High Dimensional Data Analysis with Applications in IMS and fMRI Processing

Don Hong
Department of Mathematical Sciences
Center for Computational Sciences
Middle Tennessee State University
Murfreesboro, Tennessee, USA
http://capone.mtsu.edu/dhong

1st Annual Workshop on Data Sciences (AWDS 2015), Tennessee State University Nashville, TN. April 16-17, 2015

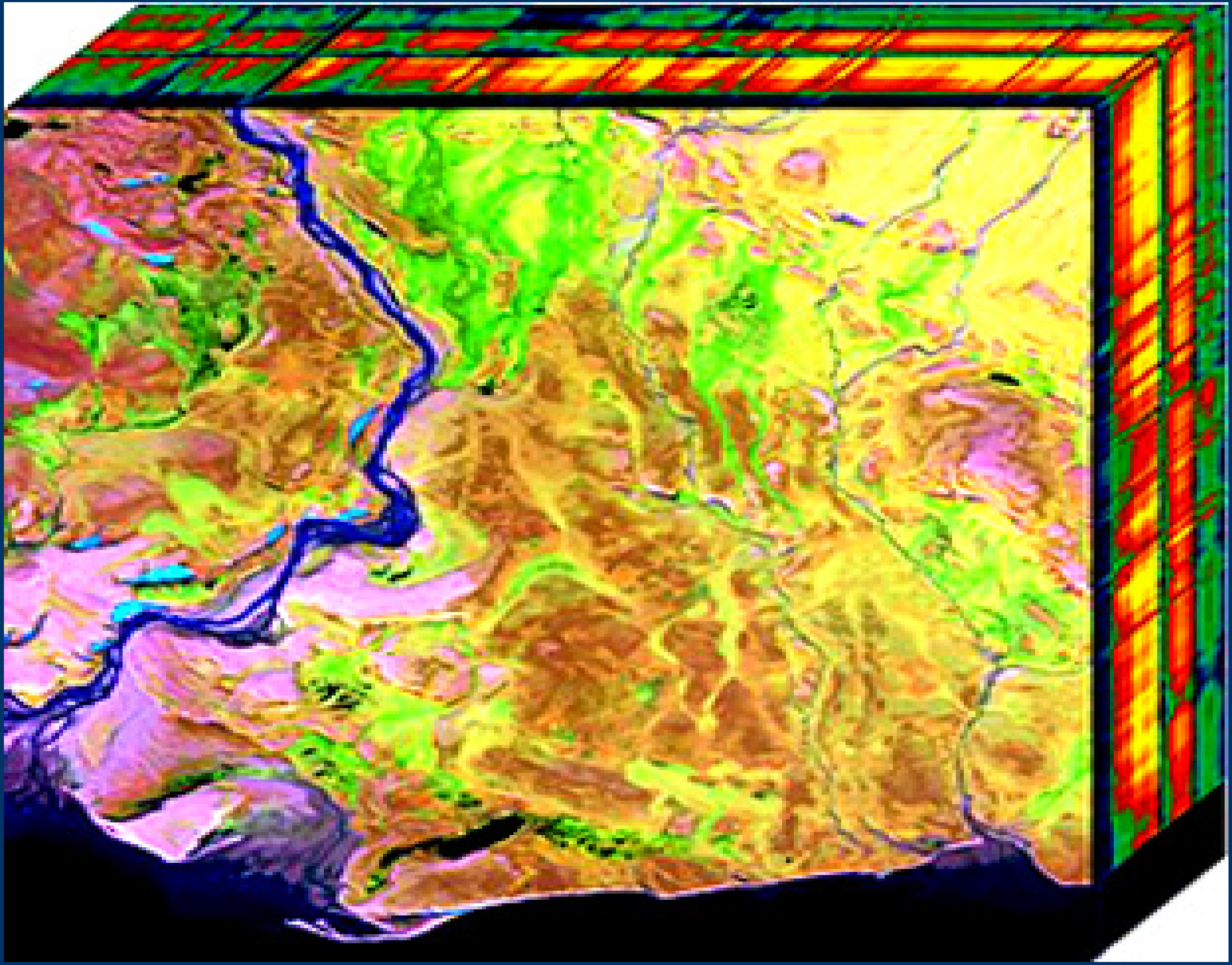
Outline

- Hyper-spectral imaging type data: IMS and fMRI
- Challenges in HSI type data processing
- Software package for IMS: IMSmining
- Markov random fields for IMS and Fusion IMS & Micoscopy
- GLM, probabilistic group ICA and Multi-task learning for fMRI data analysis

HSI Data

- Hyperspectral images: images produced by imaging spectrometers; unlike conventional color cameras that only capture light in just RGB three spectral windows, hyperspectral cameras can capture an entire section of the electromagnetic spectrum at every pixel and collect high resolution spectral detail over a large spatial and broad wavelength region.

