting)

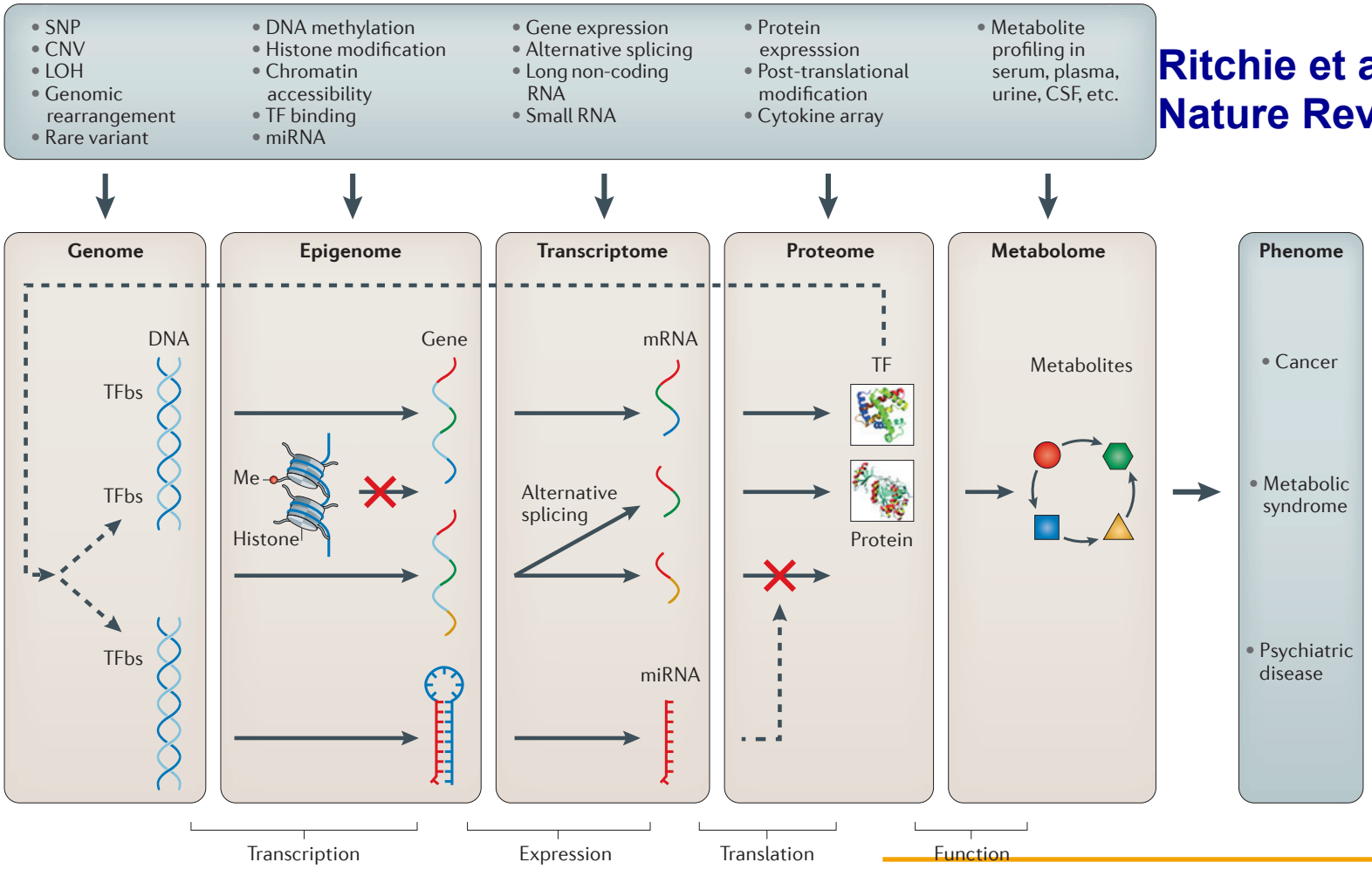
PET

EEG/MEG

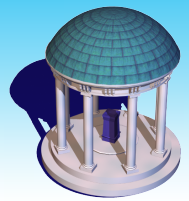
CT

Calcium

Multi-Omic Data



Ritchie et al. (2015). Nature Review Genetics



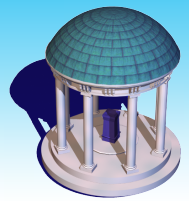
Clinical Data and Acquisition

Clinical Data: a variety of clinical sources to present a unified view of a single patient.

clinical laboratory test results, patient demographics, pharmacy information, hospital admission, discharge and transfer date, progress report, etc.

Clinical Acquisition:

- Paper or electronic medical records
- Paper forms completed at a site
- Interactive voice response systems
- Local electronic data capture systems
- Central web based systems

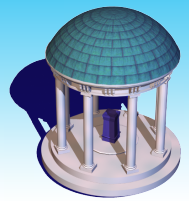


Data Exploration

Data Analysis

- **Single Level Data Analysis** for imaging or omics data, e.g., denoise, segmentation, cluster, network,
- **Multi-level Data Analysis** for across imaging or omics data
- **Prediction** by integrating imaging, clinical, and omics data.

Software/Computing Language/

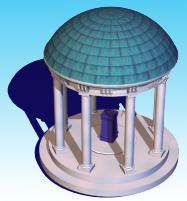


Apache Spark

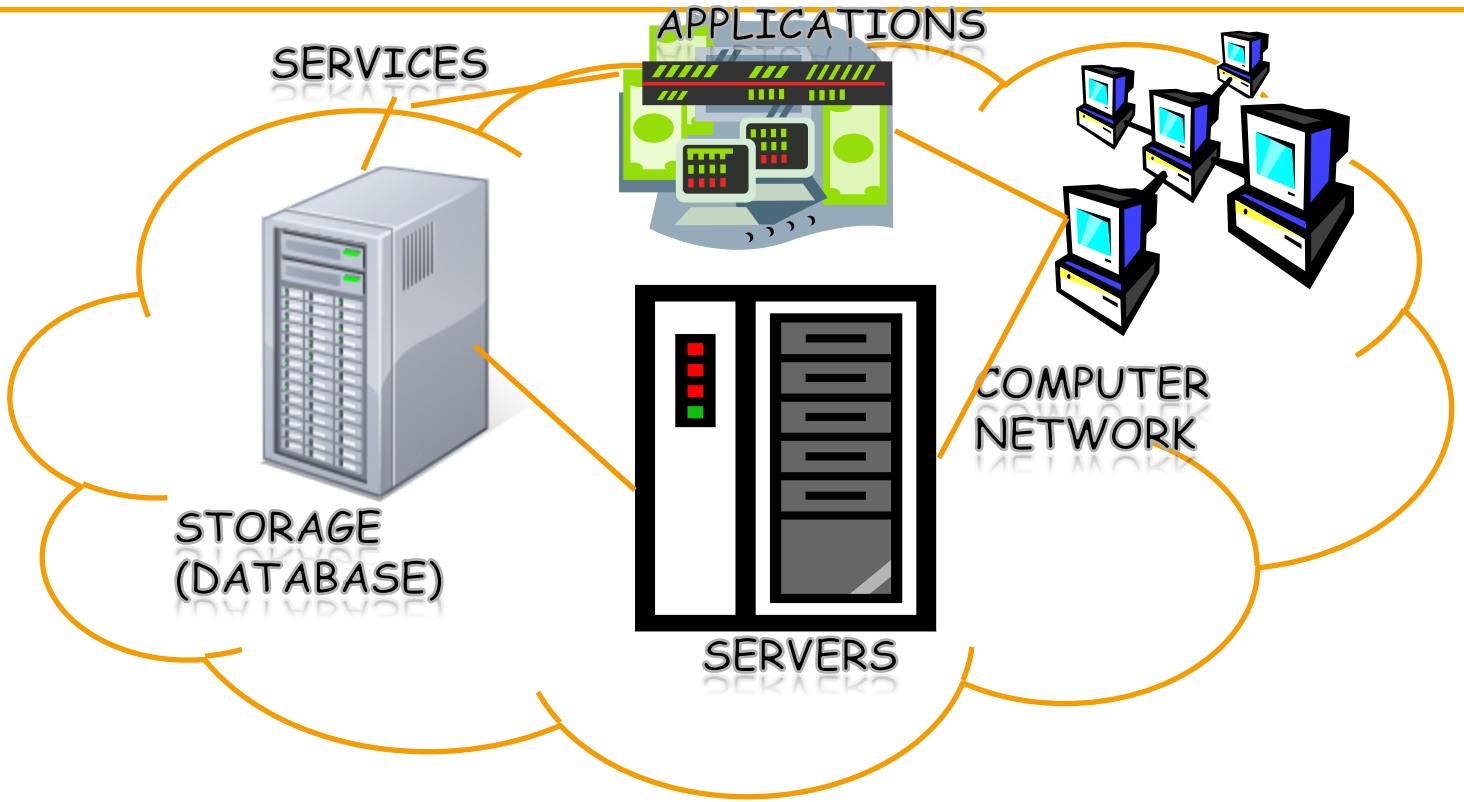
Data growing faster than processing speeds

Only solution is to parallelize on large clusters

» Wide use in both enterprises and web industry



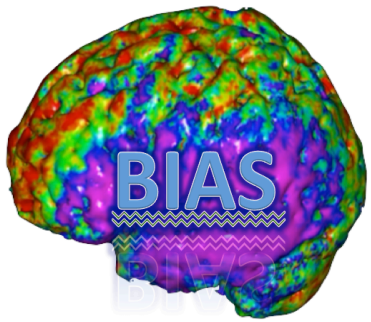
Cloud Computing



- **Shared pool of configurable computing resources**
- **On-demand network access**
- **Provisioned by the Service Provider**

Adopted from: Effectively and Securely Using the Cloud Computing Paradigm by peter Mell, Tim Grance

BDC in Neuroscience and Neuroimaging



Brain

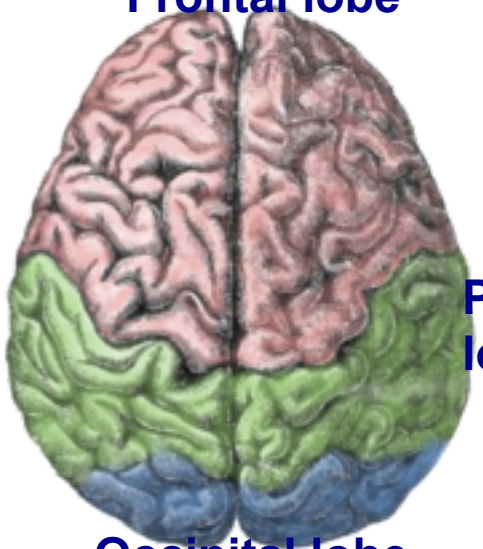
The brain is the main organ of the nervous system and is composed of neurons, glial cells and blood vessels.

