Opportunities, Challenges, and Strategies for Statistics in Large-Scale Medical Studies

Hongtu Zhu, Professor

Department of Biostatistics
The University of Texas MD Anderson Caner Center

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

BIG-S²

Biostatistics and Imaging Genomics Analysis Lab

BIG-S2=Statistics and Signal

http://odin.mdacc.tmc.edu/bigs2/
https://github.com/BIG-S2

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

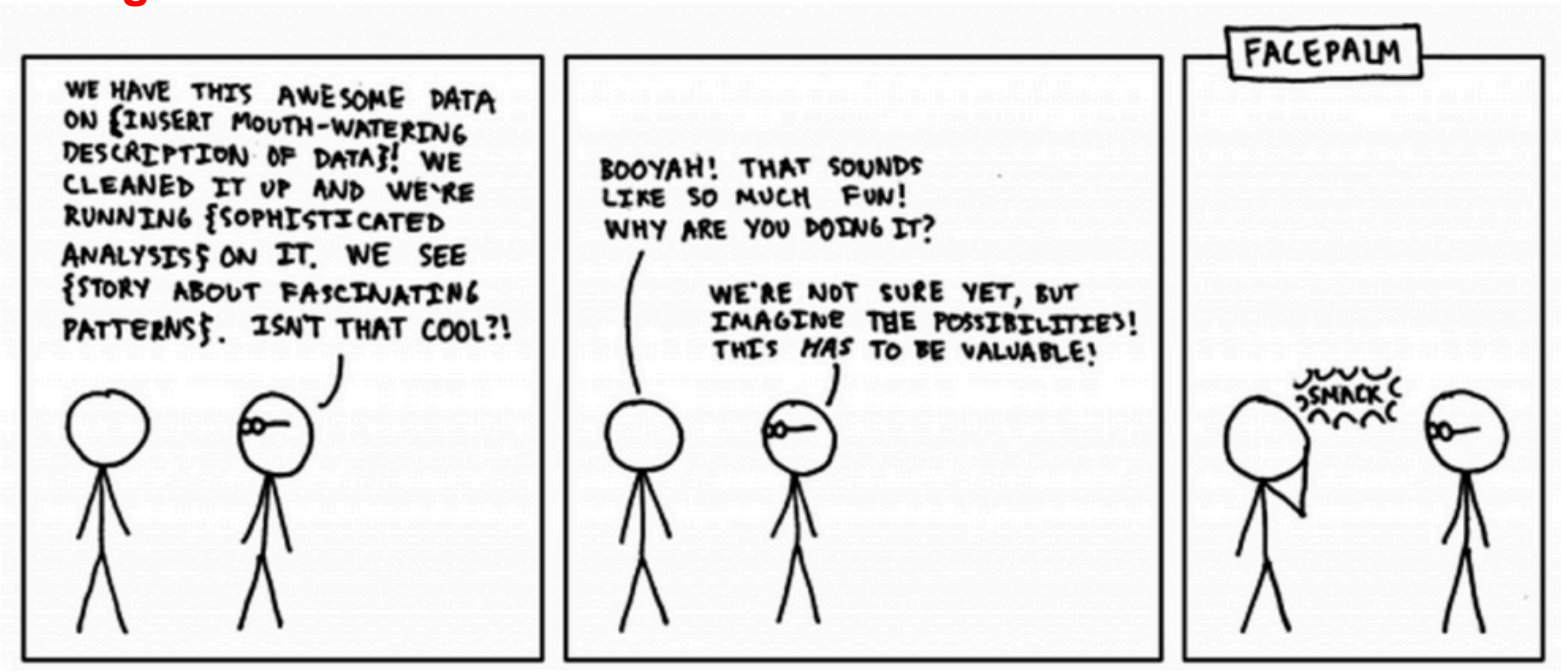
BIG-S²

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Opportunities

Why "Data Scientist" is the sexiest job of the 21st century?

Data science is all about
"using automated methods to analyze massive amounts of data and to extract knowledge from them".



https://www.import.io/post/why-data-scientist-is-being-called-the-sexiest-job-of-the-21st-century/

How much money data scientists make?

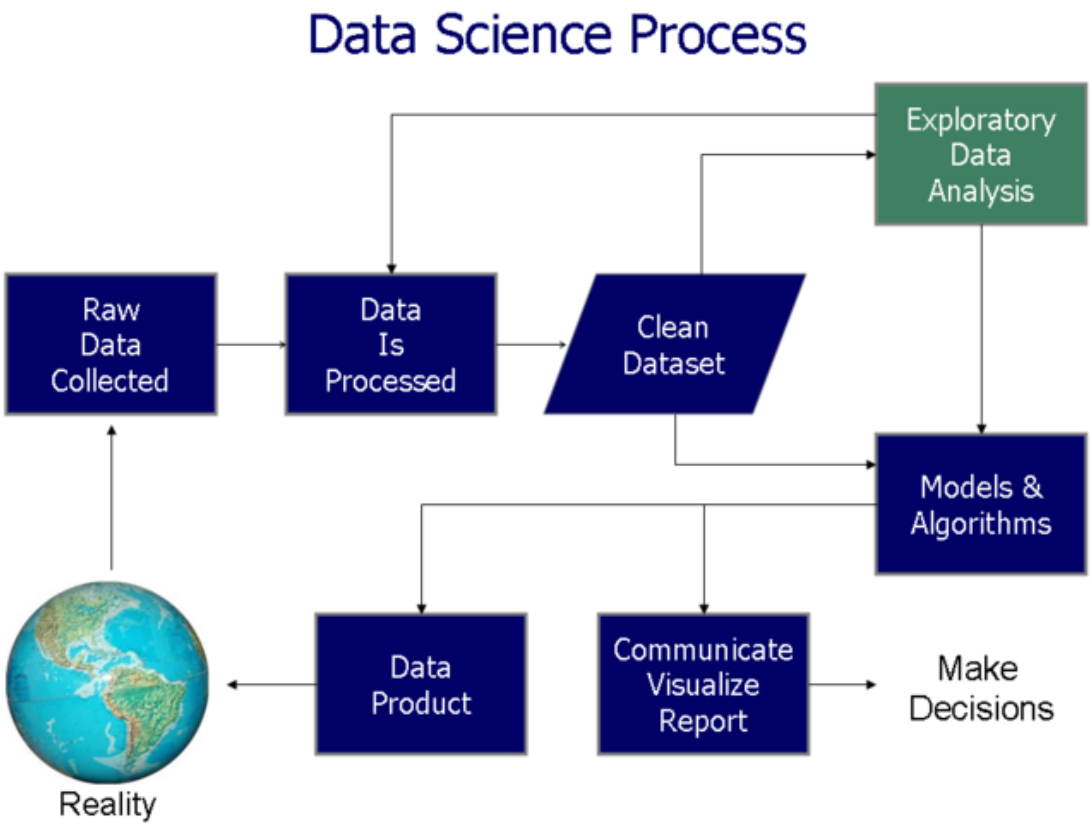
According to Glassdoor, data scientists earn a base pay of $116,840 a year, on average.

Here's how much they rake in, on average, at some of the hottest tech companies, according to Glassdoor employee salary reviews:

Facebook: $133,841
Apple: $149,963
Airbnb: $117,229
Twitter: $134,861
Microsoft: $119,129
LinkedIn: $138,798
IBM: $110,823

http://www.businessinsider.com/how-much-money-you-earn-in-the-sexiest-job-of-the-21st-century-2016-2

Data Science=Statistics?



It employs techniques and theories drawn from many fields within the broad areas of ***mathematics, statistics, information science, and computer science***.

https://en.wikipedia.org/wiki/Data_science

The Big Data Era

Big data is a term for data sets that are **so large or complex** that **traditional data processing application software is inadequate to deal with them**.

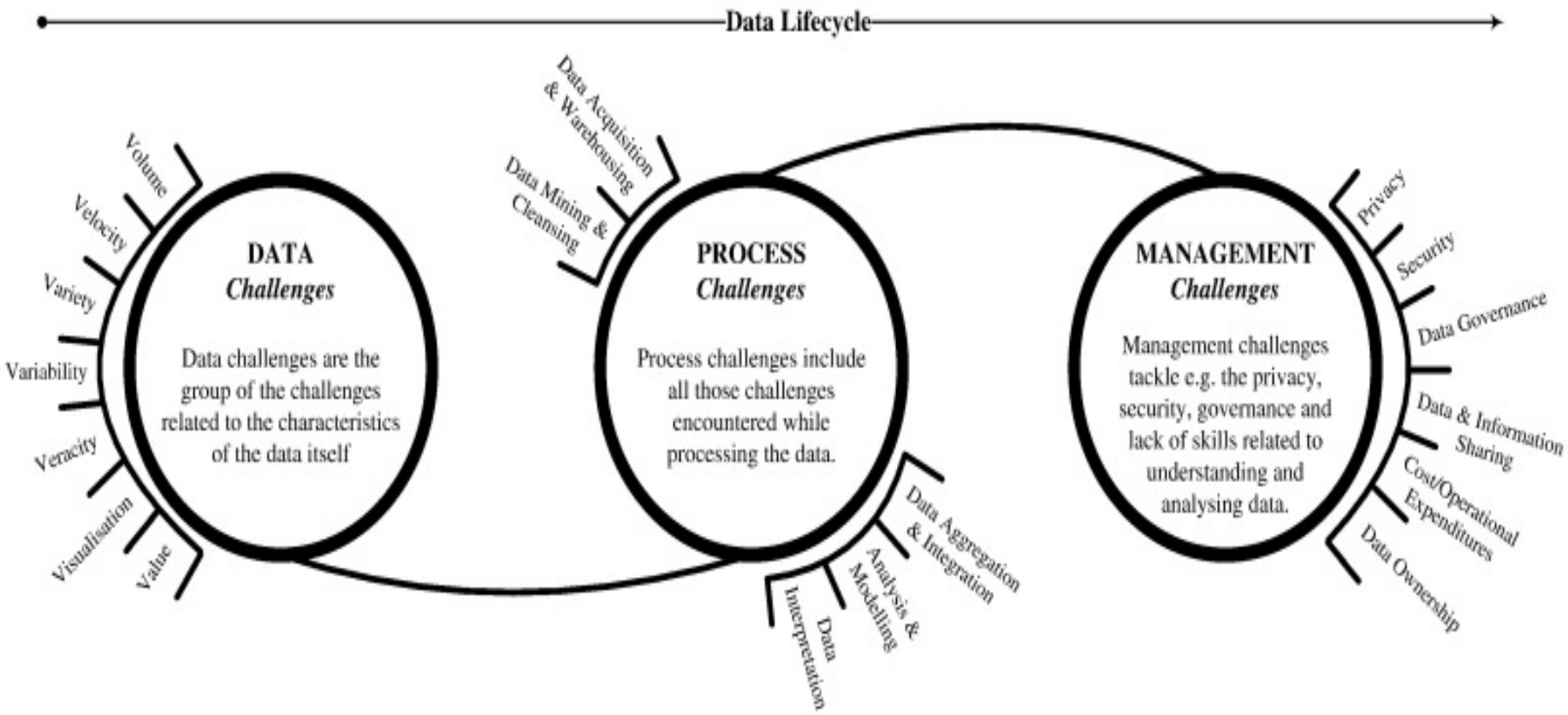
Challenges include ***capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating*** *and* ***information privacy***.

A 2011 McKinsey Global Institute report characterizes the main components and ecosystem of big data as follows:

- **Techniques for analyzing data**, such as A/B testing, machine learning and natural language processing
- **Big data technologies**, like business intelligence, cloud computing and databases
- **Visualization**, such as charts, graphs and other displays of the data

https://en.wikipedia.org/wiki/Big_data

Is Statistics ready for the Big Data era?



Sivarajah, U., Mustafa Kamal, M., Zahir Irani , V. Weerakkody (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research.* 70, 263-286.

Large-scale Medical Studies

Big Data to Knowledge (BD2K)

The four aims of BD2K are

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.

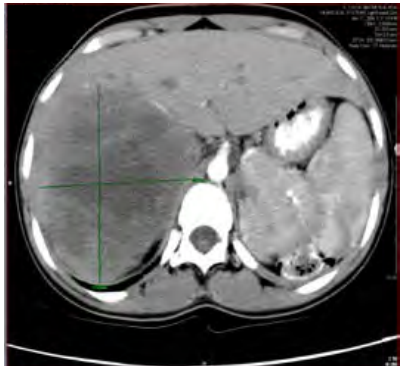
Precision Medicine

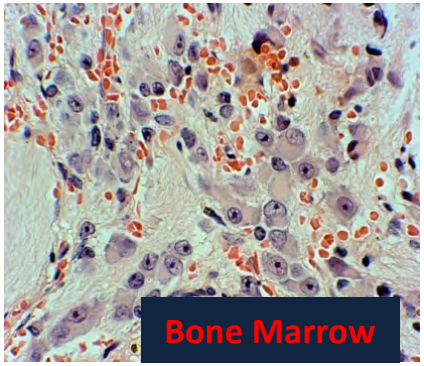
***Precision medicine* (PM) is a medical model that proposes the customization of healthcare—with medical decisions, practices, and/or products being tailored to the individual patient.**

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





Big Data Sets

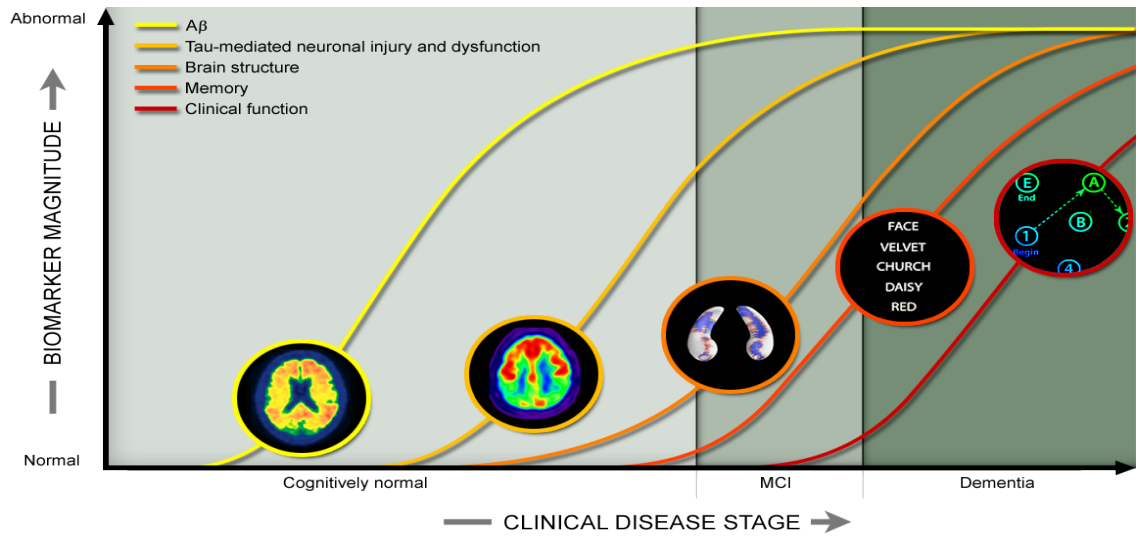
- Alzheimer's Disease Neuroimaging Initiative (ADNI)
- Human Connectome Project
- UNC Baby Neurodevelopment Study
- The Philadelphia Neurodevelopmental Cohort (PNC)
- Pediatric Imaging, Neurocognition, and Genetics (PING)
- UK Biobank
- The Cancer Genome Atlas (TCGA)
- The National Lung Screening Trial (NLST)
- Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)