Hyperspectral Cube

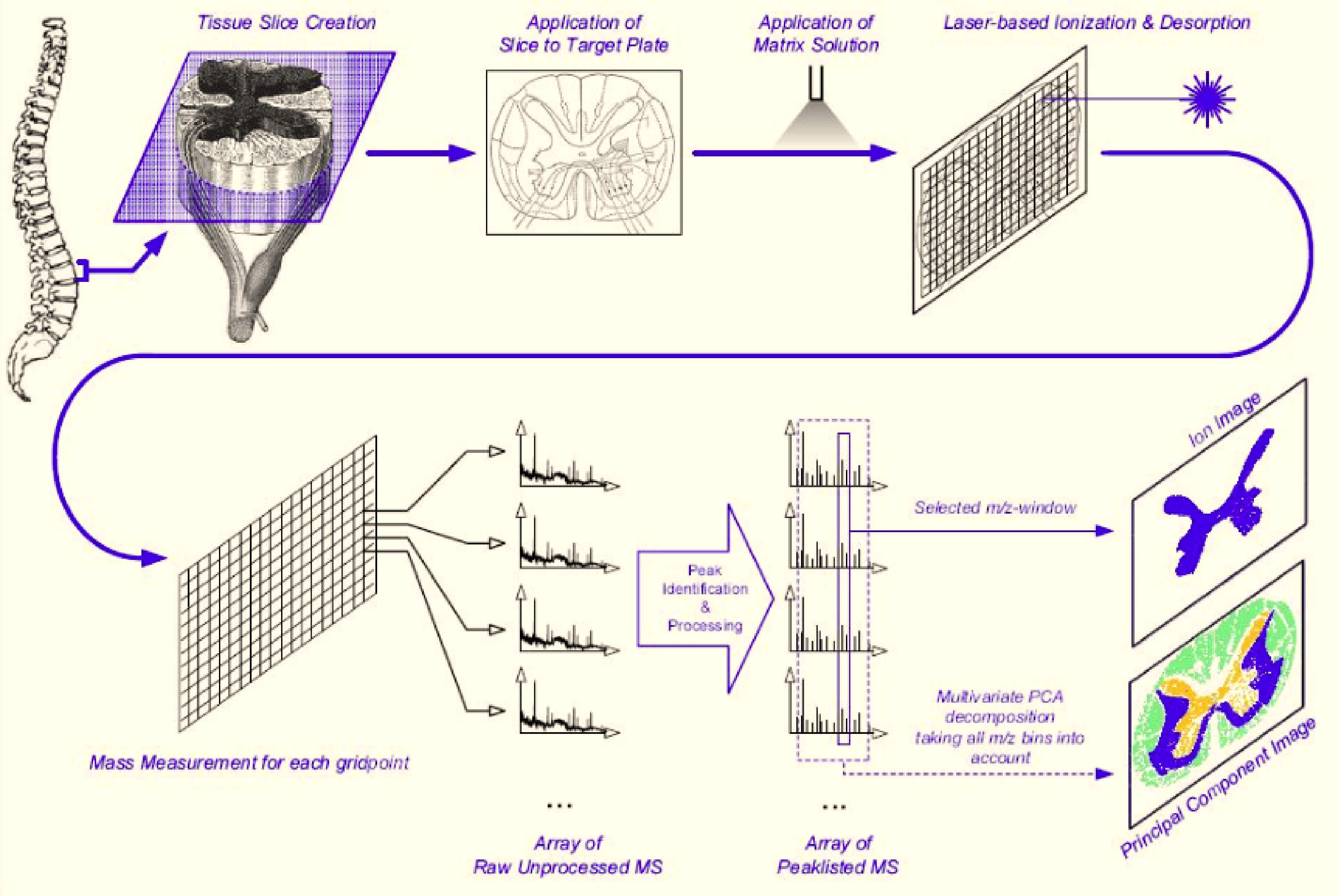


- http://en.wikipedia.org/wiki/Hyperspectral_imaging

Hyperspectral Imaging (HSI) Type: IMS and fMRI

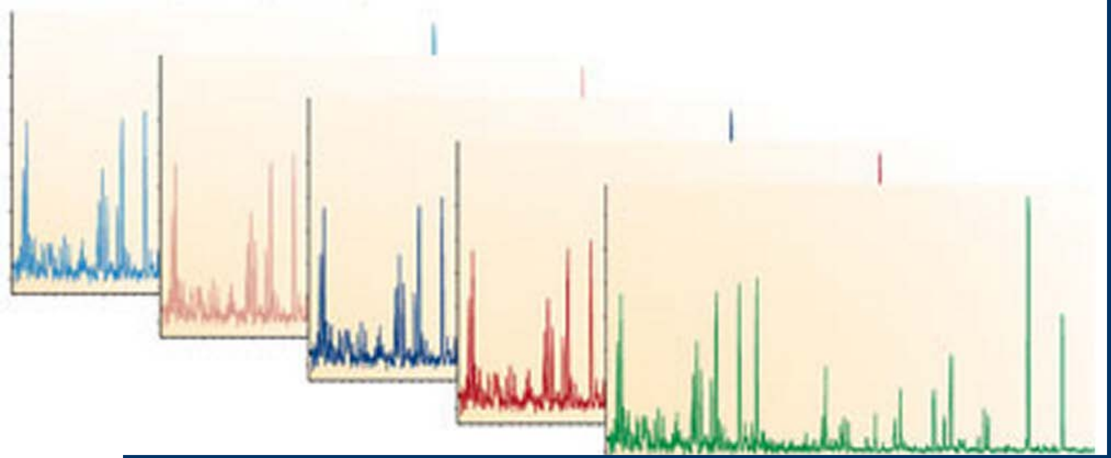
- Data Cubes
- Spectral
- Spatial
- Medical/biological specifics

Example-1: Imaging Mass Spectrometry (IMS) Data


Tissue Slice Creation
Application of Slice to Target Plate
Application of Matrix Solution
Laser-based Ionization & Desorption