Brain regions communicate with one another in complex spatiotemporal patterns, which enable

- **the formation of creative thoughts,**
- **the acquisition of new skills, and**
- **the adaptation of human behavior.**

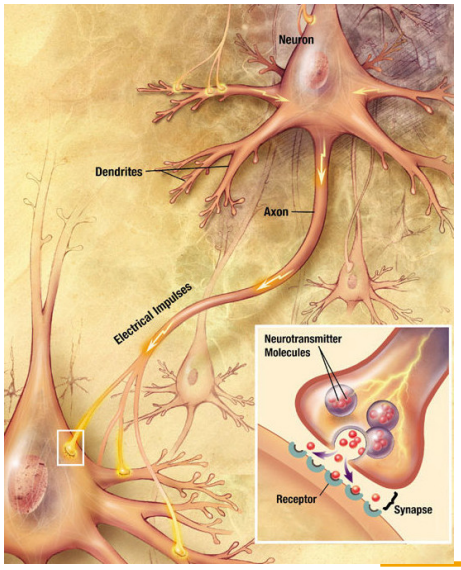
Wiki

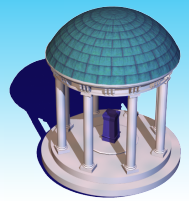
Frontal lobe



Parietal lobe

Occipital lobe



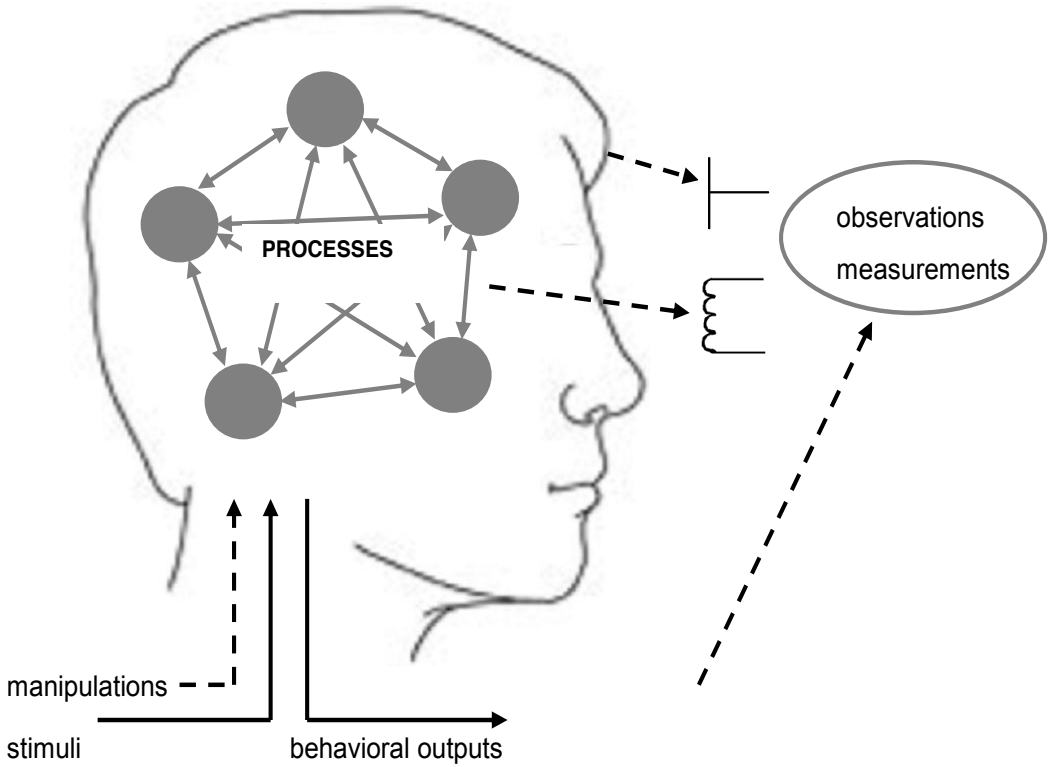


Fundamental Questions

- How do individual brain areas interact with one another to enable cognitive function?
- How is cognition constrained by white matter pathways?
- How does the brain transition between functions like memory, attention, and movement?
- How do we control the interactions between different neural circuits in our brains?
- How learned information is physically stored in the brain?
- How psychiatric diseases affect brain structure and function?
- How genetic and environmental interactions influence brain structure and its variability?
- How the brain changes over the course of development and aging may be usefully addressed?

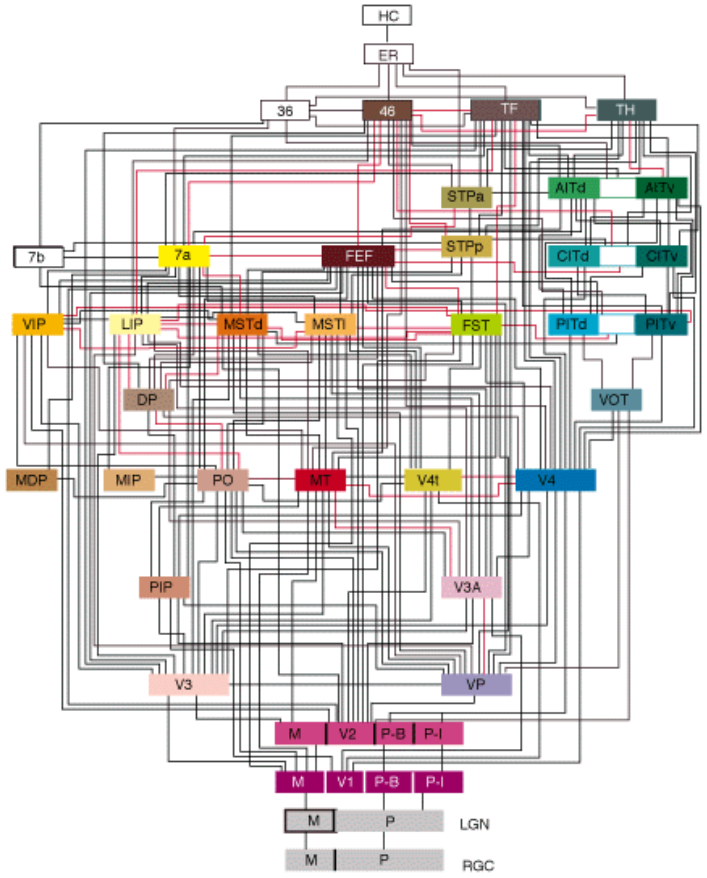
……..