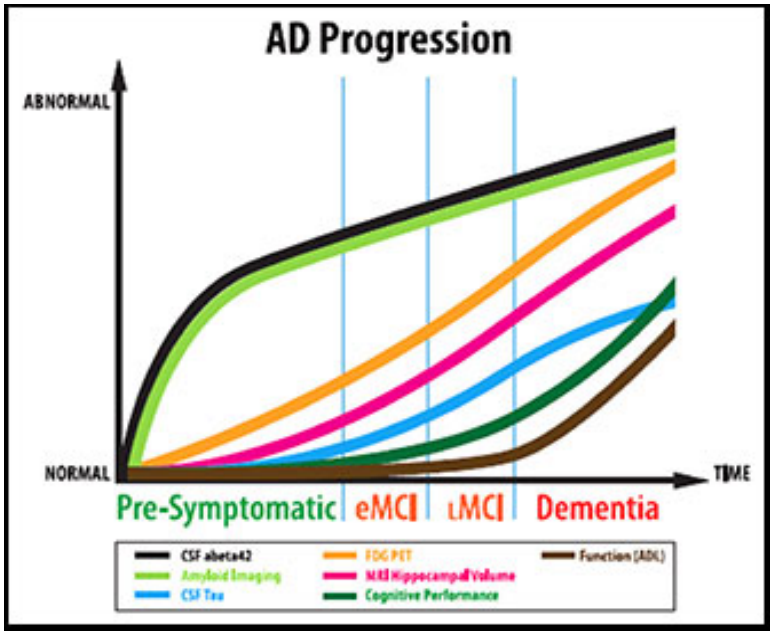
Alzheimer's Disease Neuroimaging Initiative

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





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.**

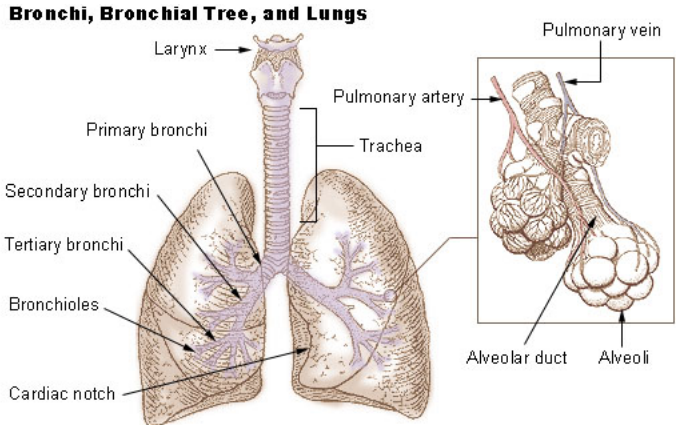
UK Biobank Project

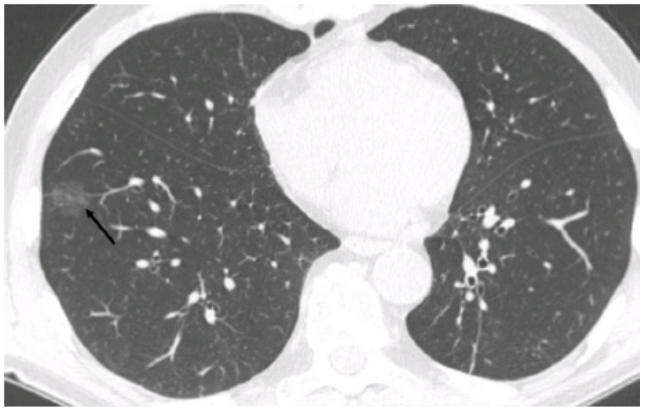
The National Lung Screening Trial

The goal of NLST is to compare two ways of detecting lung cancer: low-dose helical computed tomography (CT) and standard chest X-ray.

Launched in 2002, the initial findings were released in November 2010. On June 29, 2011, the primary results were published online in the *New England Journal of Medicine* and appeared in the August 4, 2011, print issue.

NLST enrolled 53,454 current or former heavy smokers ages 55 to 74. Participants were required to have a smoking history of at least 30 pack-years and were either current or former smokers without signs, symptoms, or history of lung cancer. Participants were randomly assigned to receive three annual screens with either low-dose helical CT or standard chest X-ray.



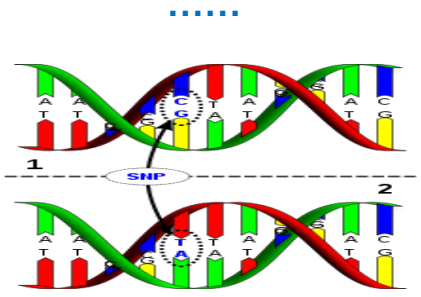


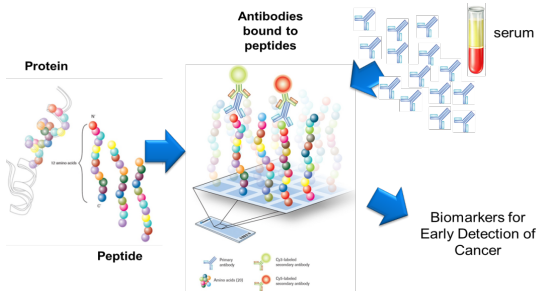
Challenges

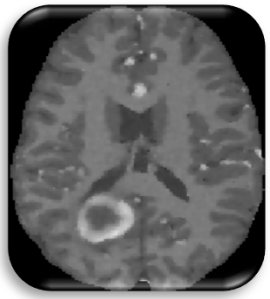
Understand the Connection between Disease Process and Data

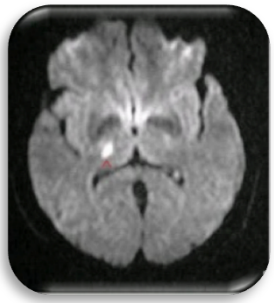
Genetic/genomic, Imaging, Clinical data

Screening	Diagnosis	Treatment	Prognosis
High/Low Risk	DZ/NC Degree	Planning	Response









Challenge 1. Is it 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 have capability to collect such information.

The Cancer Genome Atlas (TCGA)

TCGA is a project, begun in 2005, to catalogue genetic mutations responsible for cancer, using genome sequencing and bioinformatics.

TCGA applies high-throughput genome analysis techniques to improve our ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease.

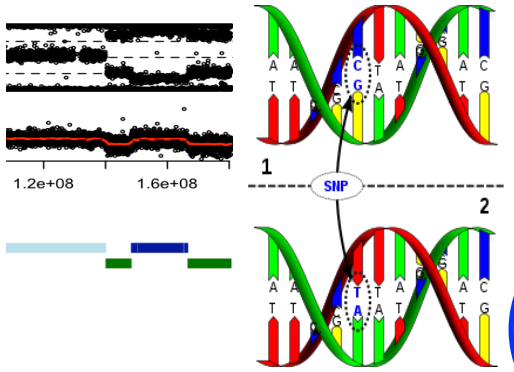
7,136 cases across 20+ tumor types

5865 with minimum clinical data set

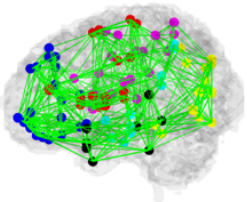
3893 with at least 1 year follow-up; ~50% with treatment data

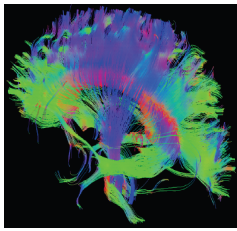
Challenge 2. How to Do Big Data Integration?

E: environmental factors



G: genetic/genomics





I: imaging/device

D: disease

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

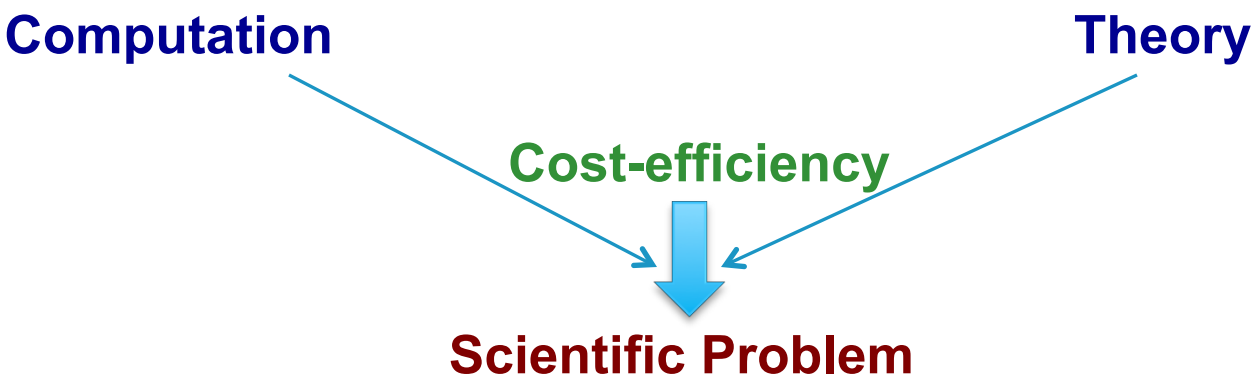
Statistical Challenges

Should we start with statistics and then real data ?

How to discover several key patterns of a complex data set for a scientific problem?

Is there a unified statistical theory associated with each statistical method?

How to evaluate a set of competing methods?



Challenge 3: Big Data Kaggle Competitions

Data Science Bowl 2017

Can you improve lung cancer detection?

Featured · 10 days to go

$1,000,000

1,889 teams

The Nature Conservancy Fisheries Monitoring

Can you detect and classify species of fish?

Featured · 10 days to go

$150,000

2,247 teams

Intel & MobileODT Cervical Cancer Screening

Which cancer treatment will be most effective?

Featured · 3 months to go

$100,000

227 teams

Google Cloud & YouTube-8M Video Understanding Challenge

Can you produce the best video tag predictions?

Featured · 2 months to go

$100,000

382 teams

https://www.kaggle.com/competitions

Grand Challenges in Biomedical Image Analysis

2017

Workshop: Apr 18, 2017
Associated with: ISBI 2017
Hosted on: grand-challenge.org

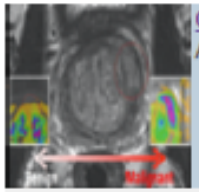
CAMELYON17

Automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections. This task has high clinical relevance and would normally require extensive microscopic assessment by pathologists.

Associated with: MICCAI 2017

ENIGMA Cerebellum

This challenge investigates progress in cerebellum lobule segmentation and labeling, structured around three MRI datasets.



Open for submissions
Associated with: SPIE MI 2017

PROSTATEx

Diagnostic classification of clinically significant prostate lesions using quantitatve image analysis methods.

Associated with: DSB 2017

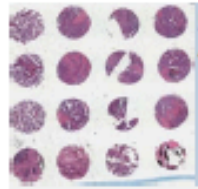
Data Science Bowl 2017

Can you improve the detection of lung cancer? Given a CT scan, predict if someone has lung cancer. The 2017 Data Science Bowl has $1 million in prizes.

Open for submissions
Associated with: ISBI 2017

Skin Lesion Analysis Towards Melanoma Detection

Segment, analyze and diagnose skin cancer from dermoscopic images provided by the International Skin Imaging Collaboration.

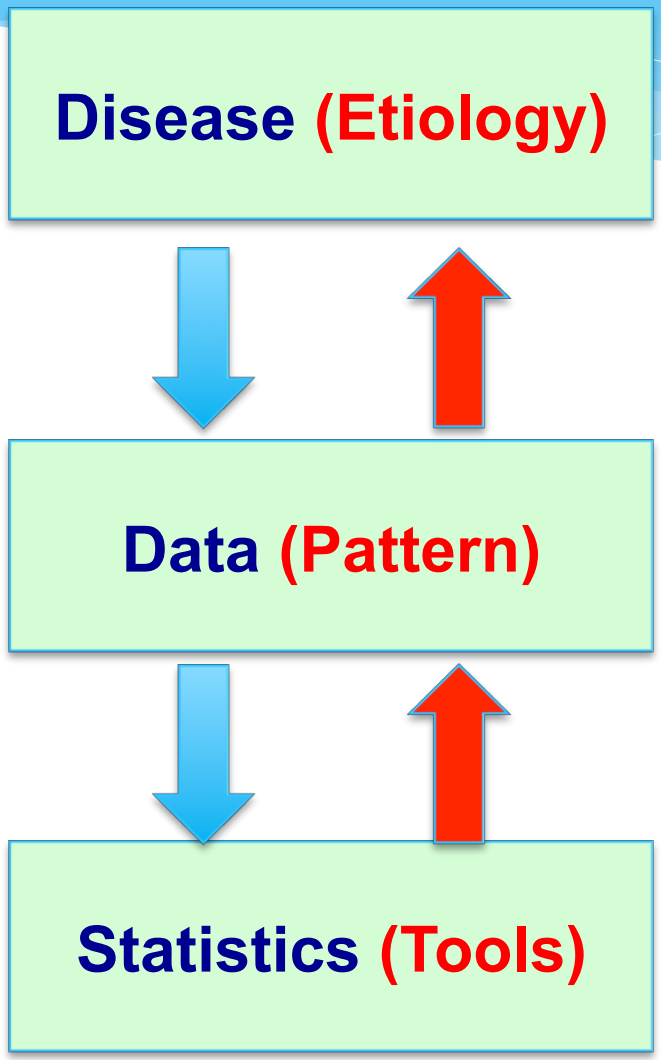


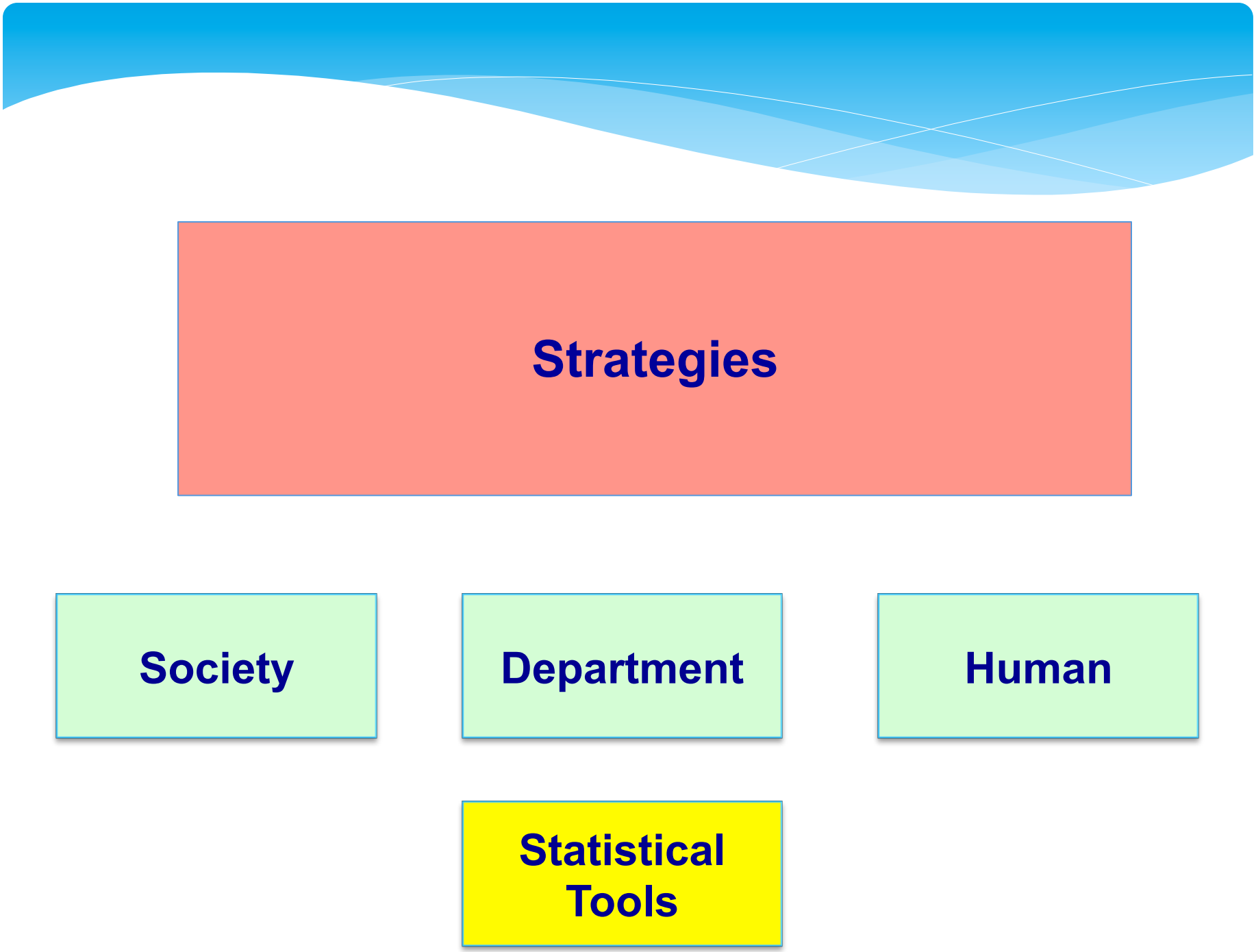
Open for submissions
Associated with: ISBI 2017

Tissue Microarray Analysis for Thyroid Cancer

Tissue microarrays (TMAs) can provide new biomarkers that could be of value in diagnosis, predicting outcome and response to therapy. Goal of this challenge is to build predictions models for thyroid cancer from TMAs.