Mass Measurement for each gridpoint
Peak Identification & Processing
Selected m/z-window
Ion Image
Multivariate PCA decomposition taking all m/z bins into account
Principal Component Image
...
Array of Raw Unprocessed MS
...
Array of Peaklisted MS

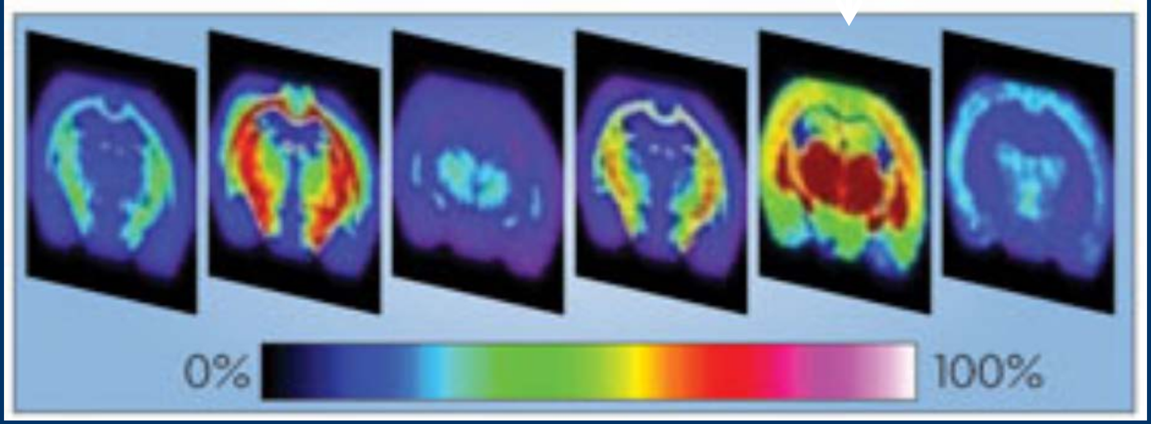
Plas R. V. et al. 2007. Imaging Mass Spectrometry Based Exploration of Biochemical Tissue Composition using Peak Intensity Weighted PCA.