A Multiscale Physical System



stimulus – activity – measurement chain

Robinson



The van Essen diagram

A Multi-modal Approach

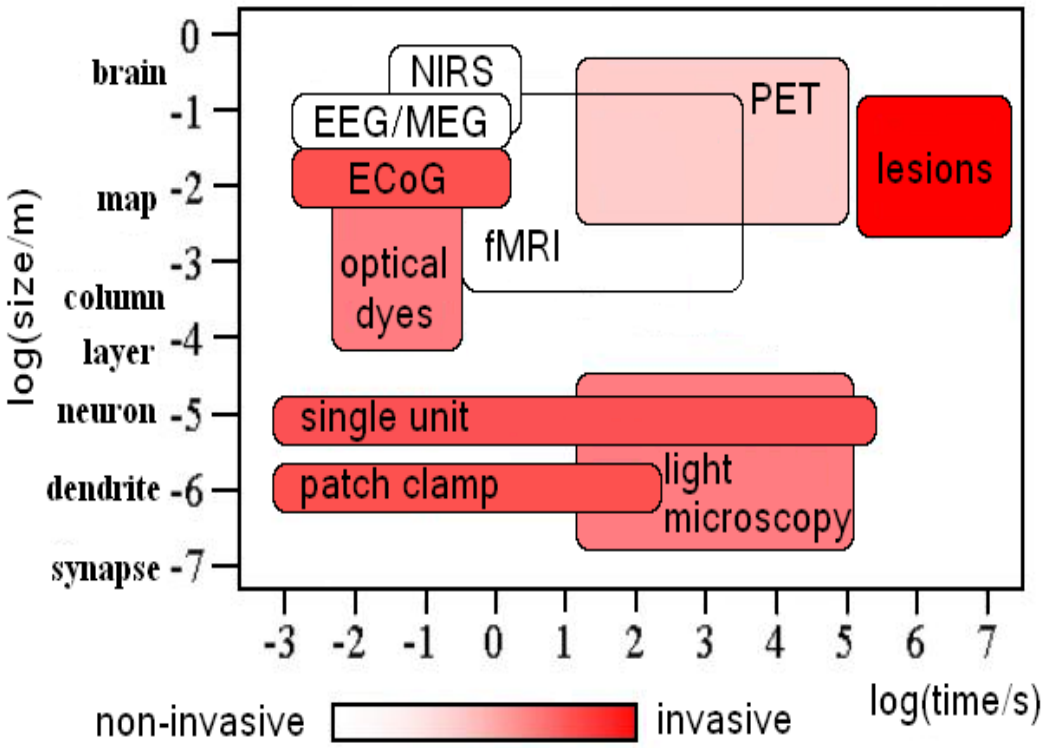
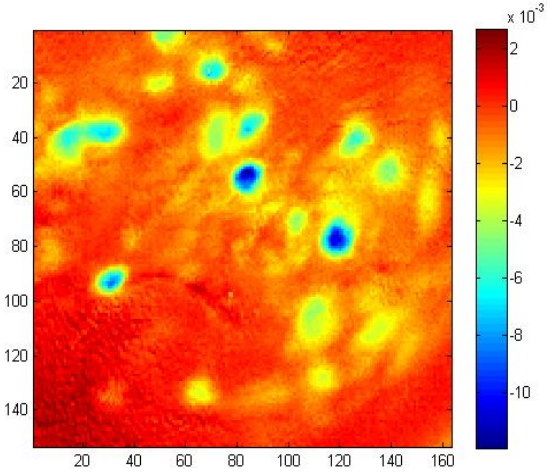
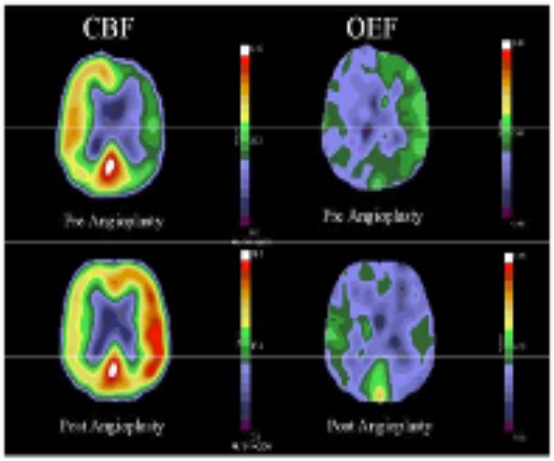


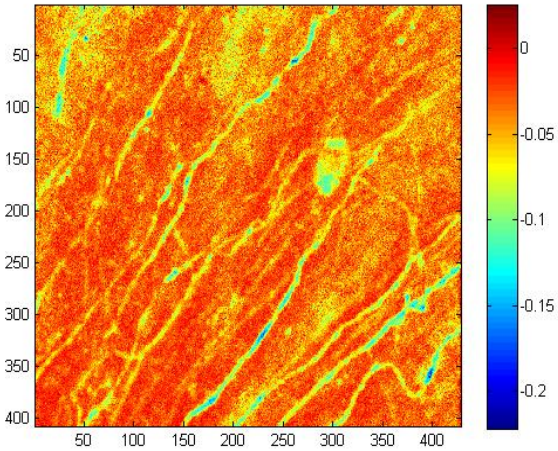
Image by A. Galka

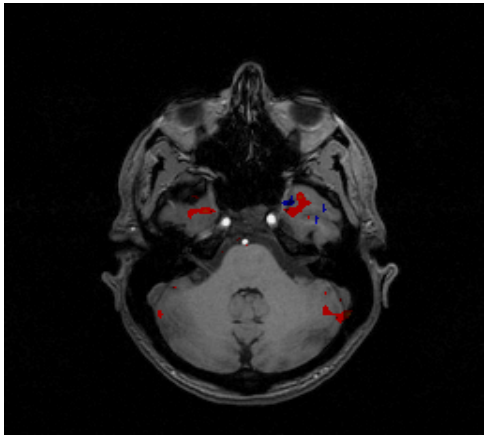
- Different models at different scales
- Ladder of overlapping models.
- Must be testable against multiple phenomena.