Challenge 4: Disease, Data and Statistics




Strategies
Society
Department
Human
Statistical Tools

Case Study I: Tumor Heterogeneity

Collaborating with

Wenyi Wang, Kaixian Yu, Jasmine Yang, Steven Marron

Tumor Heterogeneity: identification

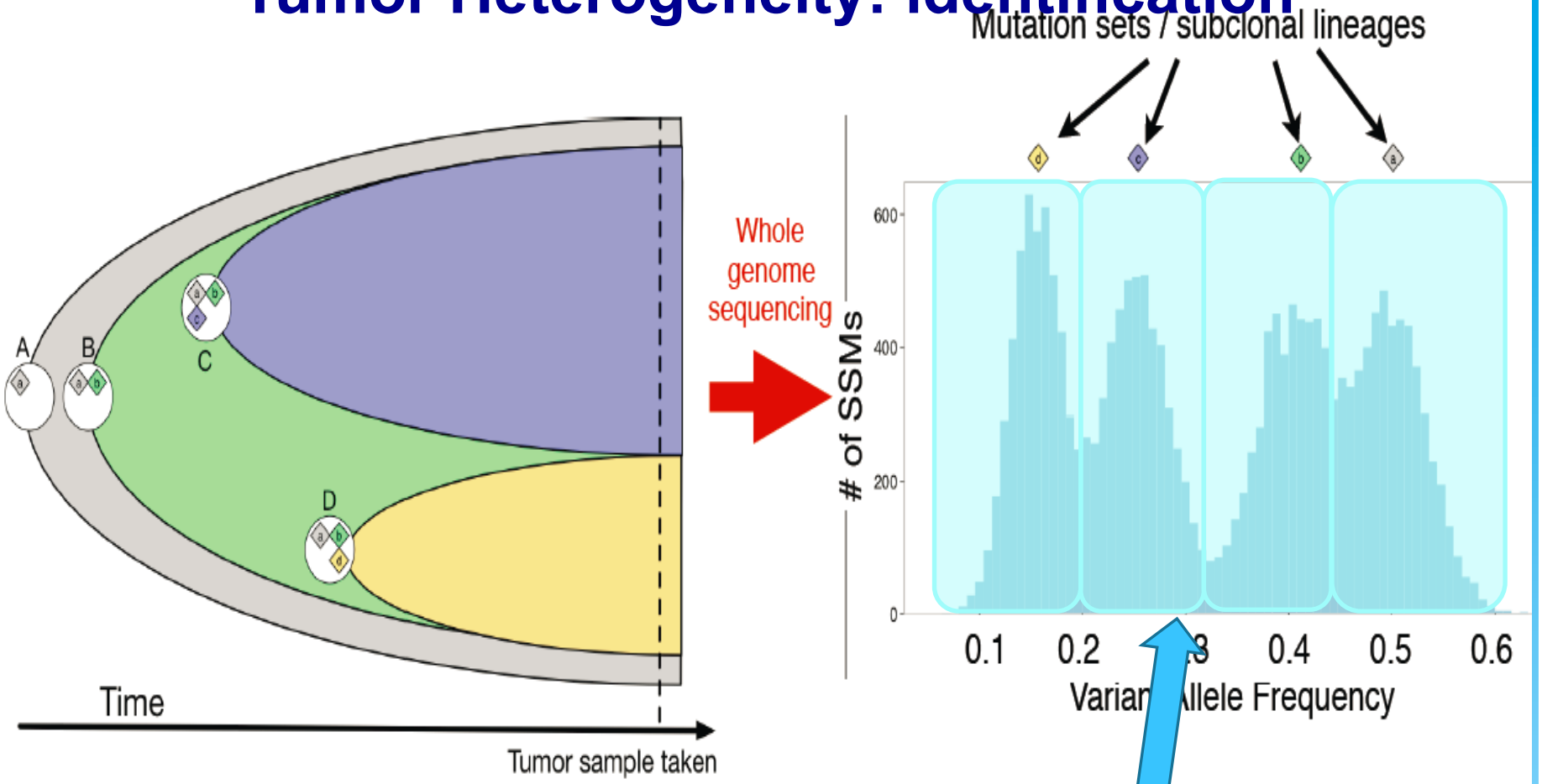


*figure by Quaid Morris

We developed a regularized model based statistical method, named CliP (stands for Clonal and subclonal structure identification through Pairwise difference penalization), to distinguish the sub-clones.

CLIP

CliP is designed for low coverage whole genome sequencing data (30-60 folds)
some key assumptions:

- Somatic mutations: the non-malignant cells are assumed to be wildtype
- Infinite sites assumption: only one type of mutation at each locus
- The clonal cancer cell frequency (CCF) is 1, the clones with CCF>1 will be considered as super-clone; hence, will not be counted for the phylogenic tree

Model:

- Let r_i be the variant reads and n_i total reads at site i. Suppose θ_i is the variant allele frequency at site i. We have

$$r_i \sim Binomial\ (n_i, \theta_i)$$

- Further, let us denote b_i as allele-specific copy number for variant and c_i total copy number at site i, so we have the following:

$$\theta_i = f(\phi_i) = \frac{\phi_i b_i}{2(1-\phi_i) + \phi_i c_i}$$

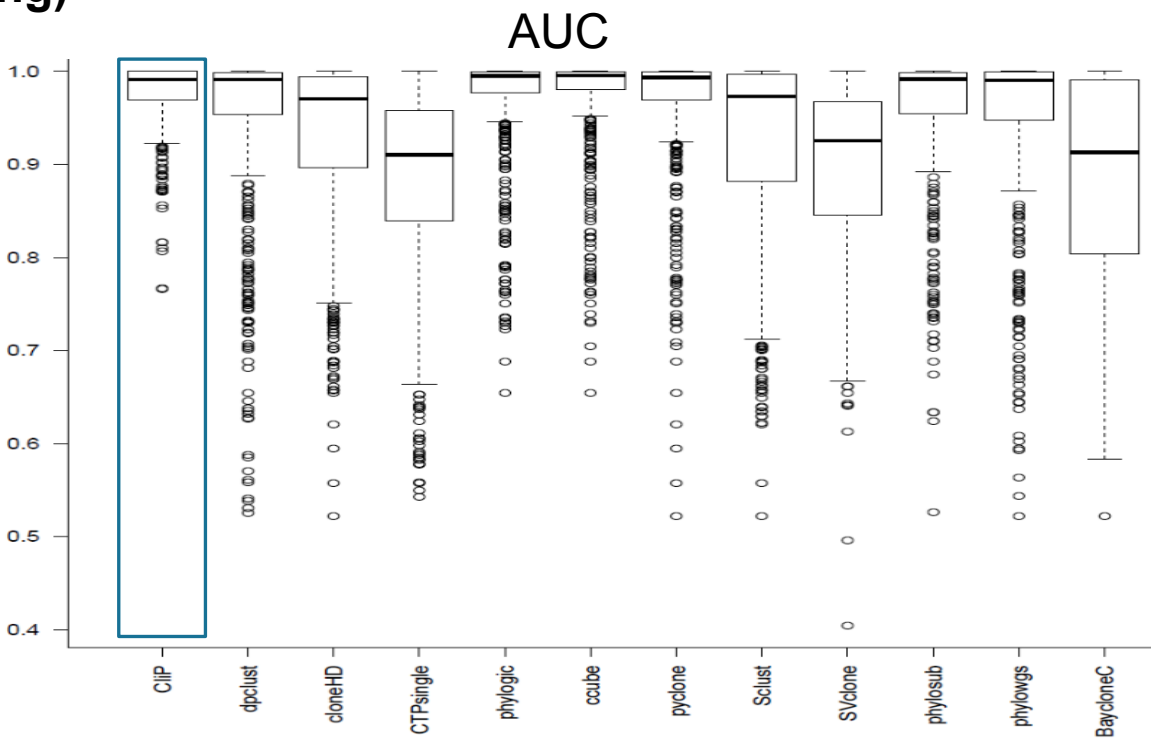
where ϕ_i, the cellular prevalence, is our main interest. Since $\phi_i = \phi_j$ if mutations i, j are in the same clone.

$$Q(\phi;\lambda) = \frac{1}{2}\sum_{i=1}^{N} \frac{n_i(\theta_i - \widehat{\theta}_i)^2}{\theta_i(1-\theta_i)} + \sum_{1\le i<j\le N} p_\lambda(|\phi_i - \phi_j|)$$

Simulated data

- **We tested CliP on 500 simulated samples generated by the Broad Institution, all samples are generated using copy number profiles from actual patients samples.**

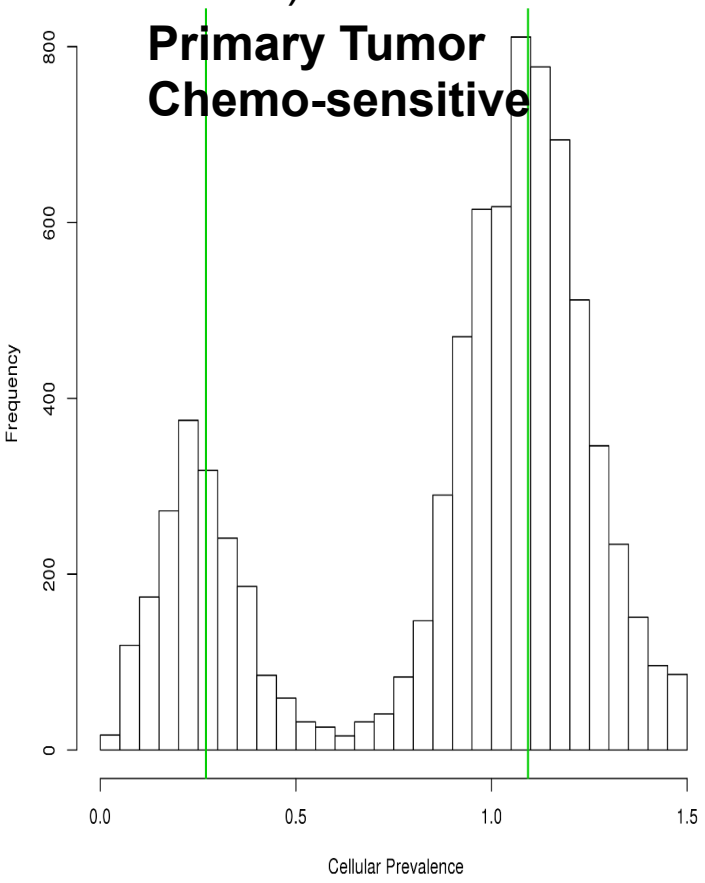
- **The evaluation is done on AUC (ability to separate clonal and subclonal mutations), Adj-rand index (overall mutations assignments), and Earth Mover's Distance (mean absolute error, measures overall topology of clustering)**



Real sample results

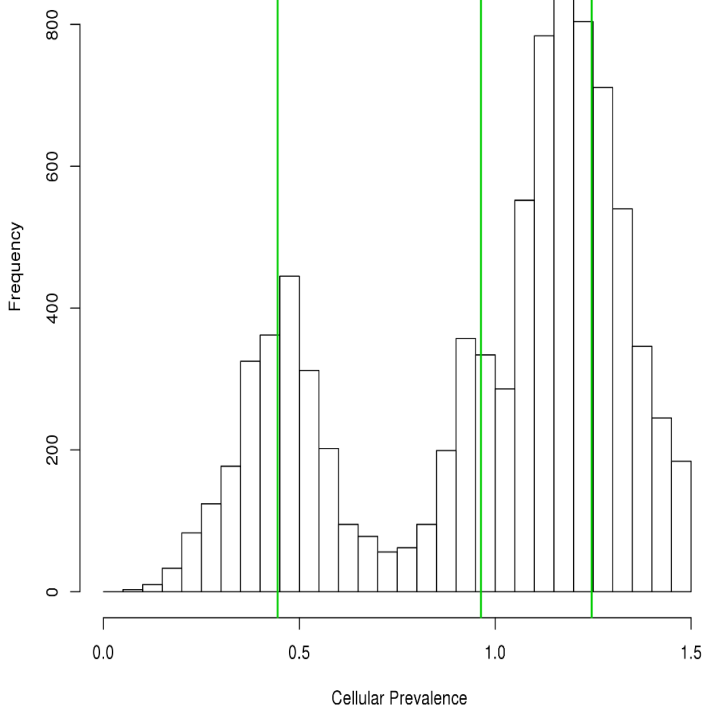
- Most patients with high grade serous ovarian carcinoma relapse and succumb to chemo-resistant disease
- The phenomena may be due to clonality/subclonality change
- We investigated the genomic profile (whole genome sequencing, DNA copy number) of 17 pre- and post-treatment ovarian cancers (by Dr. David Wheeler of Molecular and Human Genetics Human Genome Sequencing Center, Baylor College of Medicine)