MALDI-TOF



Mass spectra for each x,y coordinate

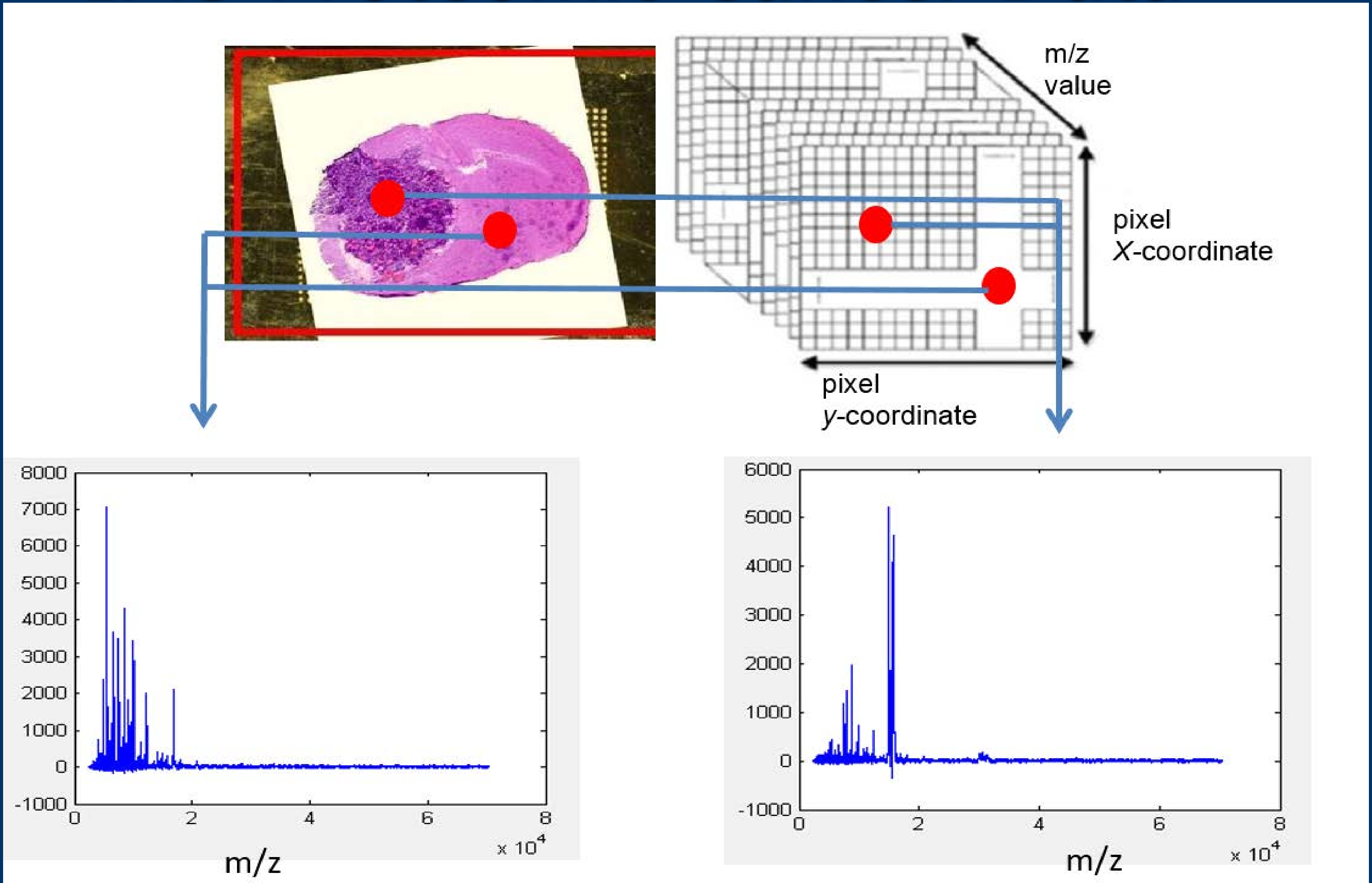
Single m/z value



MALDI-TOF MS technique was awarded the **2002 Nobel Prize in Chemistry**

http://www.mc.vanderbilt.edu/root/vumc.php?site=ims&doc=35274

IMS Proteomic Cancer Data



Zhang and Hong, Statistics in Medicine , 30 (2011), 753-768

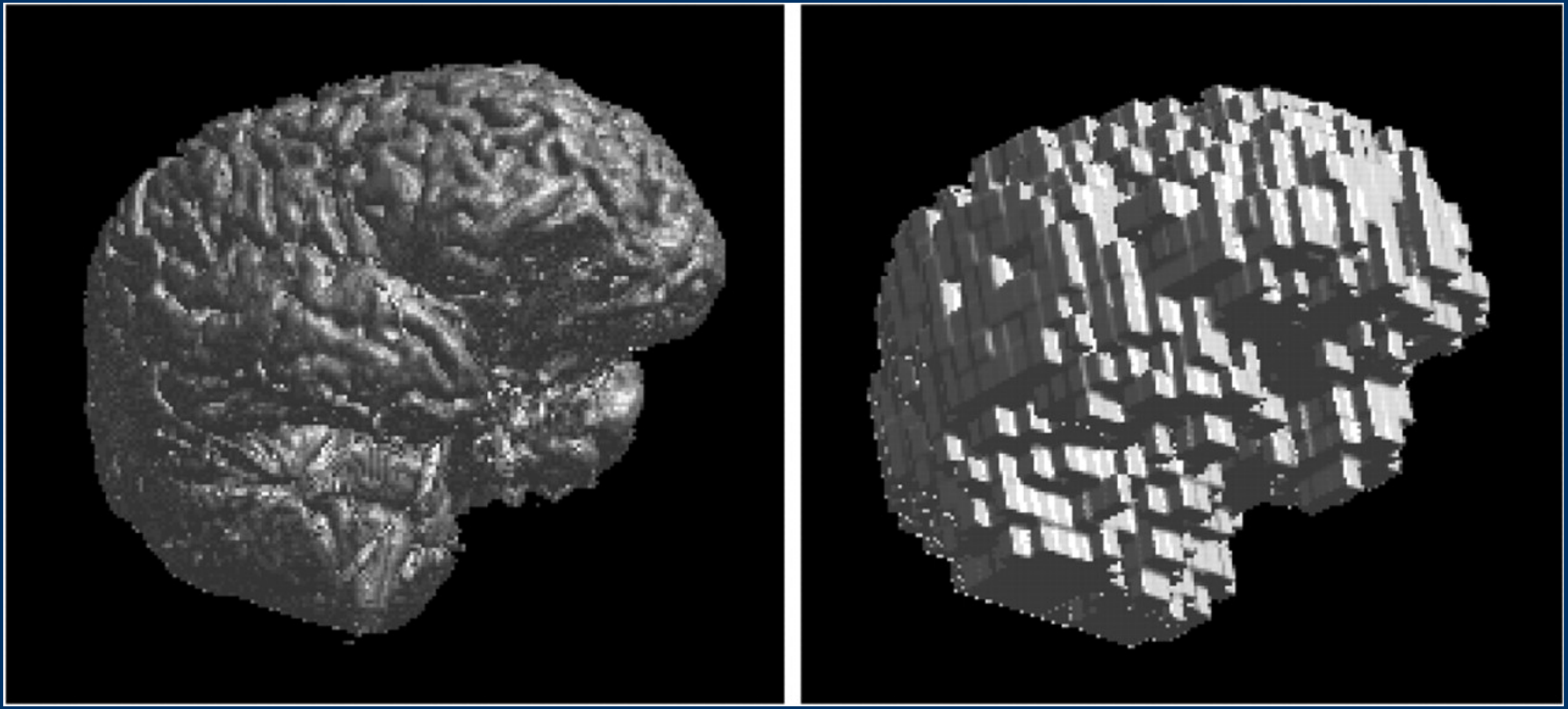
Example-2: MRI→fMRI

- **Magnetic resonance imaging** (**MRI**), different from computed tomography (CT) or X-rays, can visualize internal structures of the body in detail, provide good contrast between the different soft tissues of the body, and thus, become very useful in imaging the brain, muscles, the heart, and cancers.
- Reflecting the fundamental importance and applicability of MRI in medicine, Paul Lauterbur of the University of Illinois at Urbana-Champaign and Sir Peter Mansfield of the University of Nottingham were awarded the 2003 Nobel Prize in Medicine for their "discoveries concerning magnetic resonance imaging".

fMRI

- **Functional magnetic resonance imaging** or **functional MRI** (**fMRI**) is an MRI procedure that measures brain activity by detecting associated changes in blood flow.
- The main principle of fMRI is using the difference of blood oxygenation level dependent (BOLD) to observe the local changes of deoxyhemoglobin concentration in the brain vasculature.
- Then, fMRI data can be measured at voxel by voxel, with a voxel representing a three dimensional cube unit inside the whole brain.

The fMRI Data Cube is of HSI Type



- [http://bjr.birjournals.org/content/77/suppl 2/S167/F1.large.jpg]

Important Characteristics and Challenges in HSI Type Data Processing

- Data sets are very high dimensional
- Many challenges in data processing and data analysis: data preprocessing, dimension reduction, spatial information incorporation for feature selection and pattern recognition
- Requiring interdisciplinary collaboration
- Advanced mathematical tools and statistical techniques can not only provide significance analysis of experimental data sets but also can help in finding new data features/patterns, guiding biological experiments designs, as well as leading computational tools development.