A Multi-modal Approach







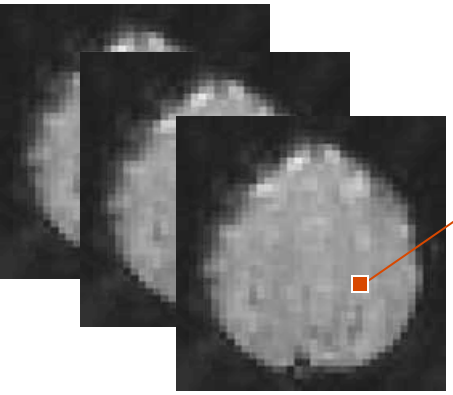


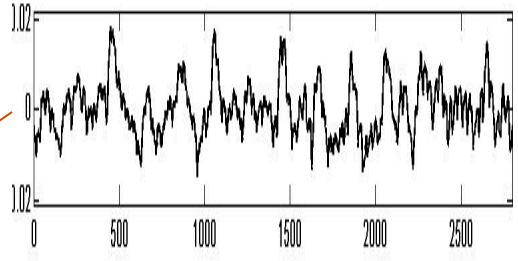
Single Level Analysis

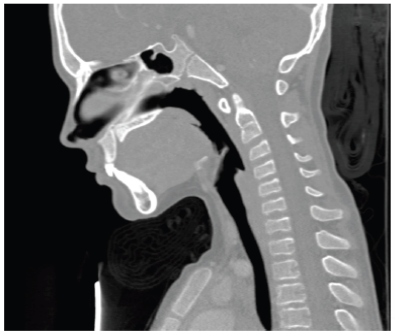
Imaging Construction

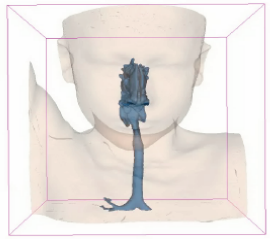
Image Segmentation

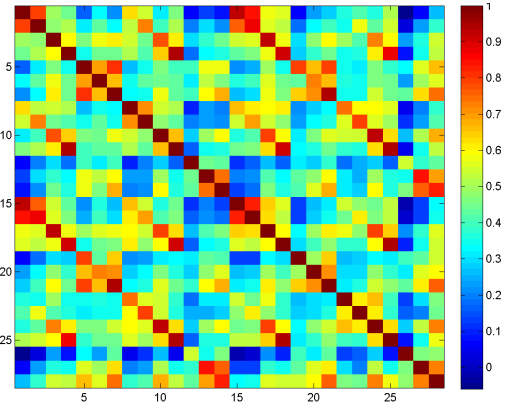
Example: Airway Segmentation from CT

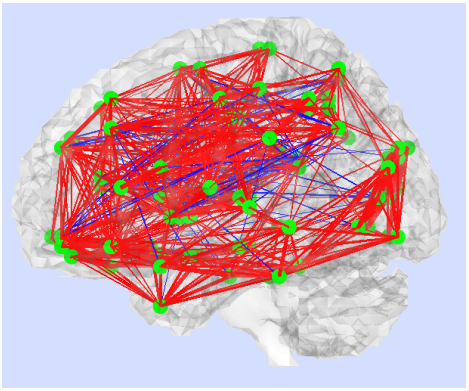


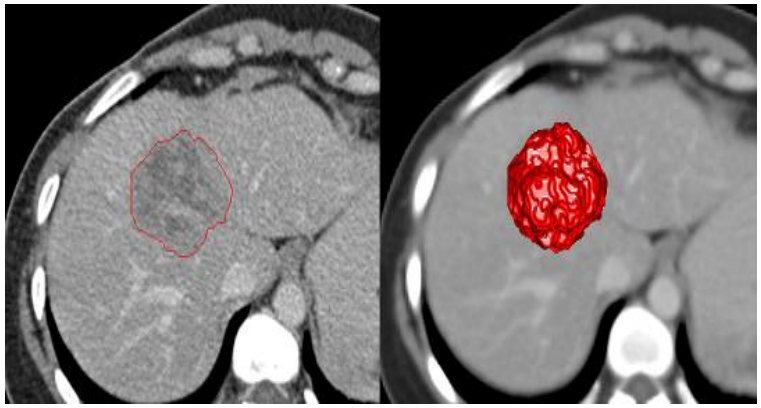








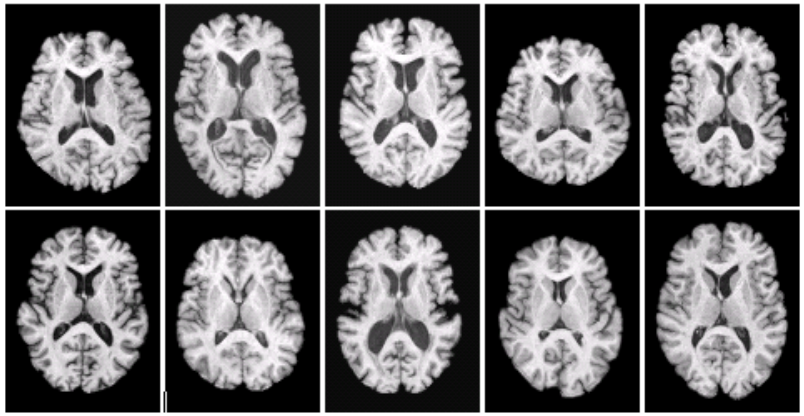


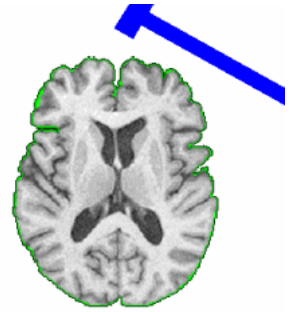


Marc

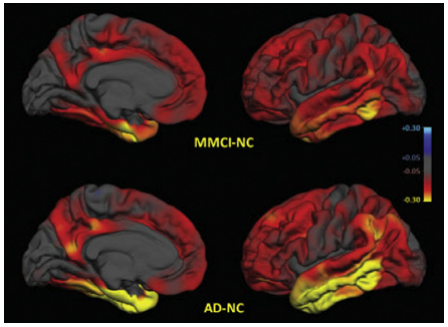
Single Level Analysis

Registration





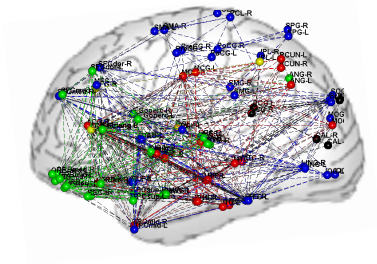
Group Differences



Longitudinal/Family Brain

Hibar, Dinggang, Martin

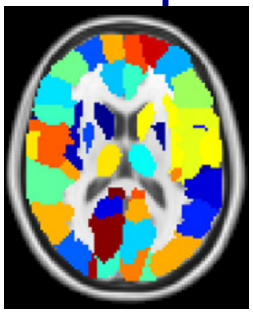
SLA: Brain Network Analysis



A parcellation of the whole brain is needed

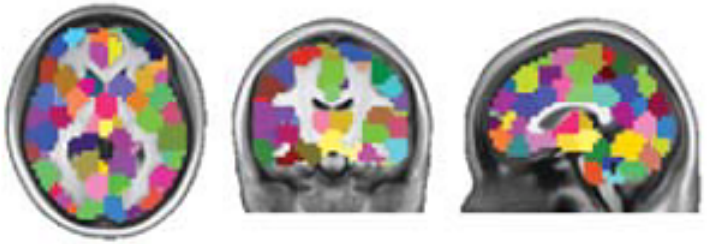
Volume based

AAL template



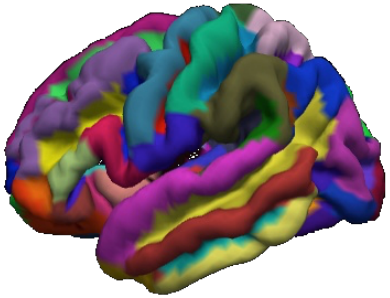
Salvador (2005)

fMRI-based parcellation
(spatial constrained spectral clustering)



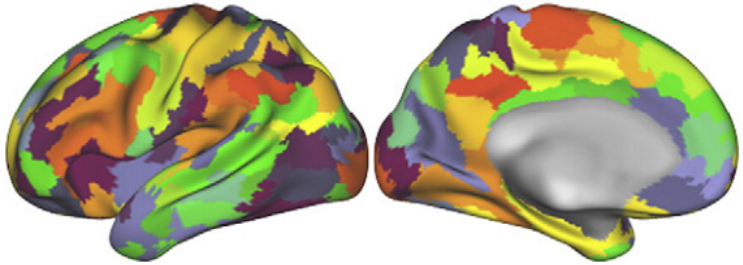
Surface based

Gyral-based parcellation



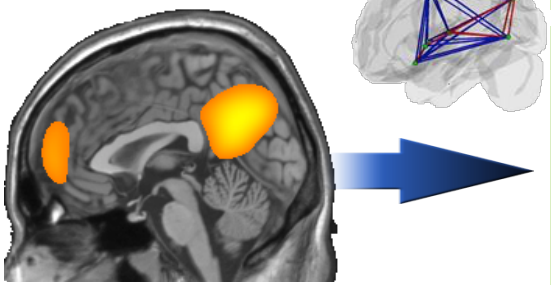
Desikan (2006)

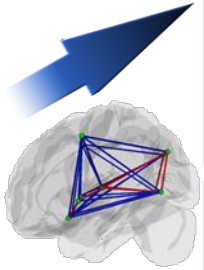
fMRI-based parcellation
(spatial constrained hierarchical clustering)

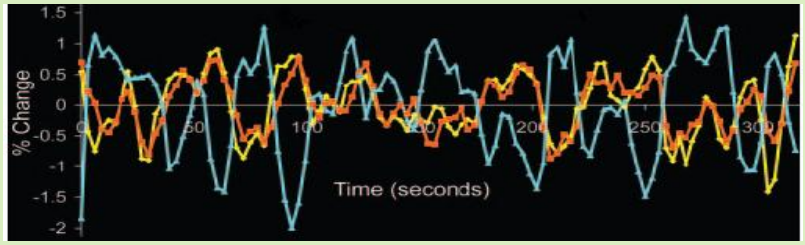


SLA: Brain Network Analysis

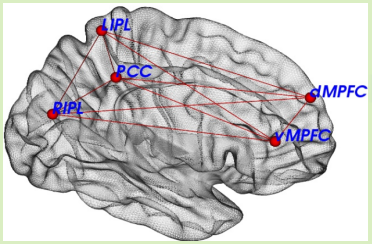
- **Brain connectivity analysis is a promising tool for investigating the human brain's structural and functional organization.**

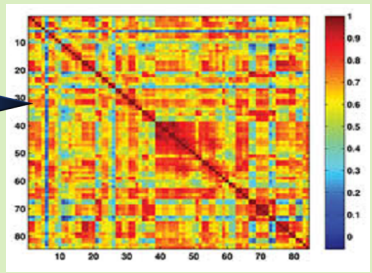






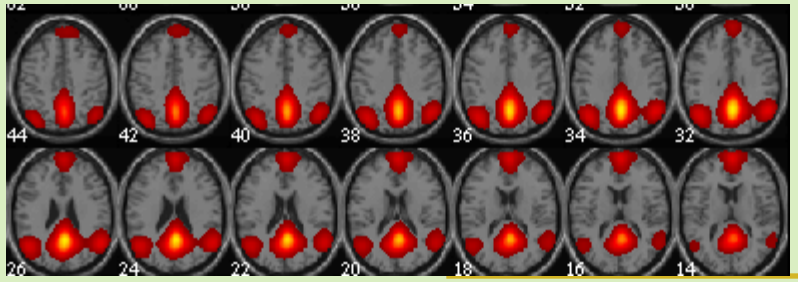
seed correlation analysis





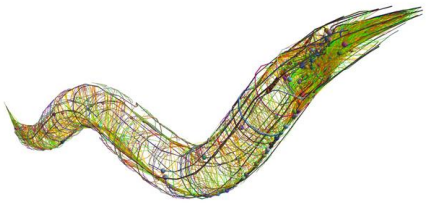
graph theoretic analysis

$X = A \cdot S$



independent component analysis (ICA)

Connectomics



Worm (*C. elegans*) with **302** neurons
Manual reconstruction took ~10 years

~**85,000,000,000** neurons
How many years?

PERSPECTIVE FOCUS ON BIG DATA

The big data challenges of connectomics

nature neuroscience

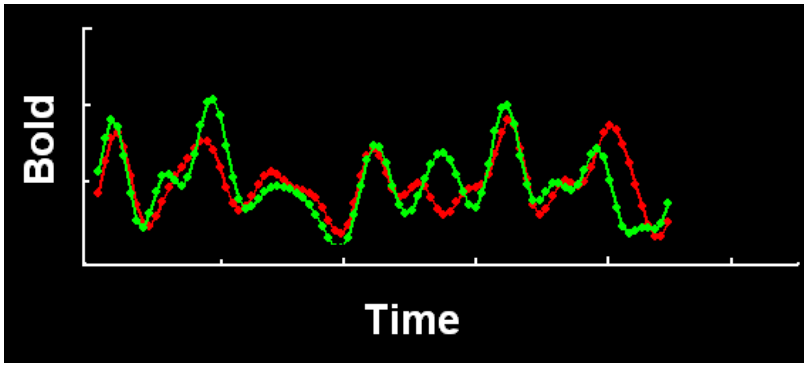
Jeff W Lichtman[1,2], Hanspeter Pfister[2,3] & Nir Shavit[4,5]

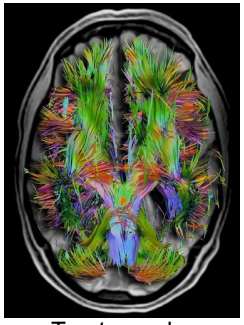
A complete human cortex will require a zetabyte (1,000 exabytes) of data, an amount of data approaching that of all the information recorded globally today.