Primary Tumor
Chemo-sensitive



Relapse after chemotherapy

Recurrent Tumor
Chemo-resist

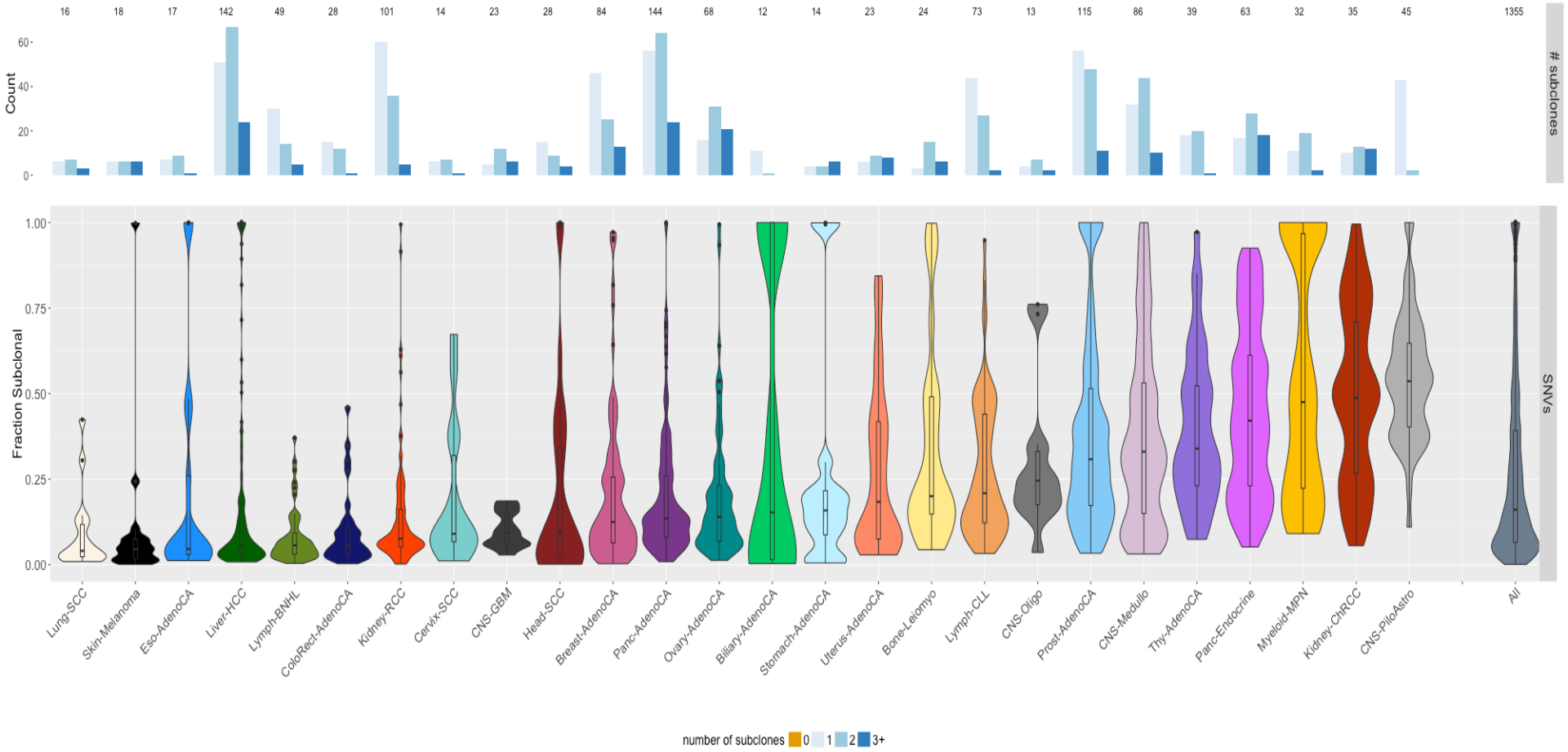


*The green lines give the approximated clone/ subclone positions

Results on ICGC samples

The International Cancer Genome Consortium has collected whole genome sequencing for over 2,700 samples. The clonality study shows that the clon/subclonality compositions are quite different across cancer types.

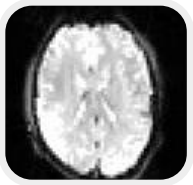
Figure: clonality composition of selected types of cancer. Both the number of subclones and subclonal fractions are different across tumor types



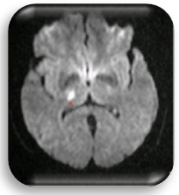
Medical Imaging Analysis

Medical Imaging Analysis and its Applications

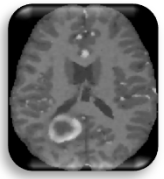
sMRI



fMRI

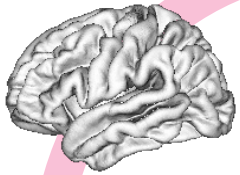


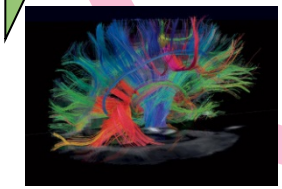
DTI



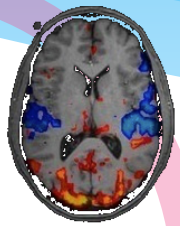
Tumor image

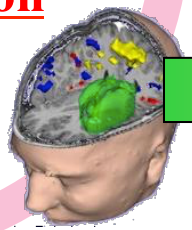
Image Analysis





Biomarker segmentation & visualization





Diagnosis

Detection

Diagnosis

Surgery Planning

Prognosis

Data Knowledge

Clinical Applications

Individual Imaging Analysis

Imaging Construction

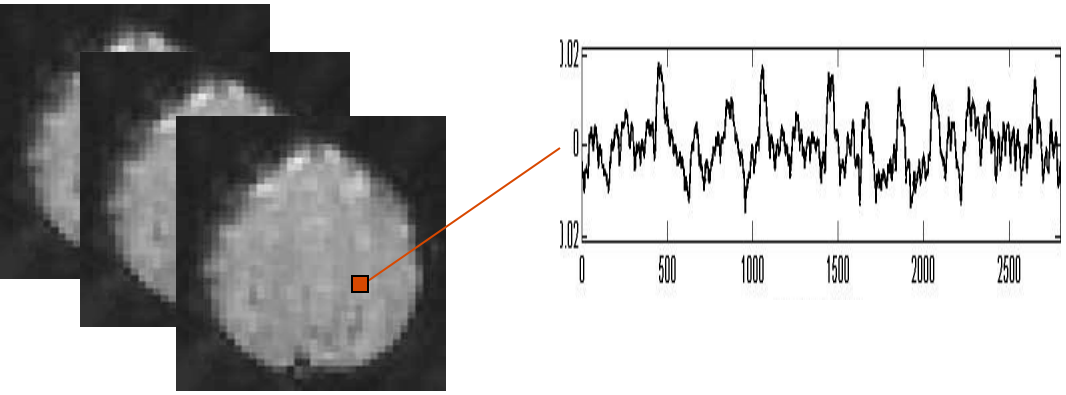
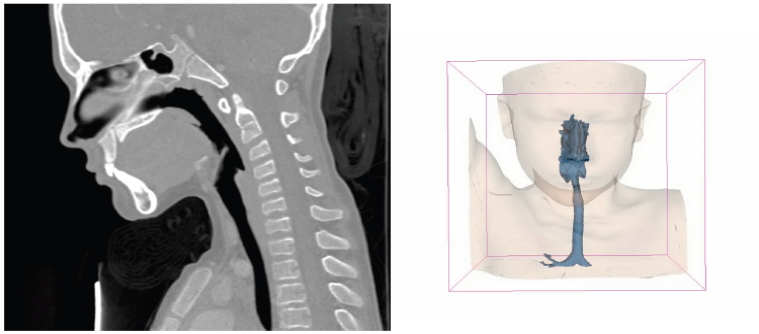