Important Characteristics and Challenges in HSI Type Medical Data Processing

- Different from typical image processing emphasis, in HSI type medical data processing, we concentrate on the following 4Cs
- Content of biology
- Contrast between groups such as normal data and cancer data
- Characteristic/feature/biomarker discovery
- Classification accuracy improvement

I. MS/IMS Data Processing

- MS Data Preprocessing

  with Shuo Chen and Yu Shyr

- IMS Data Biomarker Discovery and Classification

  with FQ Zhang and Lu Xiong

- IMSmining Software Development

  with JS Liang, FQ Zhang, and JC Zou

IMSmining Software Package

IMSmining is a newly developed software package by our research group for IMS data processing using statistical methods. It contains the following functions:

1) Data visualization

User can view the spectrum of a single pixel, the average spectrum of an area and intensity distribution matrix at a fixed m/z value.

2) Biomarkers selection and classification.

Statistical algorithm selections include PCA+SVM/LDA, EN/WEN, SPCA+SVM/LDA.

- The graphical user interface is very friendly and convenient. Users can directly select an area of pixels from the ion image as the training data sets, choose the data sets for testing or classification from folders, and export data sets, mean spectrum, selected m/z list for later usage.
- Another great convenience is the mutual response between the ion image and the spectrum figure.

IMSmining

Work flow:

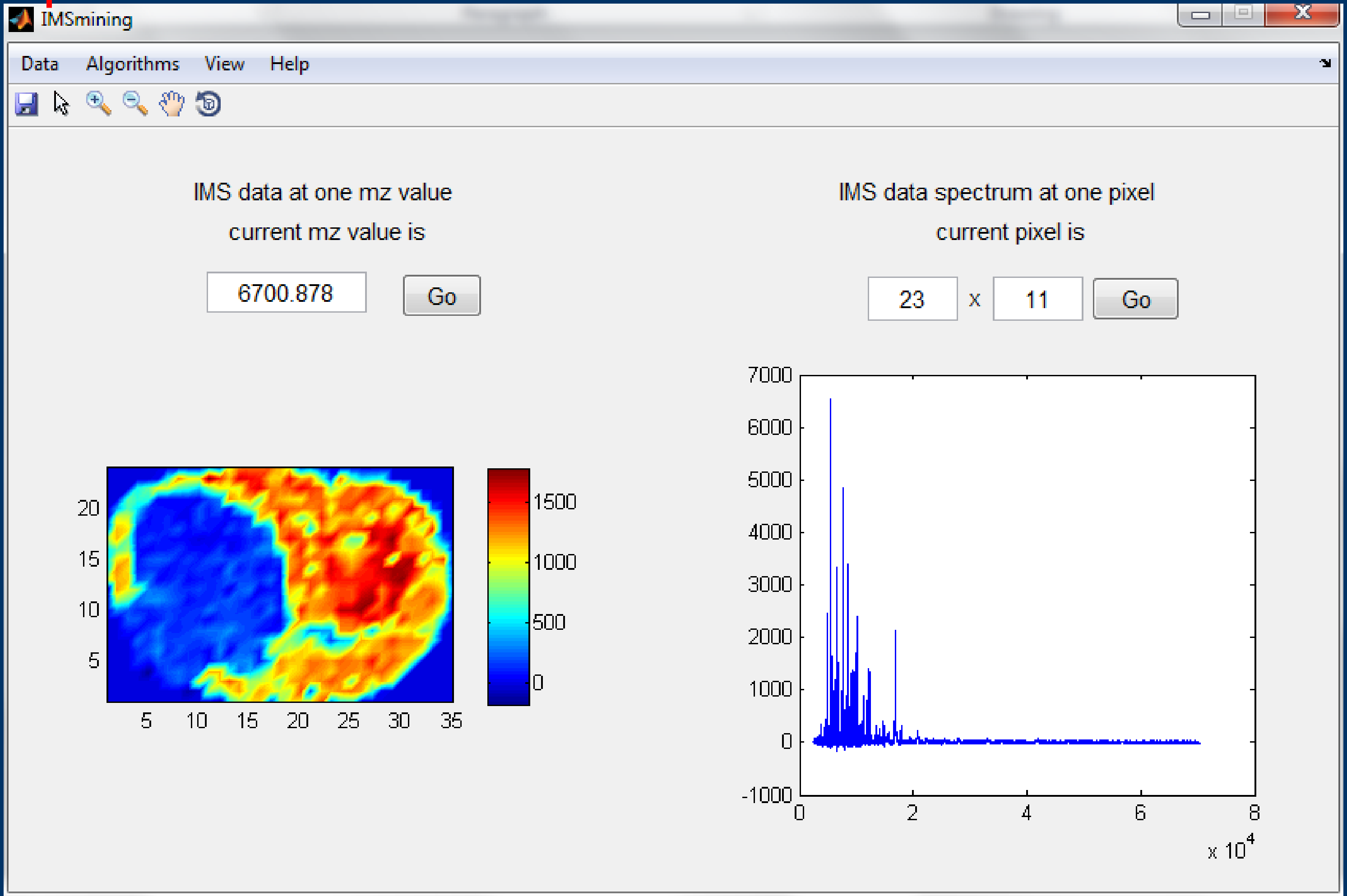
1) Import the training data and select one algorithm.
2) Use mouse to drag square areas of cancer and noncancer.
3) Import the testing data.