http://www.scientificamerican.com/article/c-elegans-connectome/

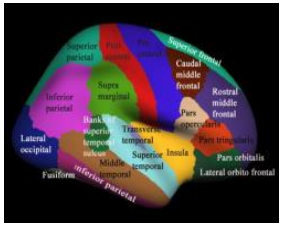
Multilevel Analysis

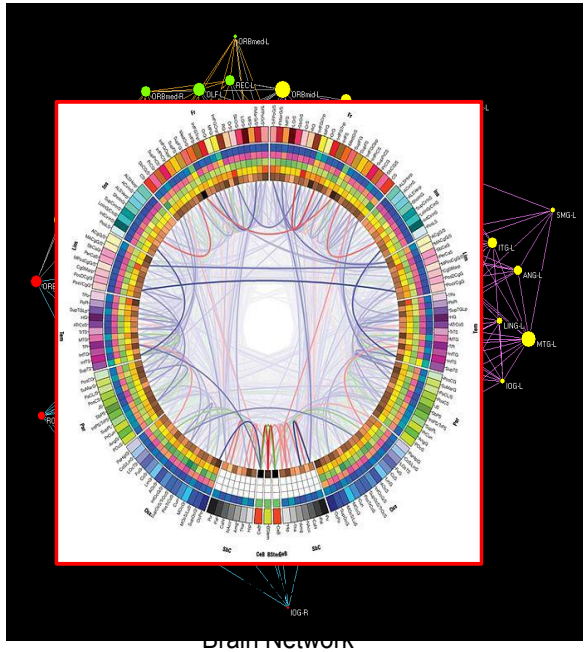
Anatomical Connectivity: a pattern of anatomical links. DTI
Functional Connectivity: statistical dependencies. rfMRI, fMRI, EEG,MEG,Cas
Effective Connectivity: causal interactions. fMRI, EEG, MEG, Cas

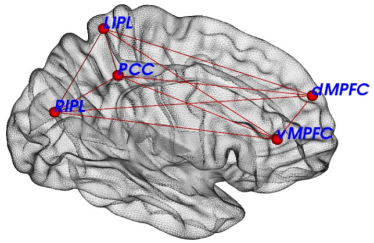


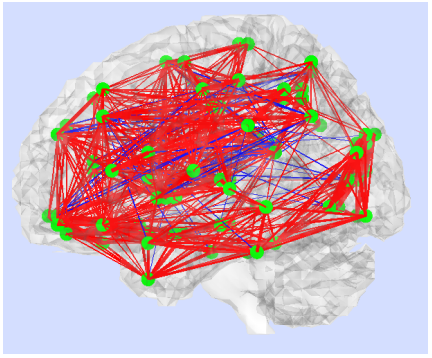

Tractography

+


Regions of Interest (ROIs)

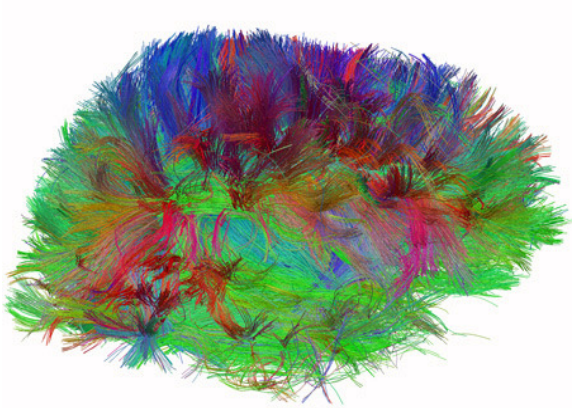

Brain Network



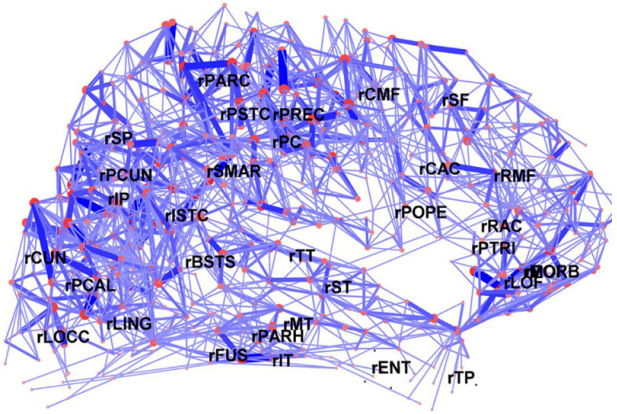


MLA:Functional and Structural Connectivity

Fact: Functional connectivity depends on structural connectivity.



Structural connectivity
(Diffusion MRI)



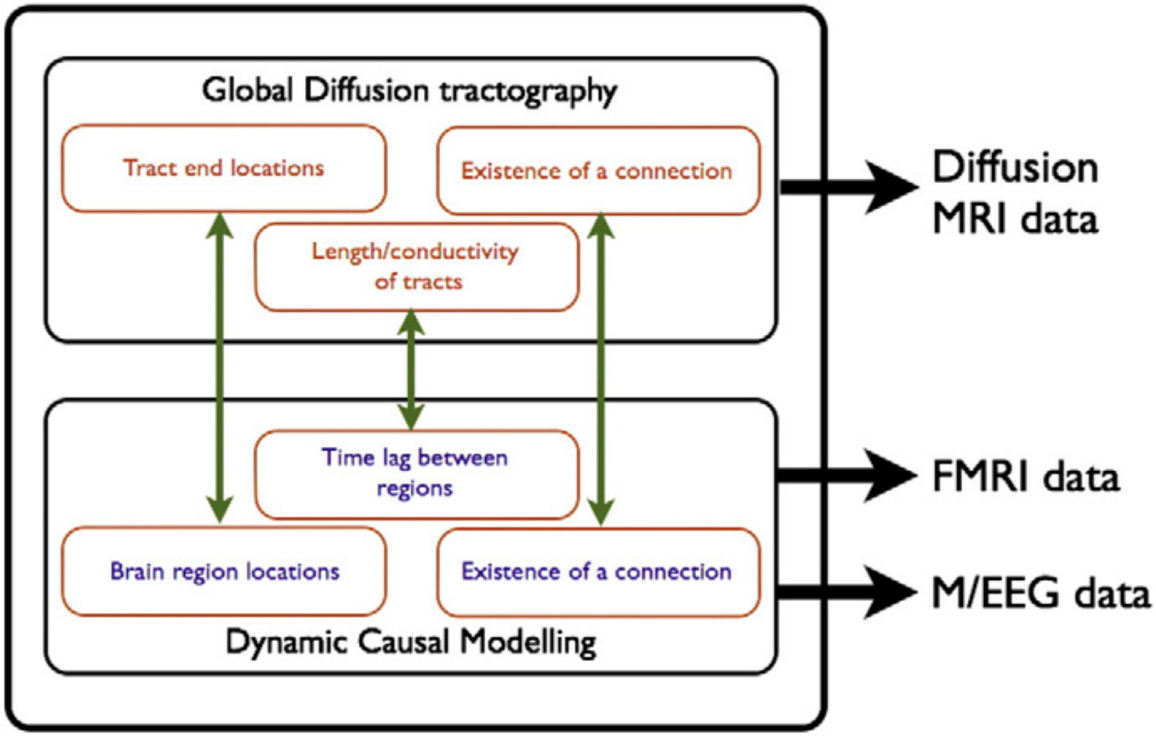
Functional connectivity
(Functional MRI)

- Diffusion MRI data has blind spots.
- Functional connectomics can help inform the anatomical connectome when structural information is missing or inaccurate.

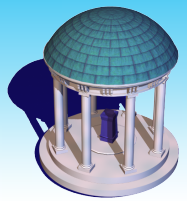
biophysical network can embody both the structural and functional architecture, and allow information from the different modalities to be fused in a mathematically principled way.

MLA:A Combined Biophysical Network Model

Schematic of a combined biophysical model



- predicts both anatomical and functional imaging data;
- can be regarded as separate generative models for anatomical and functional modalities, linked probabilistically by common parameters (green arrows).

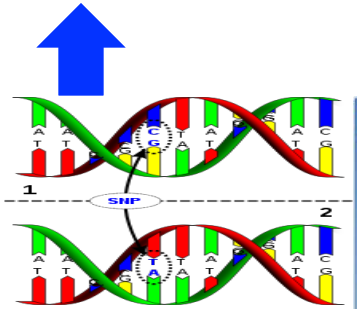


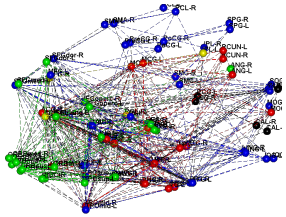
Prediction

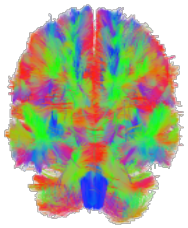
Data $\{(y_i, X_i) : i = 1, \cdots, n\}$ $X_i = \{X_i(d) : d \in D\}$

$$y_i = f(X_i) + \varepsilon_i$$

Disease Status, Survival Time, Treatment, Trajectories





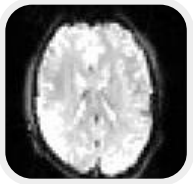


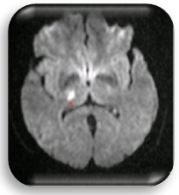
Interesting scientific questions include

- **Determine disease status**
- **Identify earlier biomarker**
- **Predict disease trajectories**
- **Predict survival time (e.g., time-to-event)**

Neuroimage analysis and its application
to computer aided diagnosis and surgery planning


sMRI


fMRI


DTI

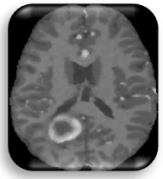
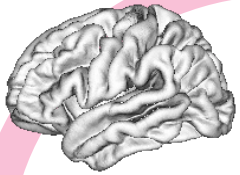
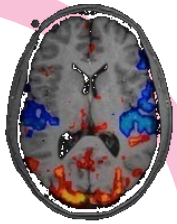