Image Segmentation

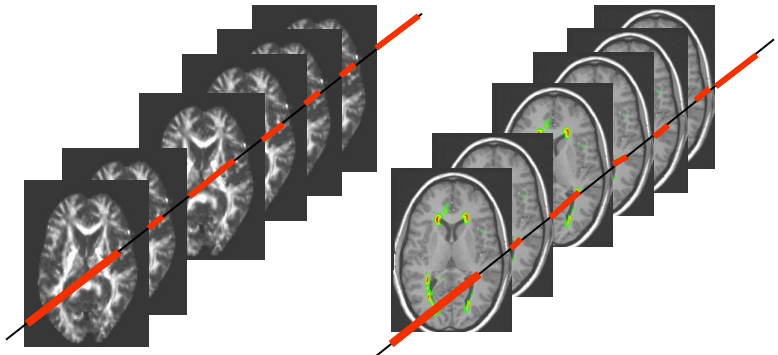
Example: Airway Segmentation from CT

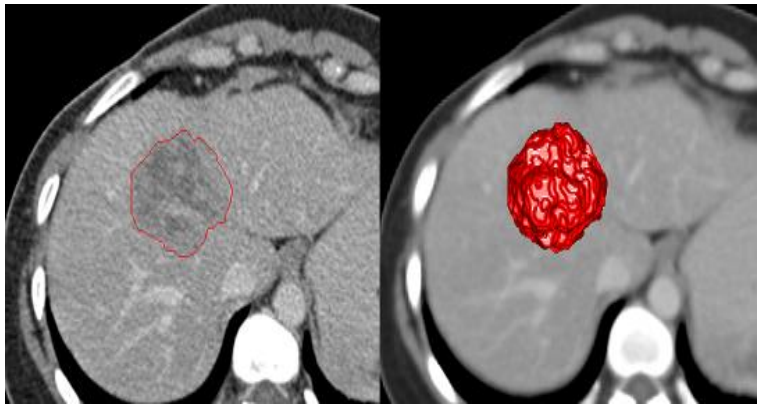


Multimodal Analysis

DTI

FLAIR

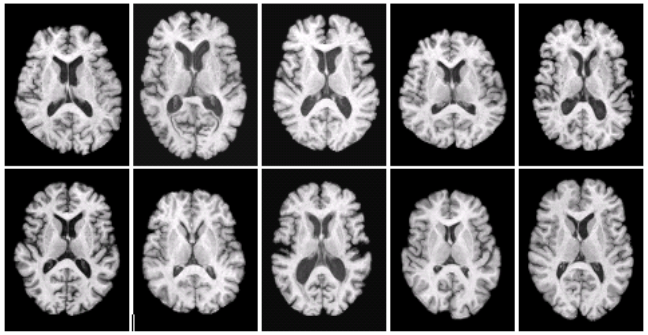


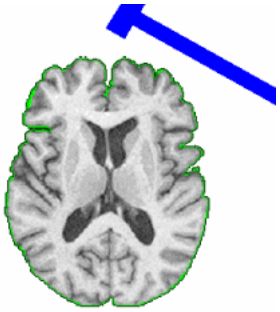


Marc

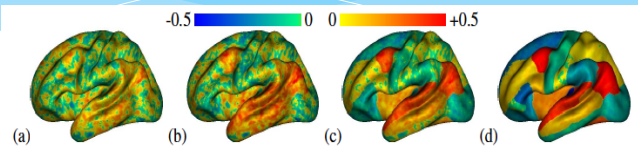
Group Imaging Analysis

Registration



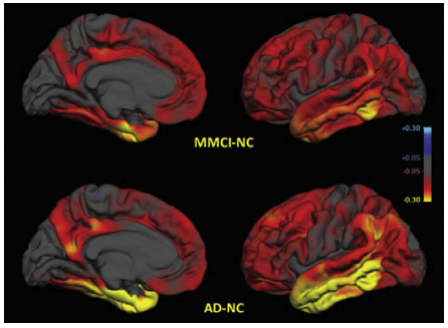


Prediction

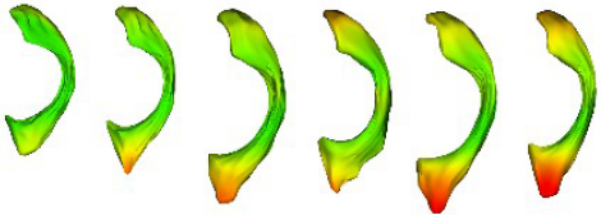


NC/Diseased

Group Differences



Longitudinal/Family Brain



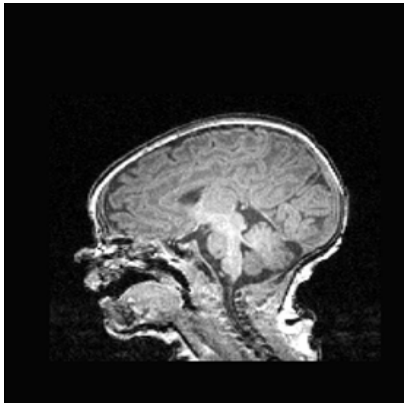
Hibar, Dinggang, Martin

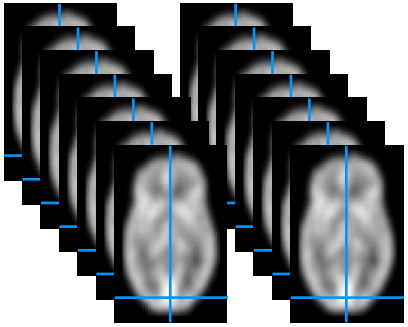
Imaging Genetics

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

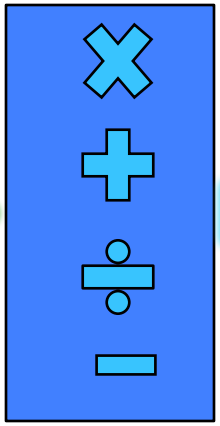
FDA: Functional Data Analysis

$$f$$

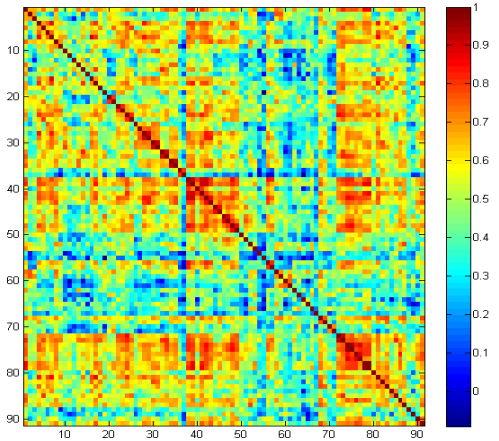


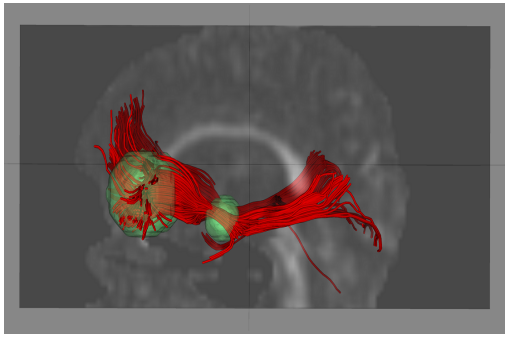


$$T$$



$$\hat{F} = T[f]$$





Case Study II: Structural Connectome Analysis

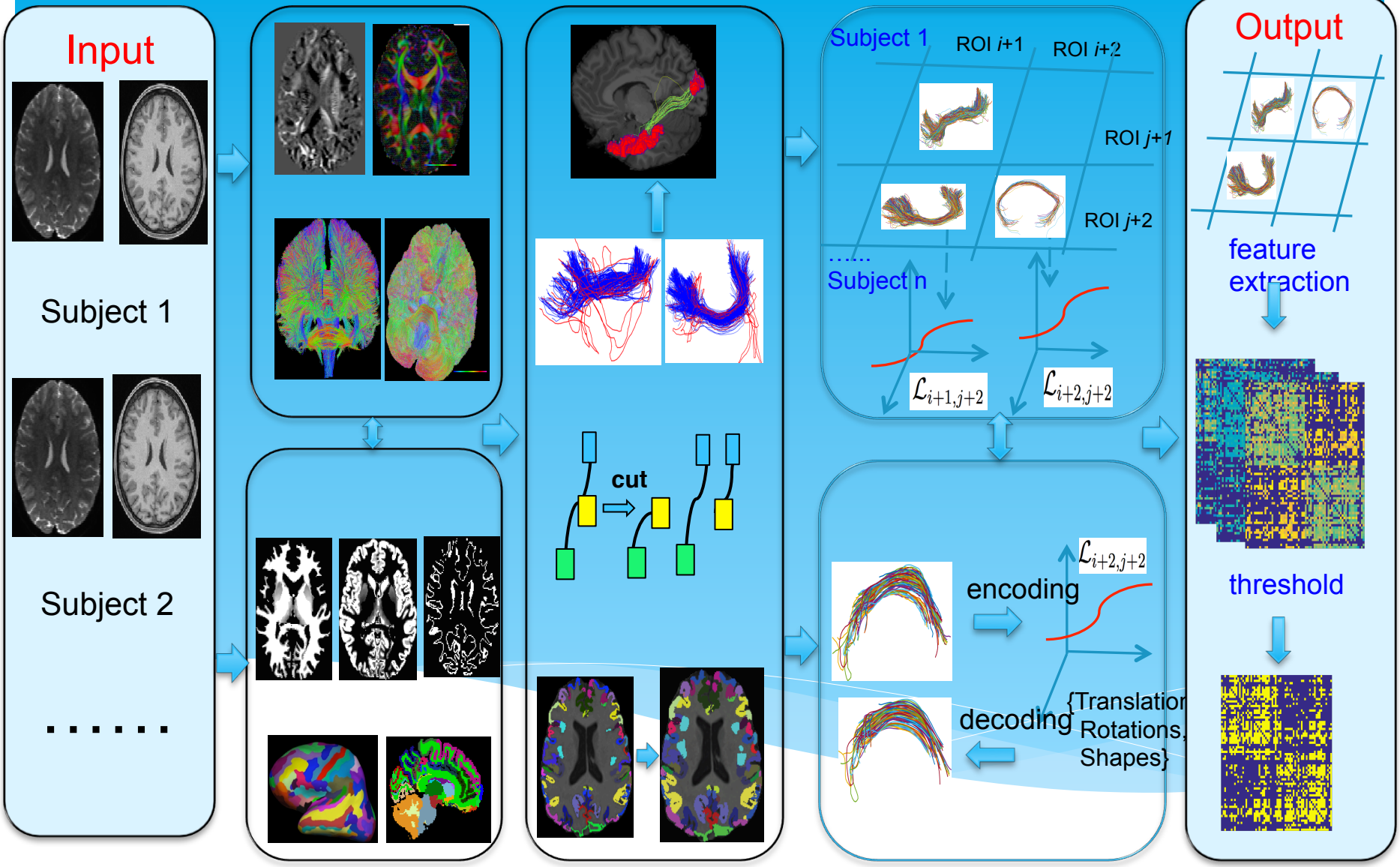
Collaborators:
Zhang, Z.W. (Rochester), Descoteaux,M. (Universite de Sherbrooke), Srivastava, A. (FSU)

Big Neuroimaging Data Sets

Motivation. Is there a robust DTI pipeline for extracting white matter features across different scales?

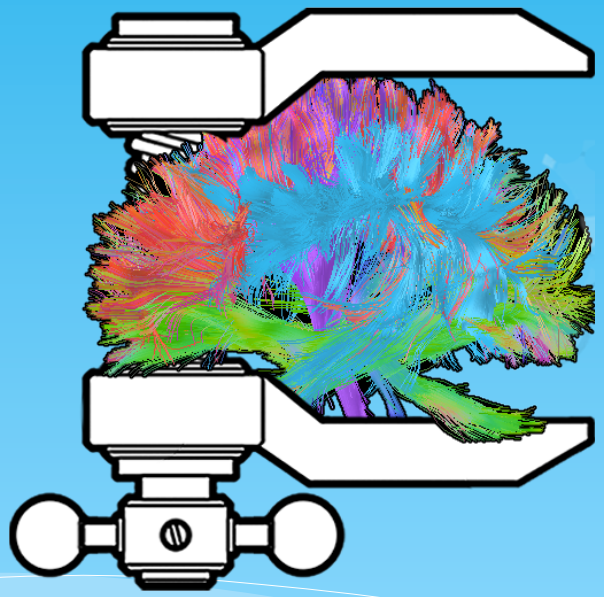
- **UK Biobank**
- **Alzheimer's Disease Neuroimaging Initiative (ADNI)**
- **Human Connectome Project**
- **UNC Baby Neurodevelopment Study**
- **The Philadelphia Neurodevelopmental Cohort (PNC)**
- **Pediatric Imaging, Neurocognition, and Genetics (PING)**

Talk overview: from raw data to tractometry & Connectomics



How to address the computational and theoretical challenges?

Structural Connectome Analysis



Efficient Representation of Streamlines

- Fiber representation: parameterized curves $f : [0, 1] \rightarrow \mathbb{R}^3$
- Examples of fibers in *CM(1,160)* for different subjects

- Observations:
  - They have similar shapes after alignment
  - These shapes can be efficiently represented

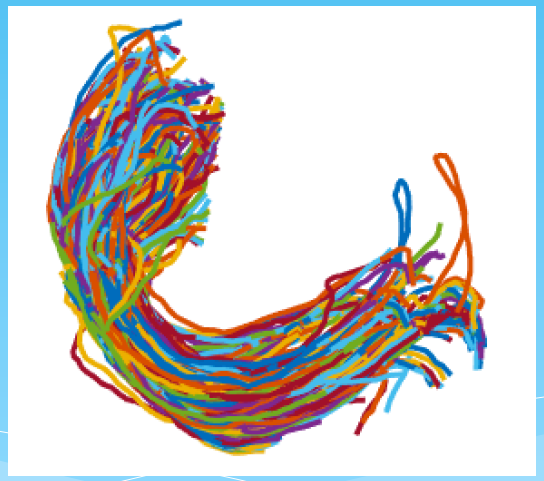
Efficient Representation of Streamlines

- Represent the streamlines through basis and coefficients
- Basis can be learnt from data to increase its representing power

Step 1. Generate atlas for streamlines connecting each pair of regions

Randomly select healthy subjects:

Merge



Efficient Representation of Streamlines

- Represent the streamlines through basis and coefficients
- Basis can be learnt from data to increase its representing power

Step 1. Generate atlas for streamlines connecting each pair of regions

Step 2. Alignment using the **Elastic Shapes Analysis** framework (Srivastava et al. 201

- rotation
- translation
- scaling
- reparameterization

Alignment

*K-means clustering may be used if these streamlines have different shapes

Square-root velocity function (SRVF) and Fisher-Rao metric

Srivastava, A. and Klassen, E. P. (2017) Functional and Shape Data Analysis. Springer.

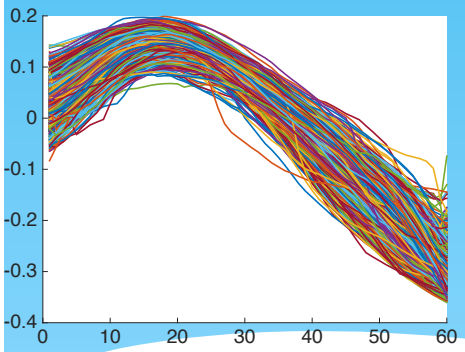
Efficient Representation of Streamlines

- Represent the streamlines through basis and coefficients
- Basis can be learnt from data to increase its representing power

Step 1. Generate atlas for streamlines connecting each pair of regions

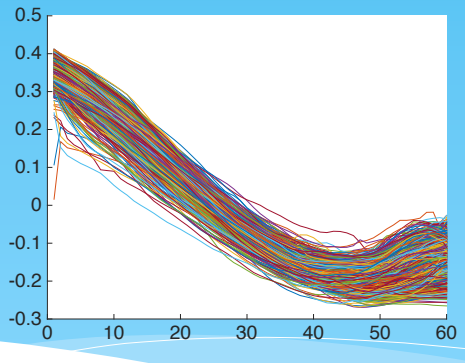
Step 2. Alignment using the Elastic Shapes Analysis framework (Srivastava et al. 2012)

Step 3. Use fPCA to learn basis functions for each component



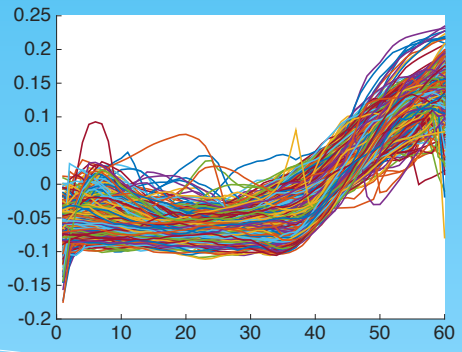
$j = 1$

X



$j = 2$

Y



$j = 3$

Z

μ_j

$\{\phi_{i,j}\}$

Efficient Representation of Streamlines

- Represent the streamlines through basis and coefficients
- Basis can be learnt from data to increase its representing power
- Efficient representation (compression): given a new fiber f

Step 1. Align f to the mean fiber in the atlas

$$\underset{O\in SO(3),C\in\mathbb{R}^3}{\operatorname{argmin}} \|O*(f-C)-\mu\|$$

$$g = O*(f-C)$$

rotation

translation

Step 2. Represent the aligned fiber using basis functions

$$g_j = \mu_j + \sum_{i=1}^{M_j} c_{j,i}\phi_{j,i} + \epsilon_j, j = 1,2,3$$

coefficients

$\|\epsilon_j\|$ determines M_j, the number of coefficients for representing fibers

Step 3. Parameters need to save $\{O, C, c_{j,i}\}$

Efficient Representation of Streamlines

- Represent the streamlines through basis and coefficients
- Basis can be learnt from data to increase its representing power
- Efficient representation (compression): given a new fiber f
- Recover f from saved parameters $\{O, C, c_{j,i}\}$

$$\hat{f} = O' * \hat{g} + C$$

where each component

$$\hat{g}_j = \mu_j + \sum_{i=1}^{M_j} c_{j,i}\phi_{j,i}$$

- Comparison of streamlines can be done through saved parameters

Efficient Representation of Streamlines

- Compression ratio: $\rho = 100 * (1 - \frac{N_c}{N_r})$

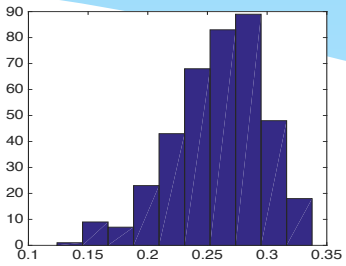
N_c -- # para. after compression, N_r -- # para. before compression

- Example of compressing fibers in *CM(1,160)* and *CM(115,160)*

Error ϵ (mm)	0.1	0.2	0.5	1.0	2.0
Ratio	95.7	97.3	98.4	98.8	99.1

recover

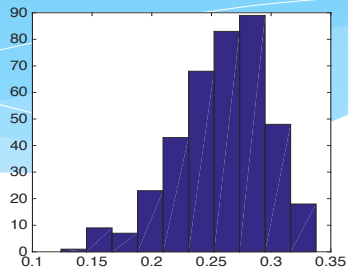
Raw data



Error ϵ (mm)	0.1	0.2	0.5	1.0	2.0
Ratio	93.7	95.6	98.9	97.4	97.6

recover

Raw data



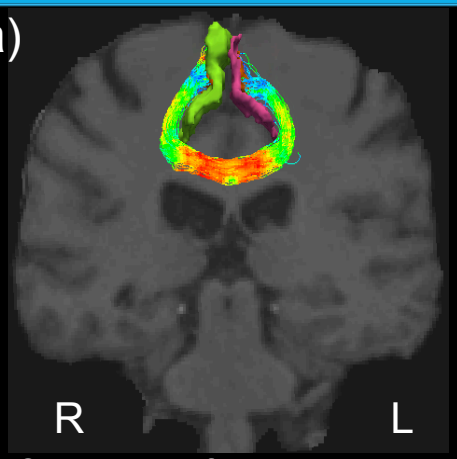
Network Level Analysis

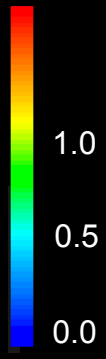
- More robust coupling strength measures
  - Diffusion properties: FA values, AFD values along fibers
  - Geometry properties: Shapes, Loops, Clusters
  - Nodes information: Volume of nodes, Connected surface areas

  …

Network Level Analysis

(a)





Connection of R and L Paracentral lobules

(b)

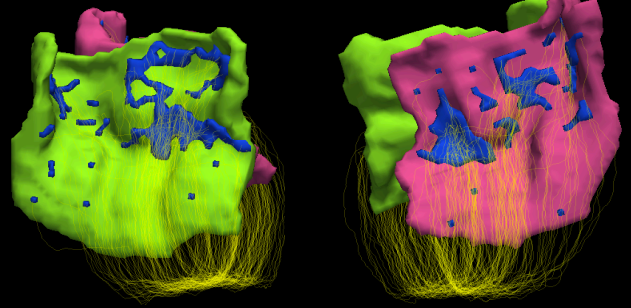
Geometric features:

- Average length
- Average deviation from the mean fiber
- # of clusters

(c)

Endpoint features:

- Fiber count
- Connected surface area



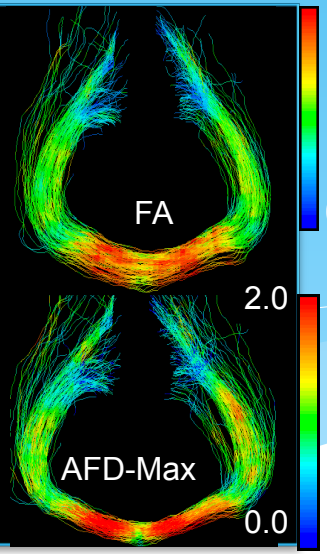
R Paracentral lobule

L Paracentral lobule

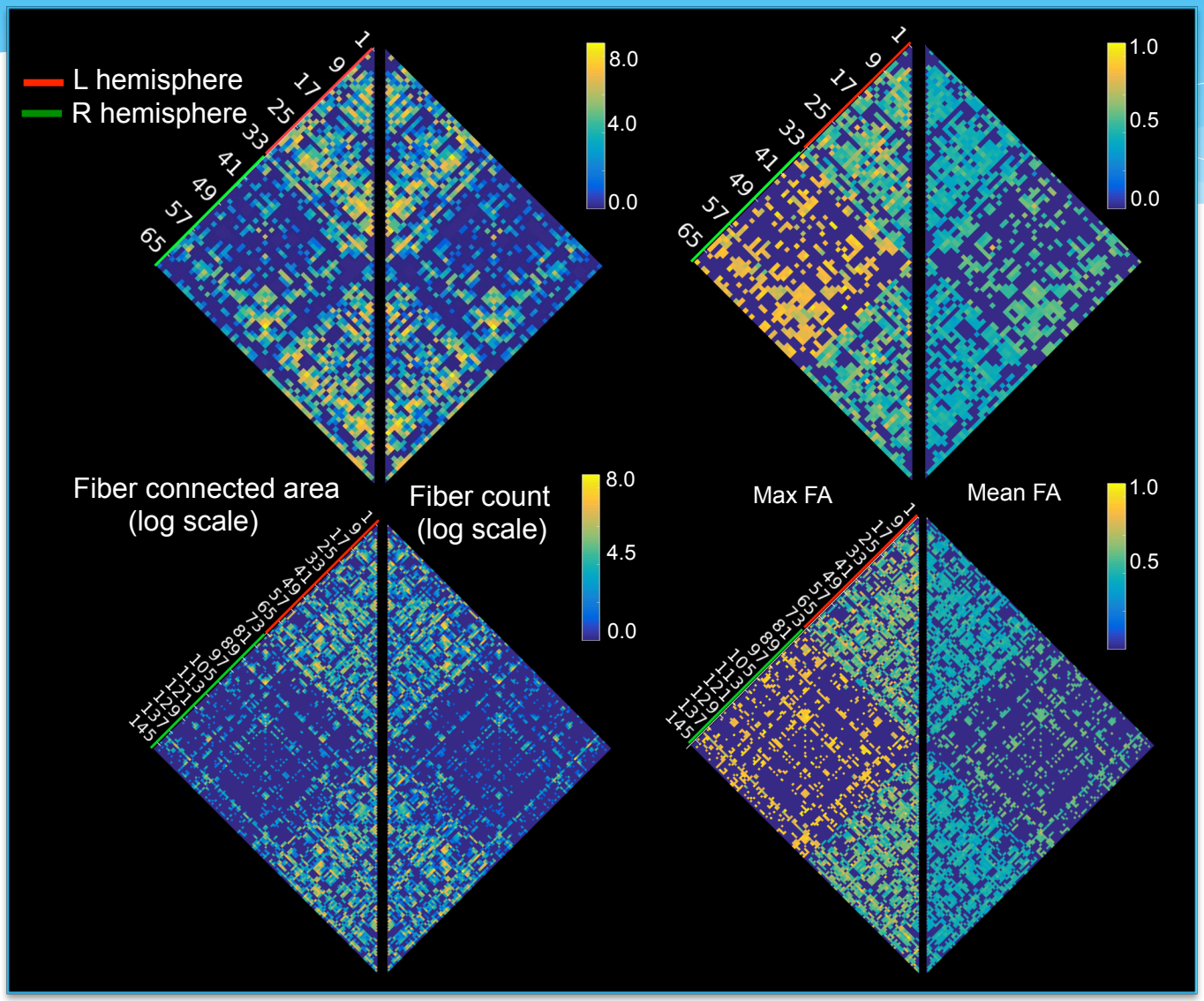
(d)

Diffusion features:

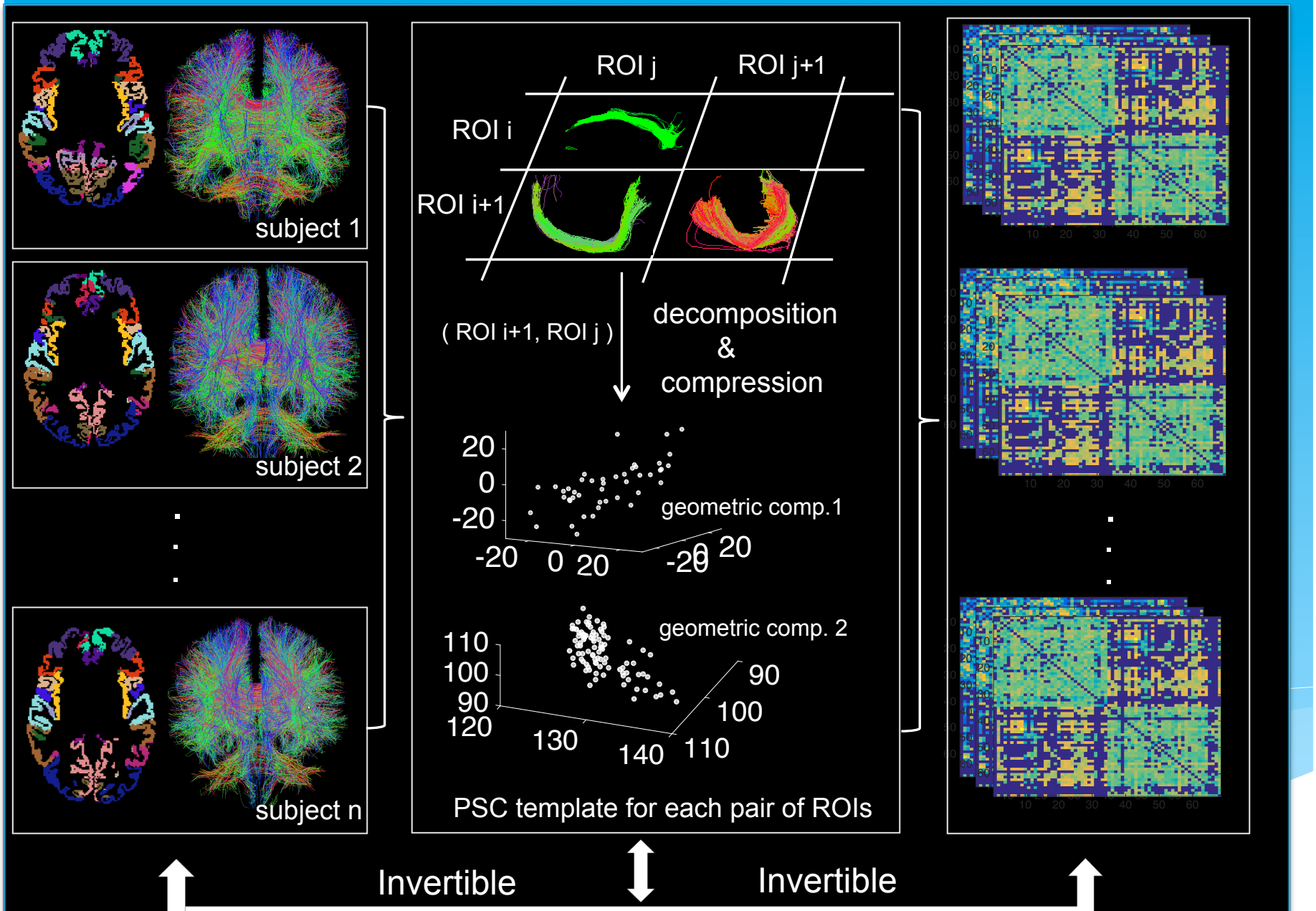
- DTI metrics, such as FA, MD, RD, et al.
- ODF metrics, such as GFA and AFD.



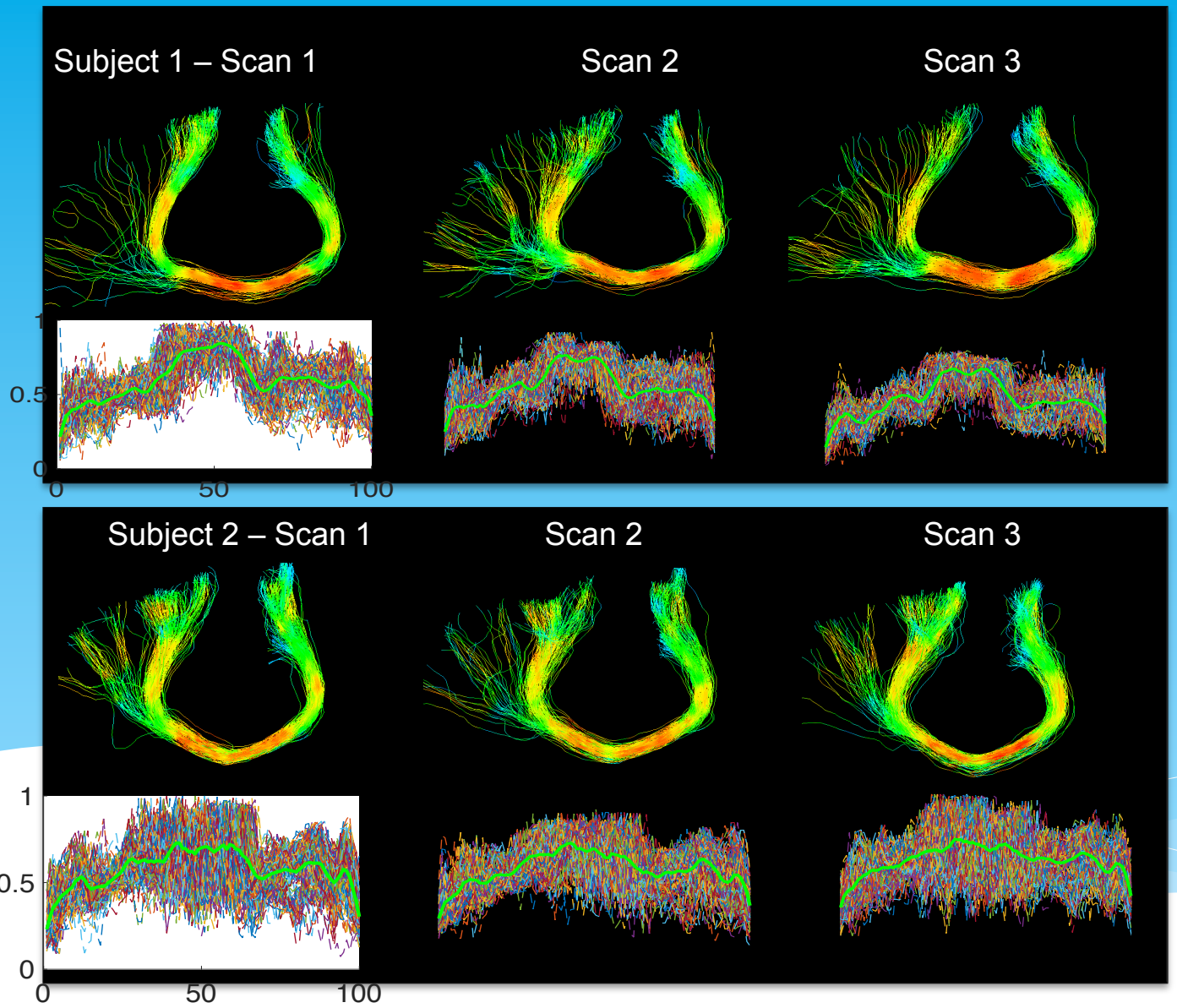
Network Level Analysis



Overview: Our Method

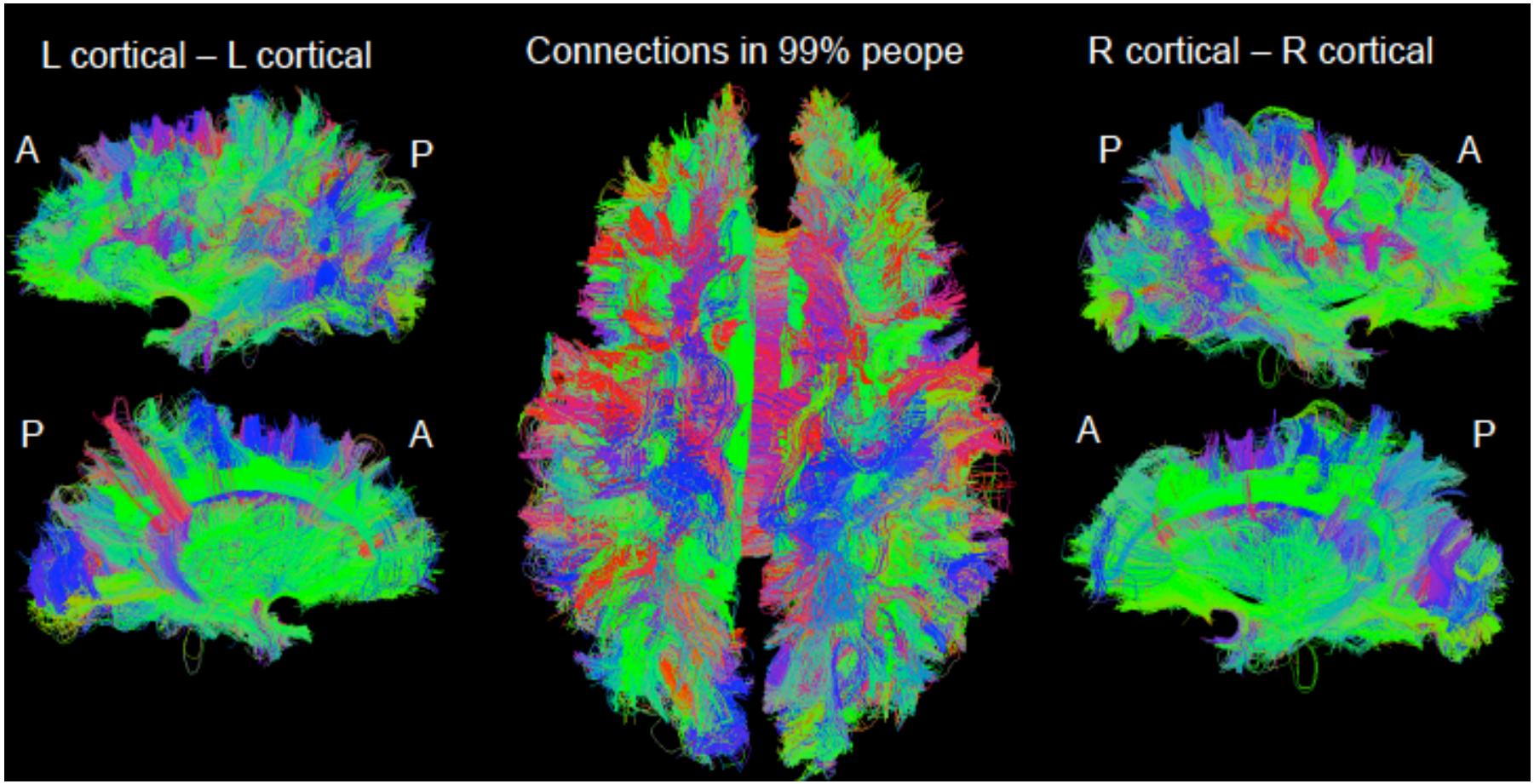


Reproducibility



The Human Connectome Project

857 Subjects

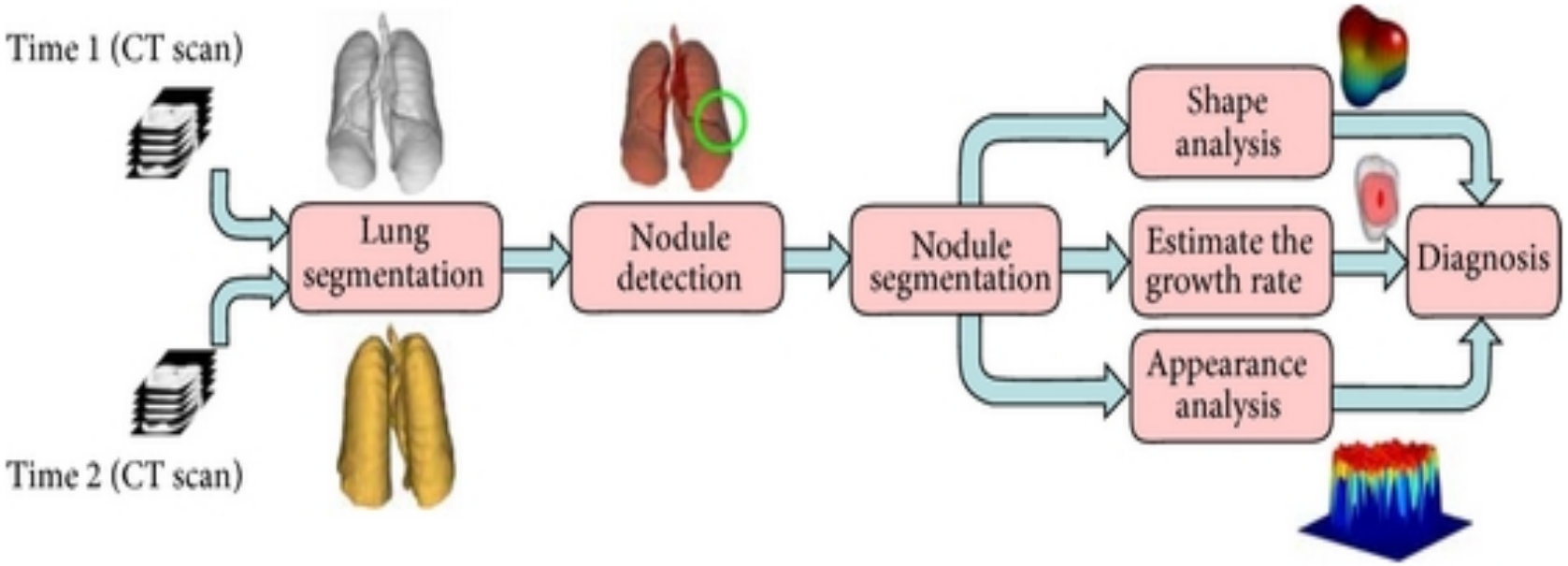


Case Study III: Biomarker Detection

Collaborators: Xiao Wang (Purdue), Y.F. Liu (UNC) and Bigs2 (Kaixian Yu, RJ Liu, M. Ding, Chao Huang, Leo Y.F. Liu, Yu Yang, Heng Li, Fan Zhou, Yue Wang)

Computer-aided Diagonsis

Processing → Segmentation → ROIs → Prediction



https://www.hindawi.com/journals/ijbi/2013/942353/fig1/

Predictive Modelling

Predictive models can either be used directly to estimate a response (output) given a defined set of features (input), or indirectly to drive the choice of decision rules.