Tumor image

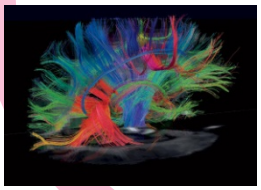
Image Analysis

Morphological analysis

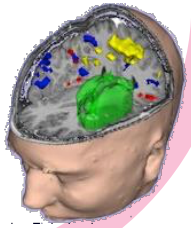


Functional analysis





Connectivity analysis



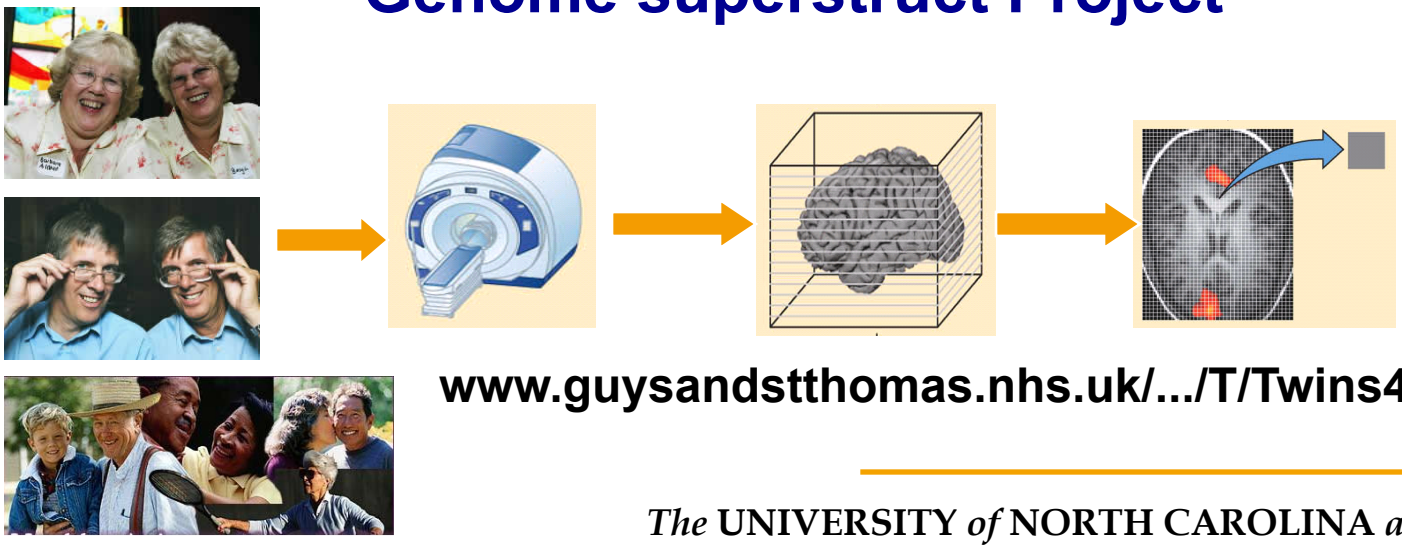
Tumor segmentation & visualization

Diagnosis Surgery Planning

Data → Knowledge

Big Neuroimaging Data

NIH normal brain development
1000 Functional Connectome Project
Alzheimer's Disease Neuroimaging Initiative
National Database for Autism Research (NDAR)
Human Connectome Project
Philadelphia Neurodevelopmental Cohort
Genome superstruct Project



www.guysandstthomas.nhs.uk/.../T/Twins400.jpg

The Human Connectome

- **Brain connectivity analysis is a promising tool for investigating the human brain's structural and functional organization.**

The Heavily Connected Brain
Peter Stern, "Connection, connection, connection…", Science, Nov. 1 2013: Vol. 342 no. 6158 P.577

- **The NIH Human Connectome Project**
  - **The Harvard/MGH-UCLA project**
  - **The WU-Minn Project**
- **The EU's 7th Framework Programme for Research**
  - **Consortium Of Neuroimagers for the Non-Invasive Exploration of Brain Connectivity and Tracts**

The BRAIN Initiative
(Brain Research through Advancing Innovative Neurotechnologies)

The Human Connectome Project

- **The HCP is to elucidate the neural pathways that underlie brain function and behavior.**

The Heavily Connected Brain
Peter Stern, "Connection, connection, connection…", Science, Nov. 1 2013: Vol. 342 no. 6158 P.577

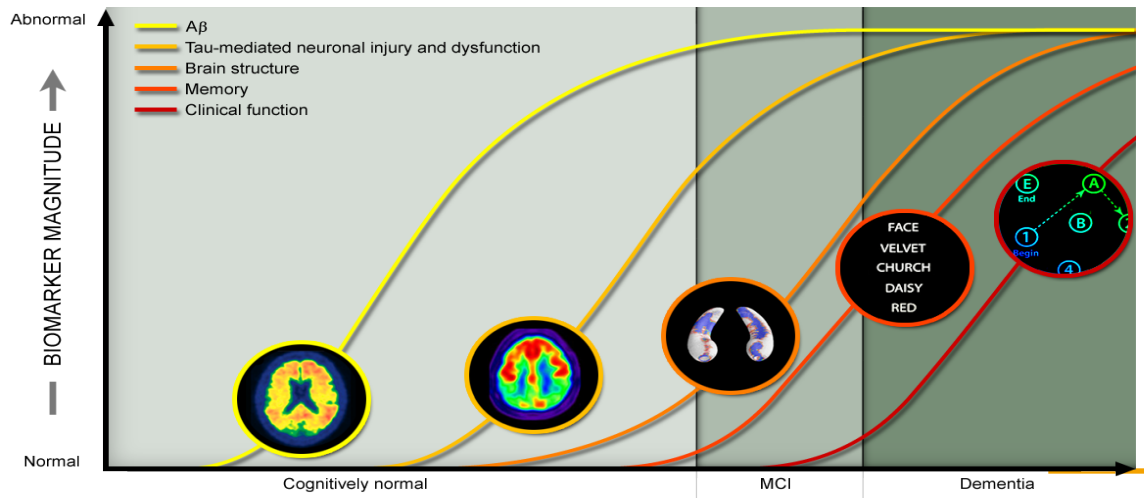
- **Resting-state fMRI (rfMRI) and dMRI provide information about brain connectivity.**
- **Task-evoked fMRI reveals much about brain function.**
- **Structural MRI captures the shape of the highly convoluted cerebral cortex.**
- **Behavioral data relate brain circuits to individual differences in cognition, perception, and personality.**
- **Magnetoencephalography (MEG) combined with electroencephalography (EEG) yield information about brain function on a milisecond time scale.**

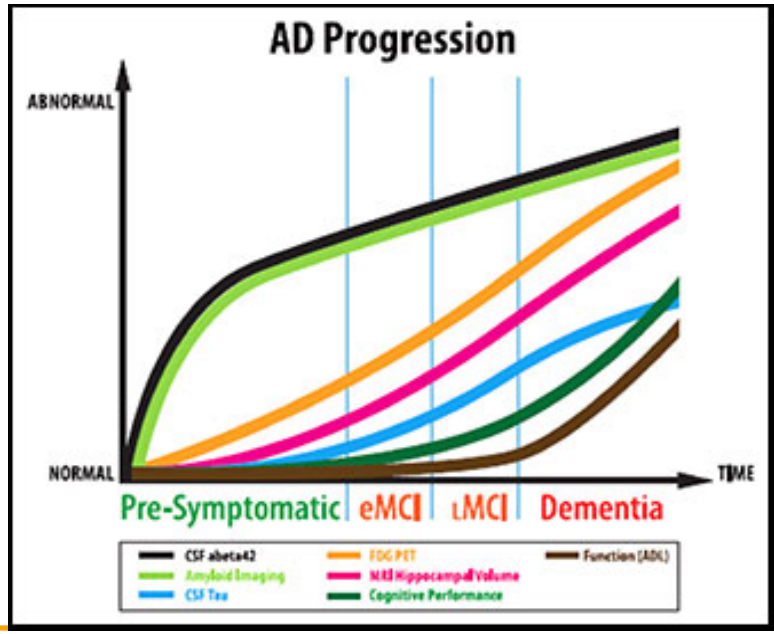
Alzheimer's Disease Neuroimaging Initiative

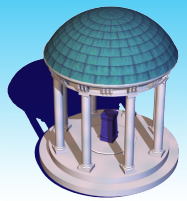
PI: Dr. Michael W. Weiner

- detecting AD at the earliest stage and marking its progress through biomarkers;
- developing new diagnostic methods for AD intervention, prevention, and treatment.

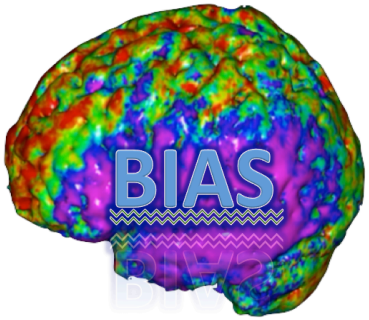
- A longitudinal prospective study with 1700 aged between 55 to 90 years
- Clinical Data including Clinical and Cognitive Assessments
- Genetic Data including Ilumina SNP genotyping and WGS
- MRI (fMRI, DTI, T1, T2)
- PET (PIB, Florbetapir PET and FDG-PET)
- Chemical Biomarker





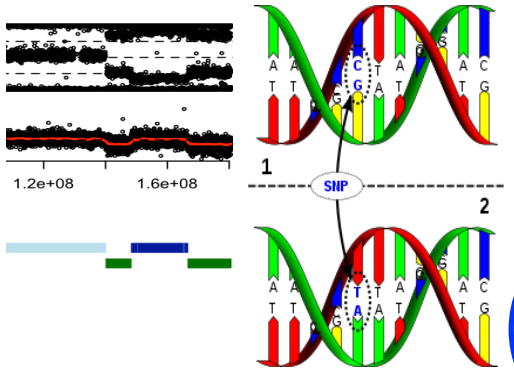


Big Data Integration



Big Data Integration

E: environmental factors



G: genetic markers

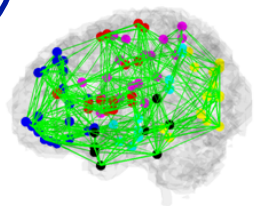
E

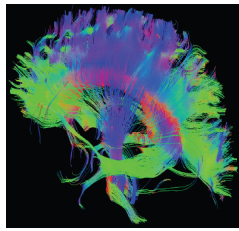
B

G

D

Selection

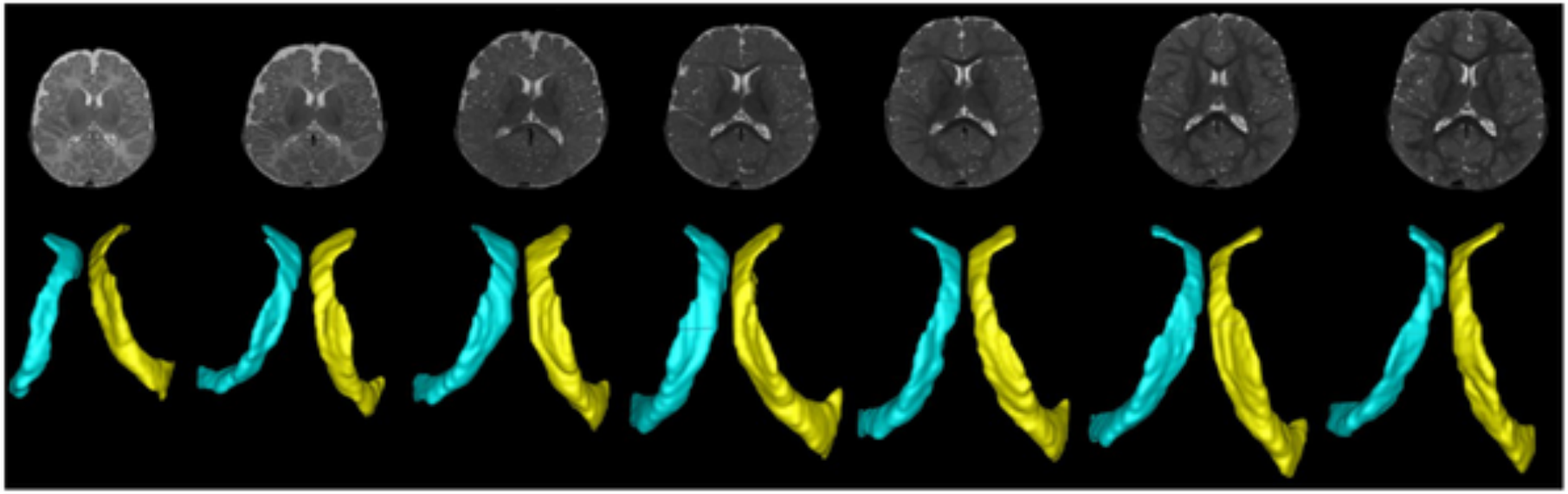




D: disease

http://en.wikipedia.org/wiki/DNA_sequence

Longitudinal Analysis of Lateral Ventricles



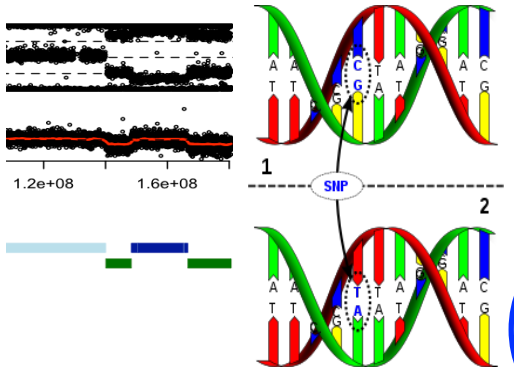
Representative T2-weighted images (upper row) from a subject imaged over the course of the first two years of life along with the segmented left and right ventricles (lower row) are shown.

Objectives: Chart changes in brain structure

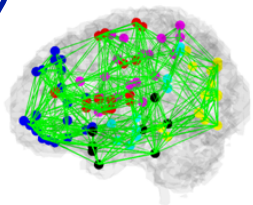
Bompard L, Xu S, Styner M, Paniagua B, et al. (2014) Multivariate Longitudinal Shape Analysis of Human Lateral Ventricles during the First Twenty-Four Months of Life. PLoS ONE 9(9):

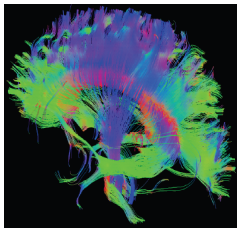
Big Data Integration

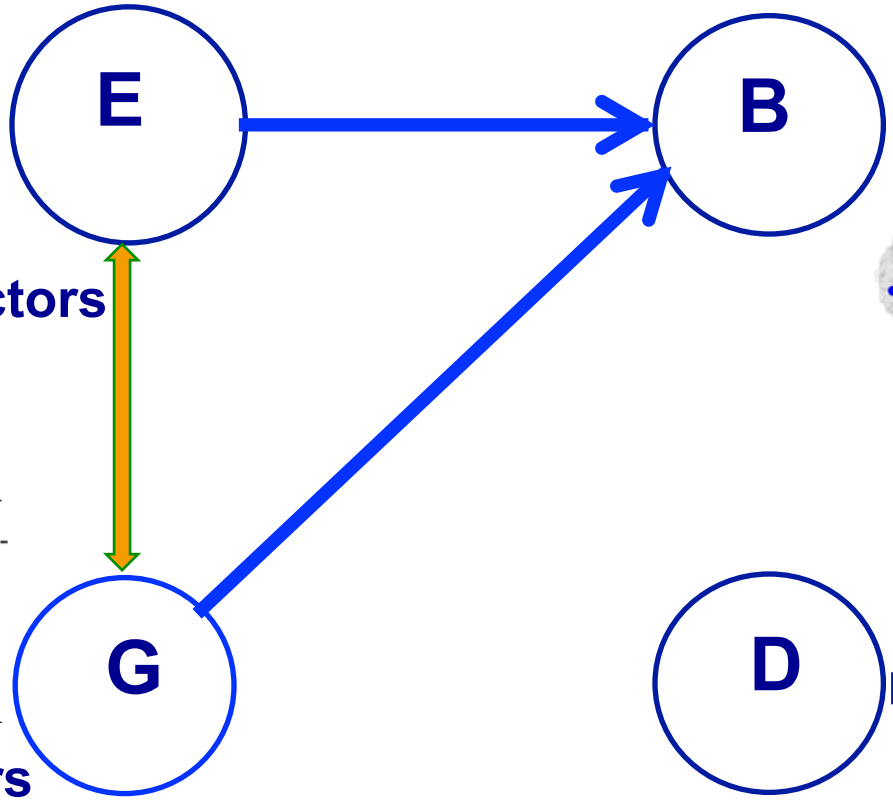
E: environmental factors

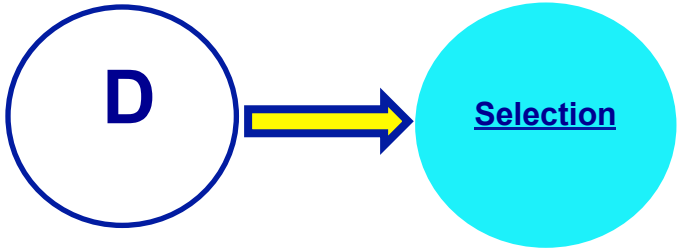


G: genetic markers





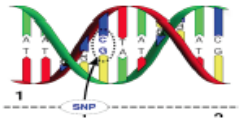
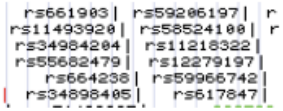




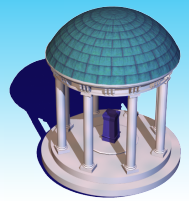
D: disease

http://en.wikipedia.org/wiki/DNA_sequence

Statistical Methods

Genetics \ Imaging	Candidate ROI	Many ROI	Voxelwise
Candidate SNP 	Imager	Imager	Imager
Candidate Gene 	Geneticist	↑	↑
Genome-wide SNP 	Geneticist	↑	↑
Genome-wide Gene 	Geneticist		