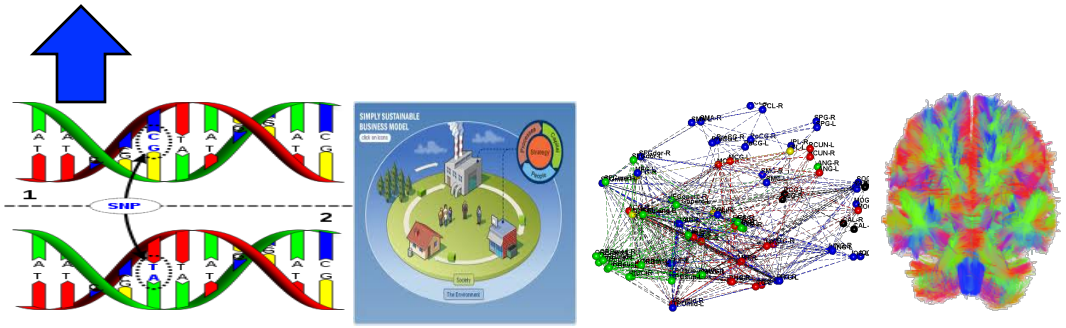
- **Determining the 'correct' features**
- **Fitting the predictive model**
- **Performance assessment**

Formulation

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)

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center
Making Cancer History®

BIG-S²

Functional Linear Models

$$y_i = < X_i, \theta > + \varepsilon_i$$

$$y_i = \theta_0 + \int_D \theta(d) X_i(d) m(d) + \varepsilon_i$$

Targets:

How to estimate $\theta(.)$?

How to achieve better prediction accuracy?

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

BIG-S²

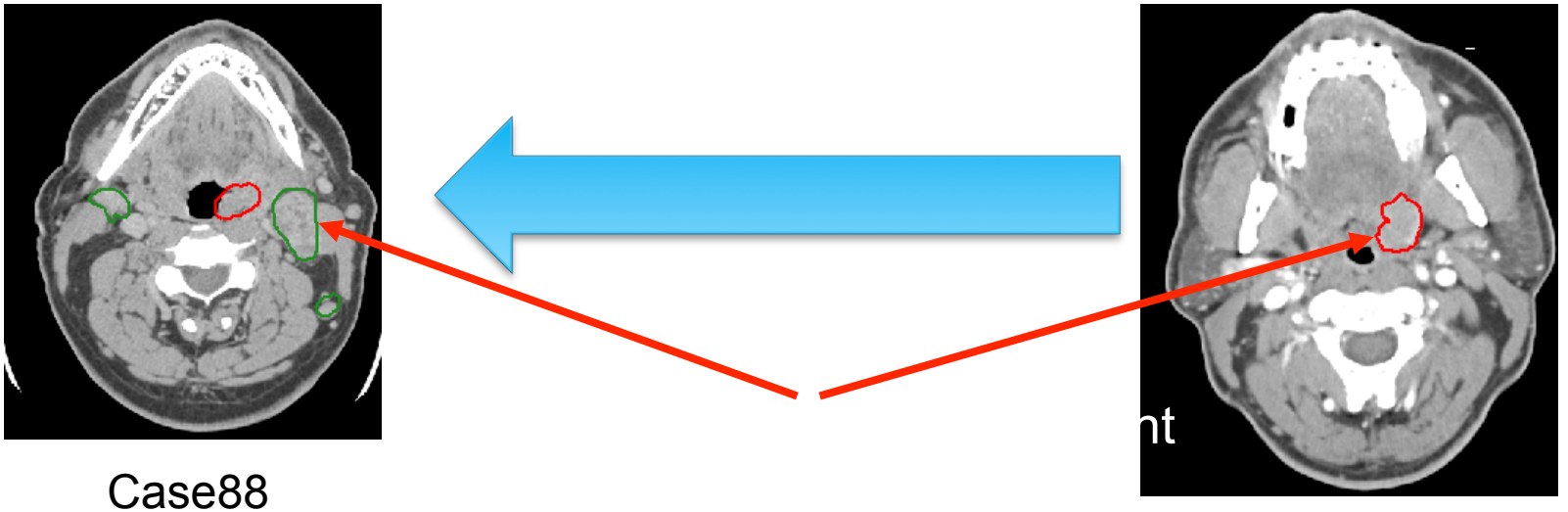
What is wrong with FLM ?

- Lack of a statistical package for handling >1 dimensional functional data
- All functional data are in a common coordinate system
- FLM is too ‘simple’ to be useful in most medical research problems
- Existing methods for computer-aided diagonsis (CADx) are not based on FLM at all.

Major Challenges

Heterogeneity: Location and Size

Tumors can appear ``anywhere''



Case88

Case102


CAMELYON17
ISBI Challenge on cancer metastases detection in lymph node
Radboudumc
university medical center
UMC Utrecht
TU/e
Technische Universiteit Eindhoven
University of Technology
Where innovation starts
PHILIPS
innovation you
trait
a ctmm project
ST
stichting
IT
PRO
IT projecten

Aims of CAMELYON17

* Evaluate new and existing algorithms for automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections.

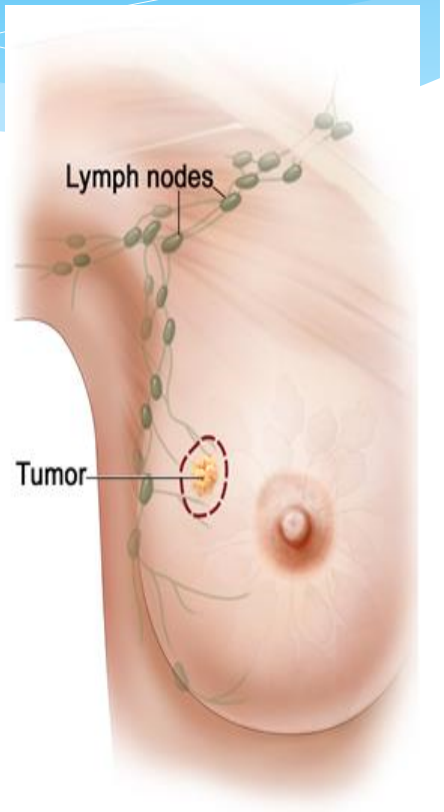
* Combines the detection and classification of metastases in multiple lymph node slides into one outcome: a pN-stage.

* Reduce the workload of pathologists, while at the same time, reduce the subjectivity in diagnosis.

Breast cancer metastases in lymph nodes and the TNM classification system

* The lymph nodes in the axilla are the first place breast cancer is likely to spread. Small metastases are very difficult to detect and sometimes they are missed.

* In breast cancer, TNM staging takes into account the size of the tumour (T-stage), whether the cancer has spread to the regional lymph nodes (N-stage), and whether the tumour has metastasised to other parts of the body (M-stage).

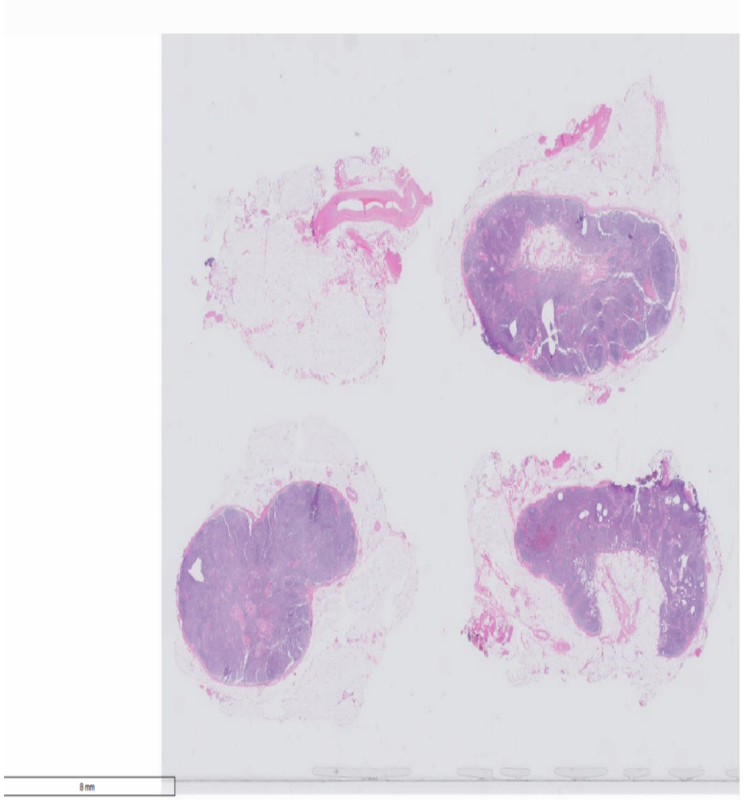


Data Summary

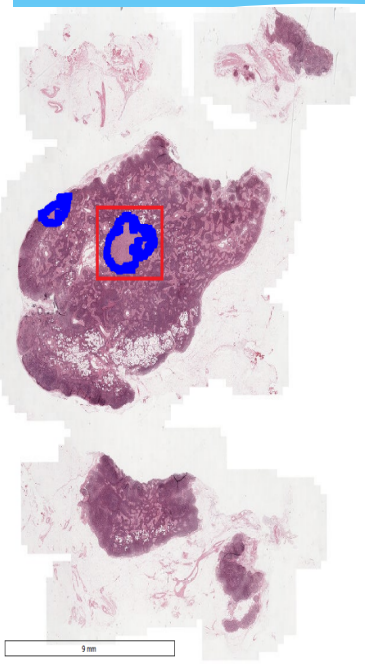
* 100 patients for training data and 100 patients for testing data. Each patients have 5 whole slides images. Some of the patients in training group have pathologists annotations.

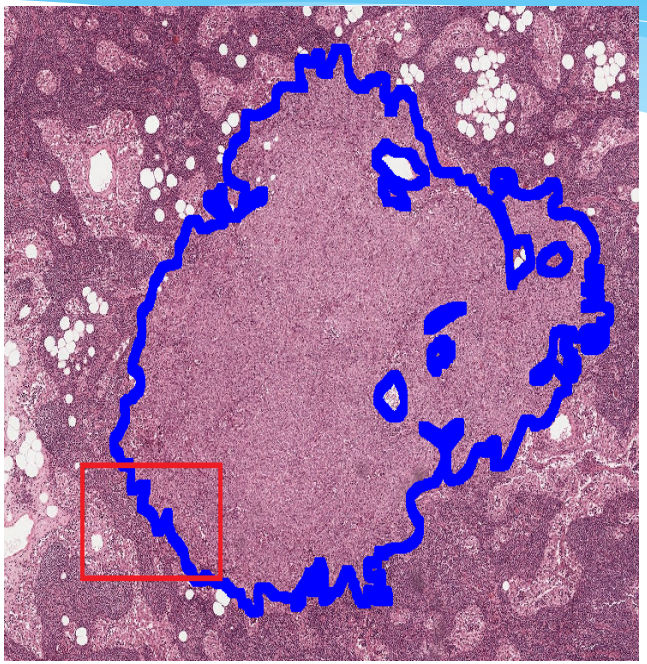
* Each whole slides image is in tiff format and about 2G in size. Approx. 100,000x100,000 pixels.

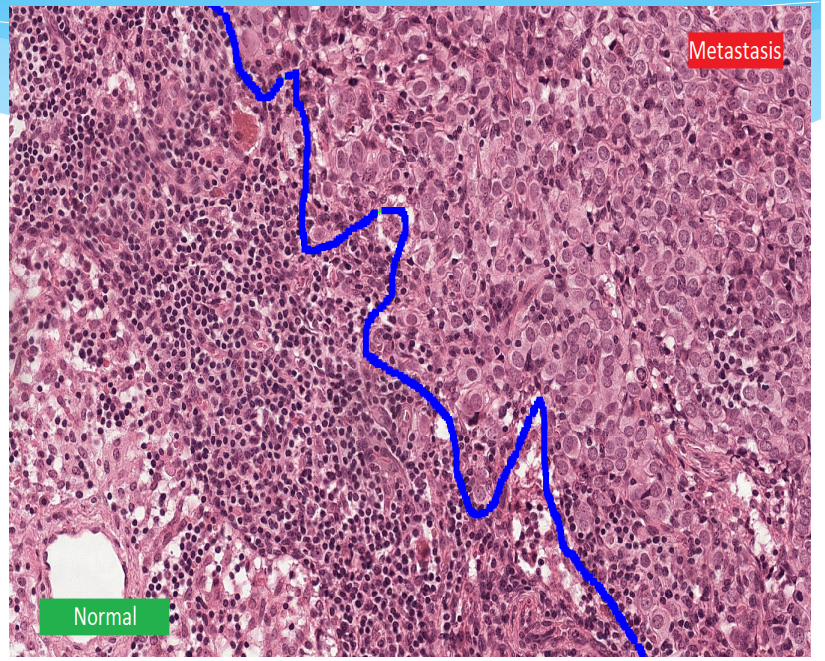
* 20 patients have pathologists' annotations for lesions(ITC,Micro- and Macro-Metastasis). The information is stored in the xml data. We extract all such regions as positive training data.