Hibar, et al. HBM 2012



High Dimensional Regression Model

Data $\{(Y_i, X_i) : i = 1, \cdots, n\}$

$$Y_i = \{y_i(v) : v \in V_0\} \qquad X_i = \{X_i(g) : g \in G_0\}$$

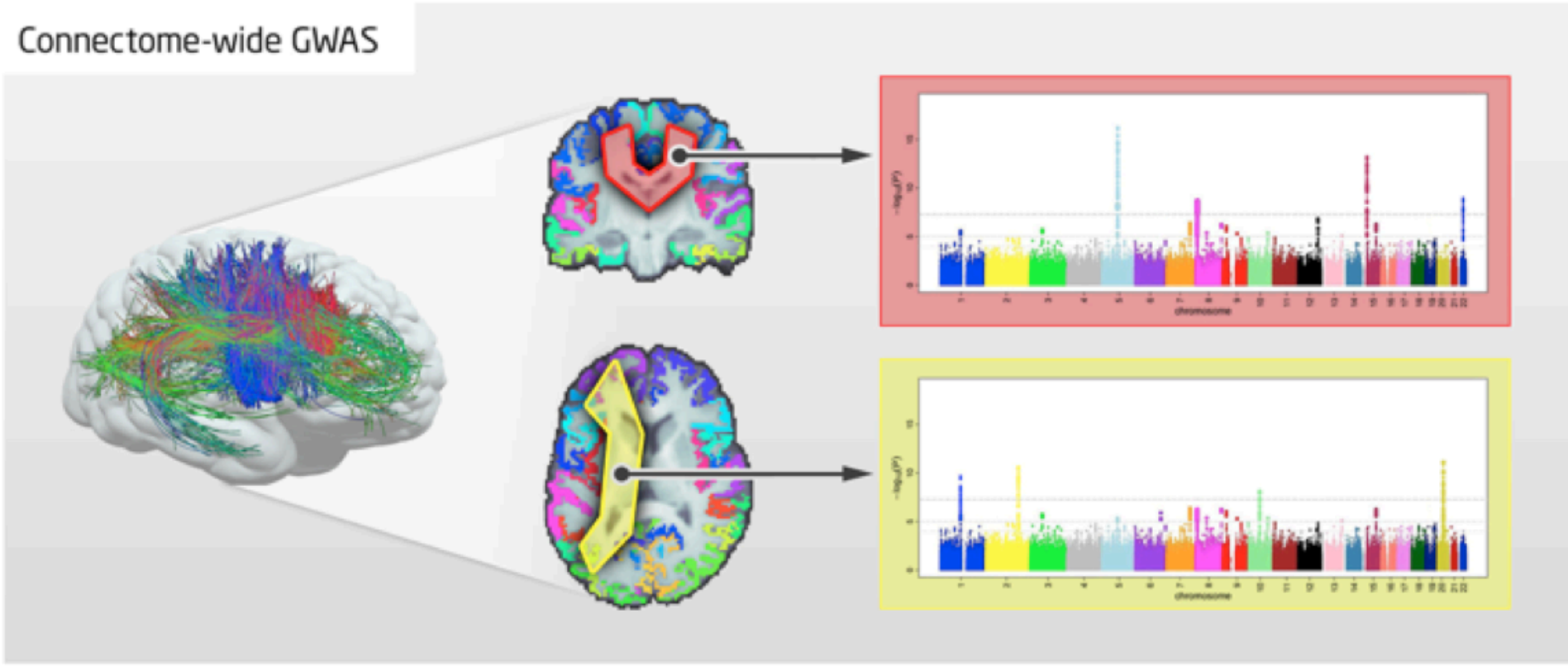
Phenotype		Genotype			Error
Y		X		B	E
$n \times p_y$	$=$	$n \times p_x$		$p_x \times p_y$	$+$ $n \times p_y$

Key Conditions:

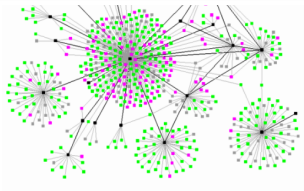
$$\max(p_x, p_y) \sim n$$

- **Sparsity of B**
- **Restricted null-space property for design matrix X**

Connectome-Wide Genome-Wide Screen
Alzheimer risk gene



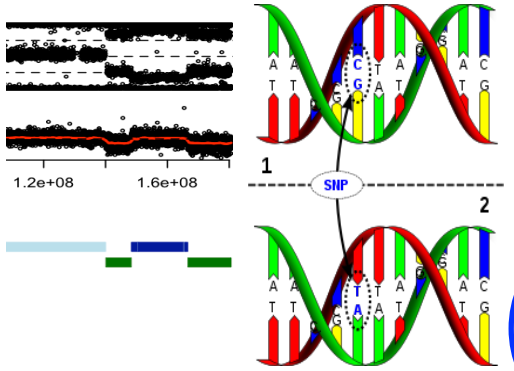
Discovery sample – Young Adults
Effect in ADNI
Within 2 weeks Sherva et al. published *SPON1*
Found in a cognitive GWAS in AD



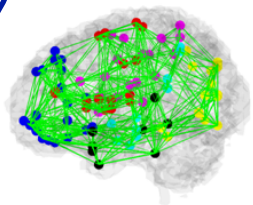
Jahanshad et al., PNAS 2013

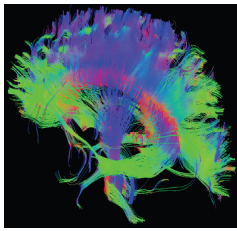
Big Data Integration

E: environmental factors



G: genetic markers



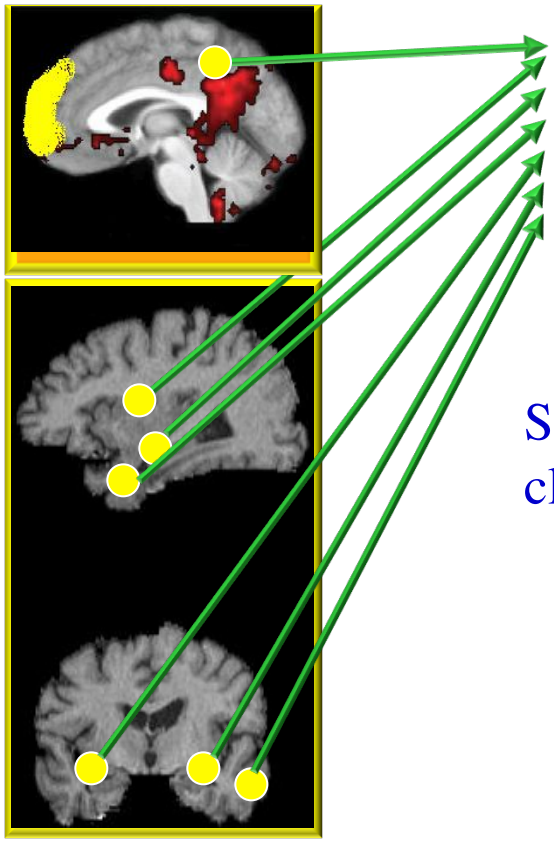


D: disease

http://en.wikipedia.org/wiki/DNA_sequence

Pattern classification of neuroimages

Functional information

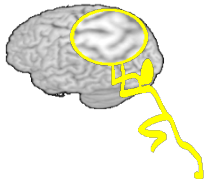


Pattern Classification

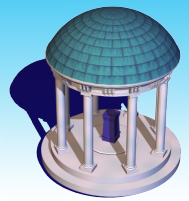
Quantitative Diagnosis

Structural, functional, and multimodality image classification

- Diagnosis of Schizophrenia
- Diagnosis of Alzheimer's disease (AD)
- Clinical outcomes



Morphological information



Alzheimer's Disease DREAM Challenge 1

Its goal is to apply an open science approach to rapidly identify **accurate predictive AD biomarkers** that can be used by the scientific, industrial and regulatory communities to improve AD diagnosis and treatment.

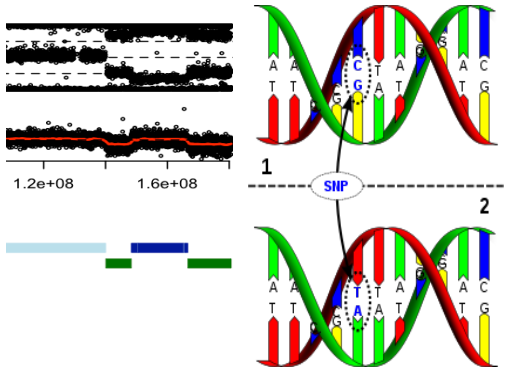
Sub 1: Predict the change in cognitive scores 24 months after initial assessment.

Sub 2: Predict the set of cognitively normal individuals whose biomarkers are suggestive of amyloid perturbation.

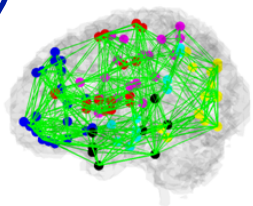
Sub 3: Classify individuals into diagnostic groups using MR imaging.

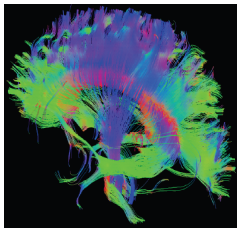
Big Data Integration in Health Informatics

E: environmental factors



G: genetic/genomics

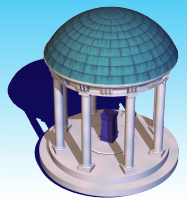




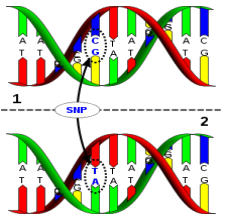
I: imaging/device

D: disease

http://en.wikipedia.org/wiki/DNA_sequence



Big Data Integration



Medical Informatics & Management

→ **Disease** → **Medical Industry**

Etiology
Prevention
Treatment

Care
Policy
System
Science
Insurance
Economics
Pharmaceutical

ASA: Statistics in Imaging Section

SAMSI

- 2013 Neuroimaging Data Analysis
- 2015-2016 Challenges in Computational Neuroscience

Thank You!!