Example of a metastatic region

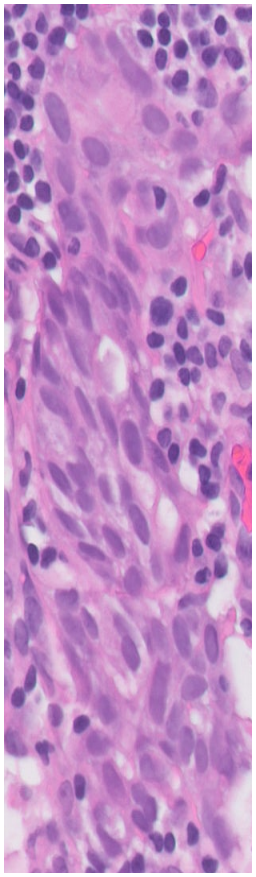


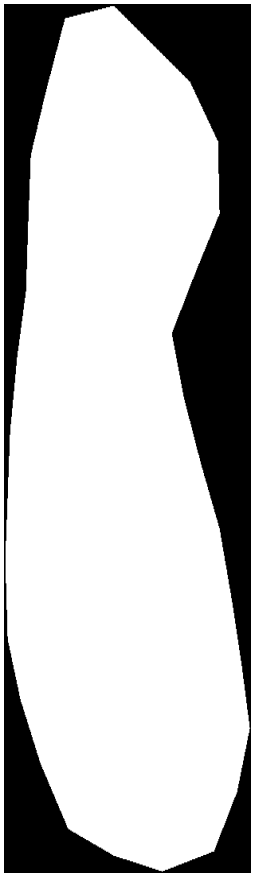




Isolated tumour cell clusters (<0.2 mm or <200 cancer cells in one section,ITC)

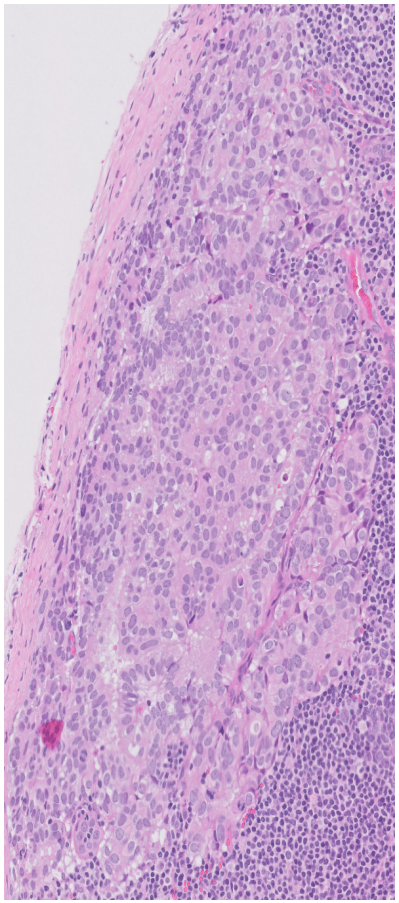
* 310x819

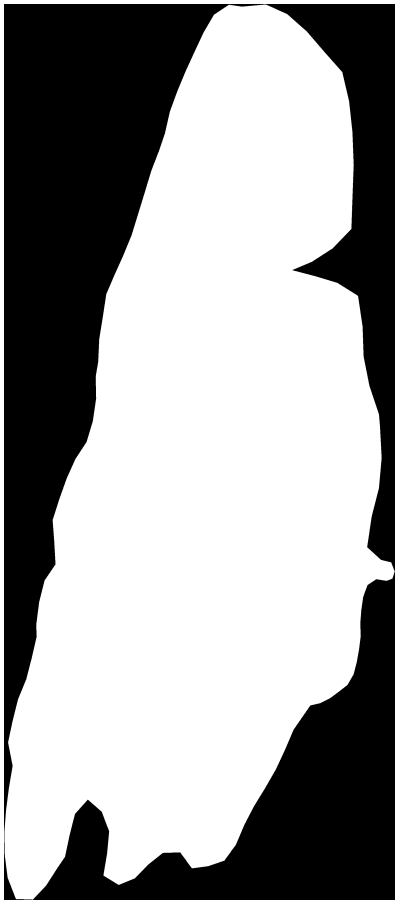


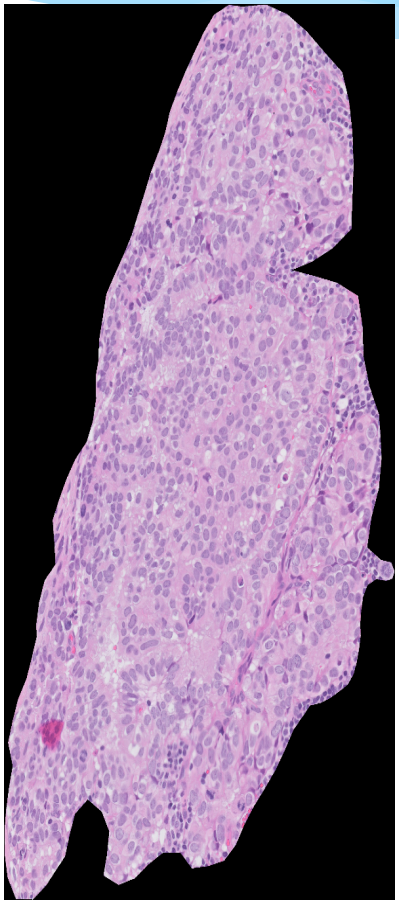


Micrometastases (0.2-2 mm)

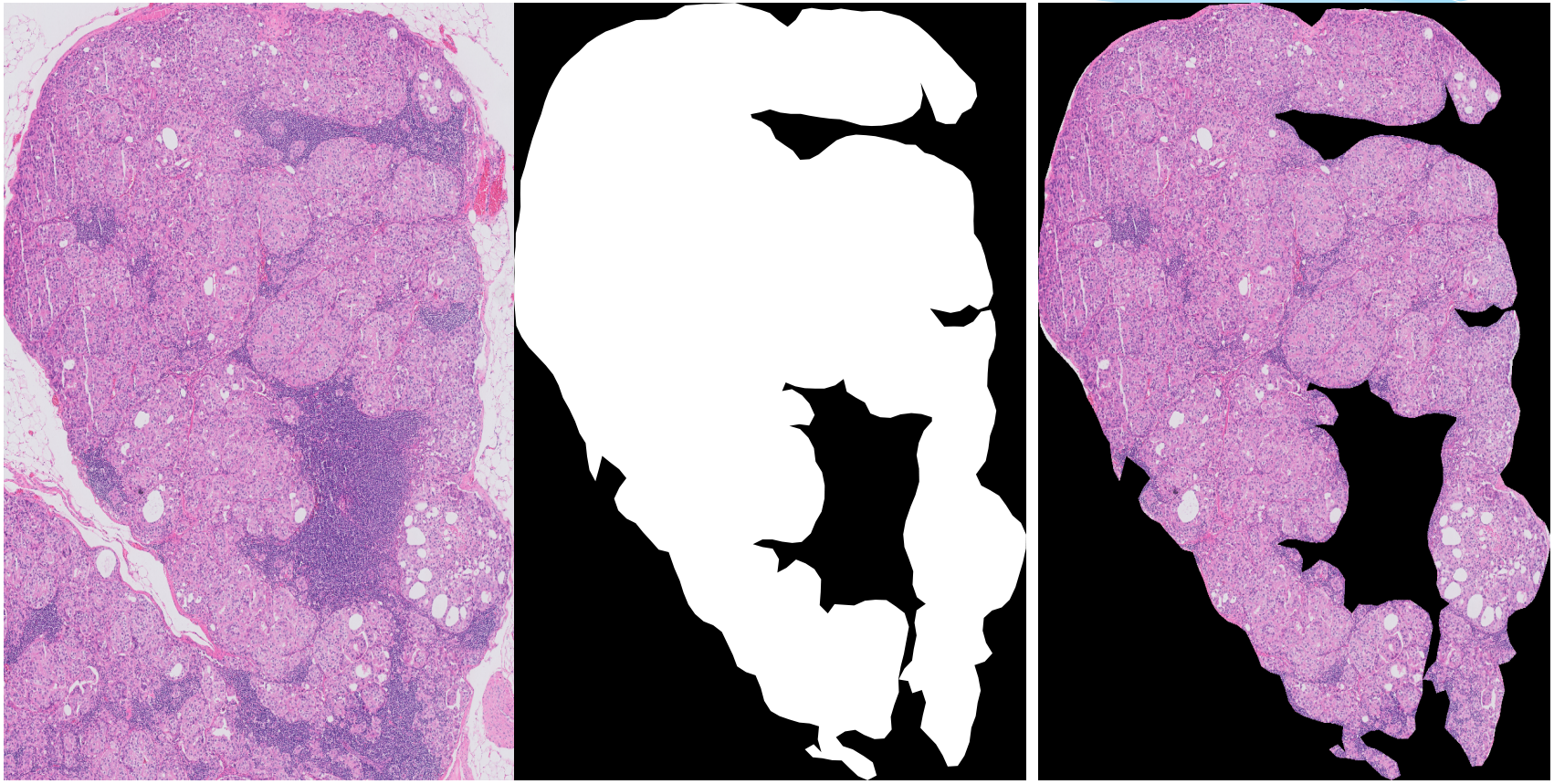
* 1887x3244





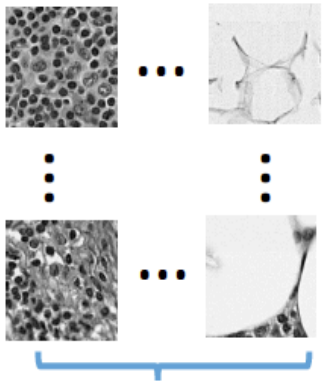


Macrometastases (>2 mm)



Dictionary Learning

Dictionary Learning (DL) approach



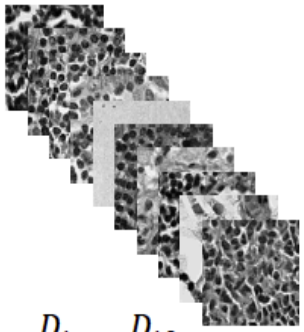
10 groups of
2000 images for normal tissue

$D_{1,1}^{N} \dots D_{1,5}^{N}$

$D_{10,1}^{N} \dots D_{10,5}^{N}$

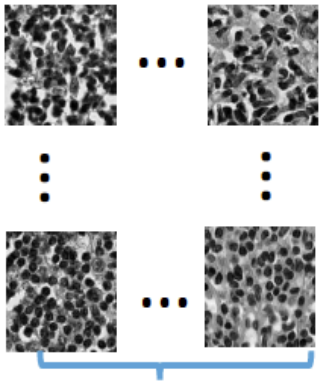
DL : image Y, dictionary D, and coefficient X:

$$\min \left|\left|Y - DX\right|\right|_2^2$$
$$s.t. \left|\left|D\right|\right|_0 < K$$



$D_1, \dots, D_{10}$

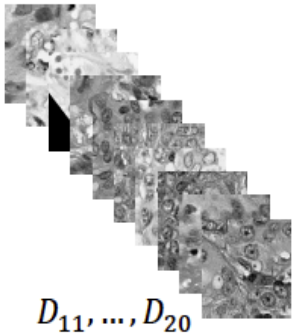
10 common basis for tumor and normal



10groups of 2000 images for tumor tissue

$D_{1,1}^{T} \dots D_{1,5}^{T}$

$D_{10,1}^{T} \dots D_{10,5}^{T}$



$D_{11}, \dots, D_{20}$

Project images to
$\Lambda := \{D_1, \dots, D_{20}\}$
The representation
$X = (\Lambda^T \Lambda)^{-1} \Lambda^T Y$
Is used to train model.

Results:
Accuracy: 0.52
Based on 10 random 50-50 split

Deep Learning Method

Issue with dictionary learning: images are not aligned (not even possible to define aligned)

$Y_{65536\times20000}$

$Z^{(1)}_{65536\times400}$

$H^{(1)}_{400\times20000}$

$Z^{(2)}_{400\times300}$

$H^{(2)}_{300\times20000}$

$\cdots$

$Z^{(N)}_{k_N\times100}$

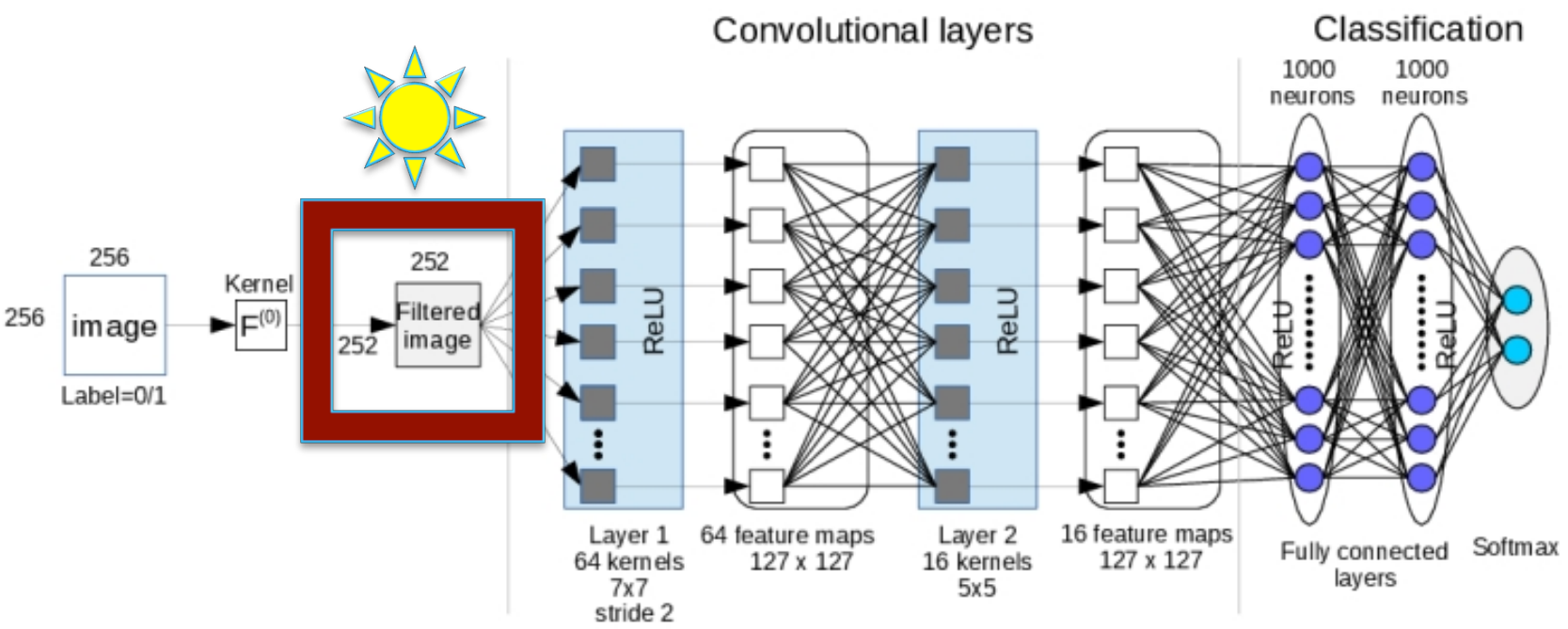
$H^{(N)}_{100\times20000}$

Number of hidden states in each layer

10,000 tumor tissues and
10,000 normal tissues images
(res: 256 × 256)

XGBoost results:
Accuracy 88.64%
AUC: 0.94
10 random 50-50 split

xxNet: Deep Learning Models



http://www.lirmm.fr/~chaumont/SteganalysisWithDeepLearning.html

Deep Learning Methods

Simple problems,

- **DL does not work efficiently.**
- **DL does not work at all.**

Complex problems,

- **DL may work.**
- **Existing statistical (ES) methods do not work efficiently.**
- **ES methods do not work at all.**

Conclusions

* Clinical usefulness is more important
* A useful and powerful statistical method may first come from some deep thinkers.
* A beautiful Mathematical framework is important, but secondary.
* Cost-effectiveness is critical for choosing a good statistical method

https://github.com/BIG-S2

To be continued

A statistical software for handling genetic and medical imaging data