Results:

1) Biomarker.
2) Visualize testing data and classification results.
3) Accuracy and Sensitivity.

IMSmining

Import the data from .mat file or .txt folder. Export the biomarker.



IMS data at one mz value

current mz value is

6700.878

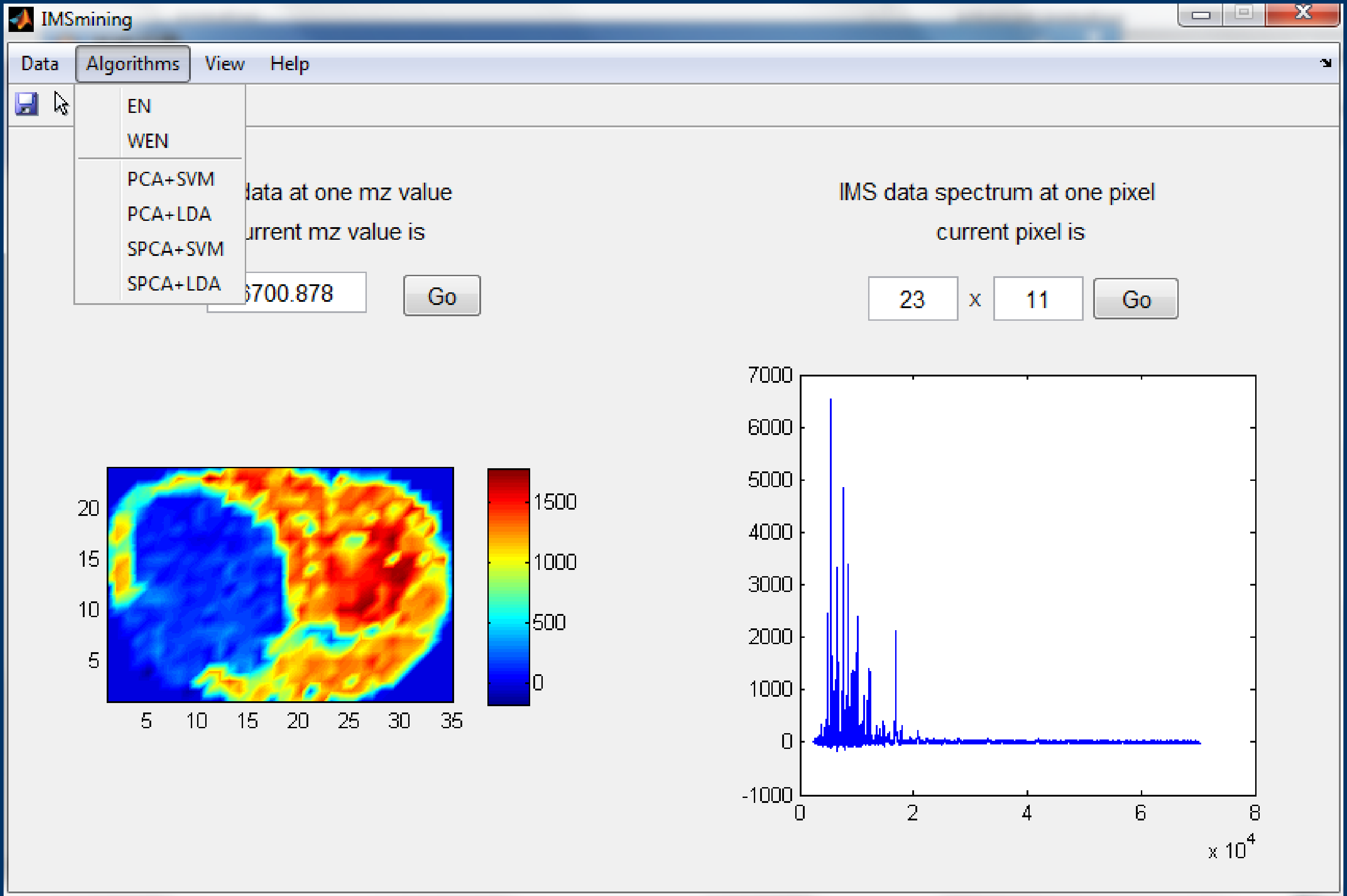
IMS data spectrum at one pixel

current pixel is

23 x 11

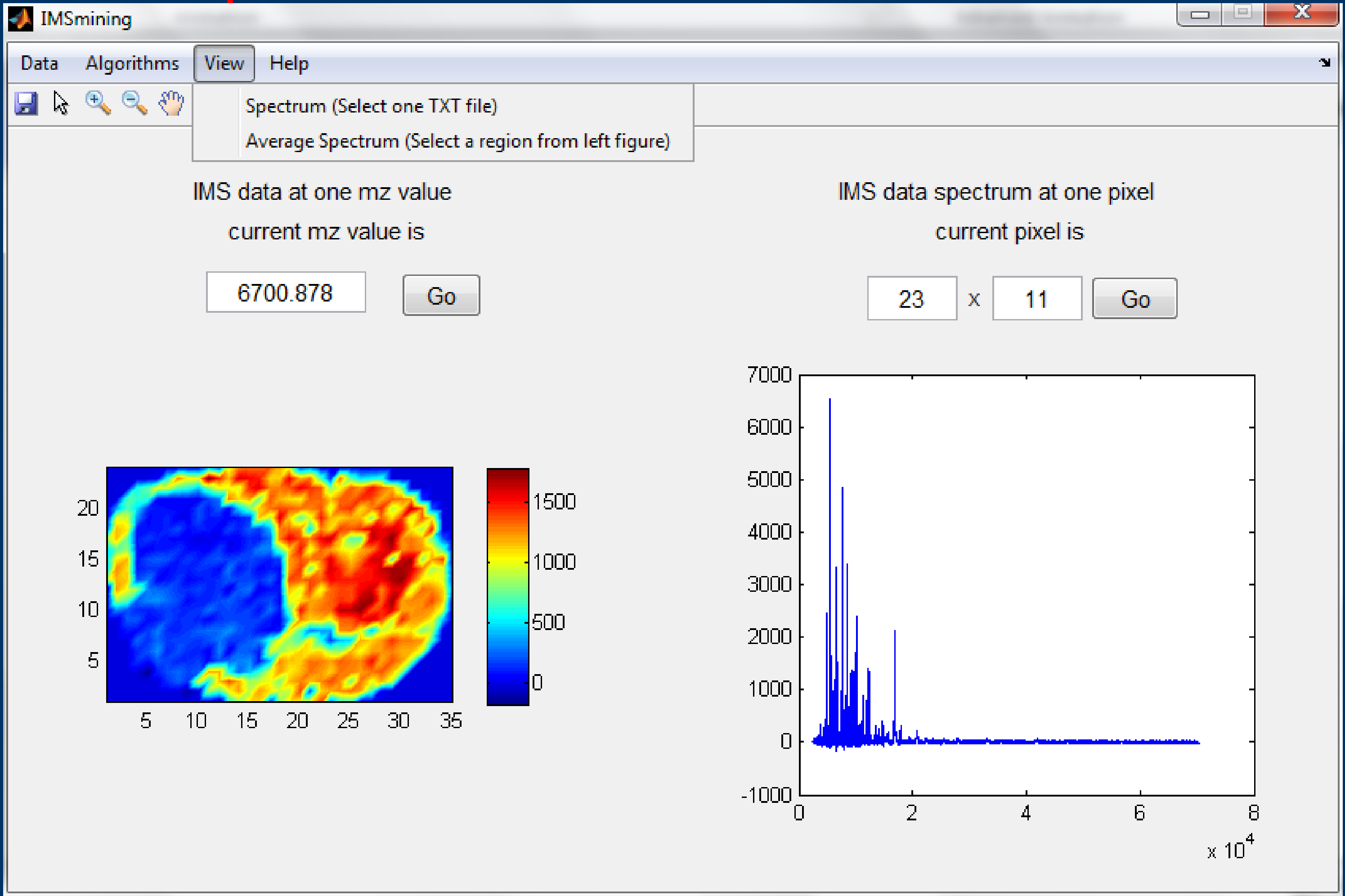
IMSmining

Execute the algorithms including PCA, SVM, EN, WEN, SPCA.



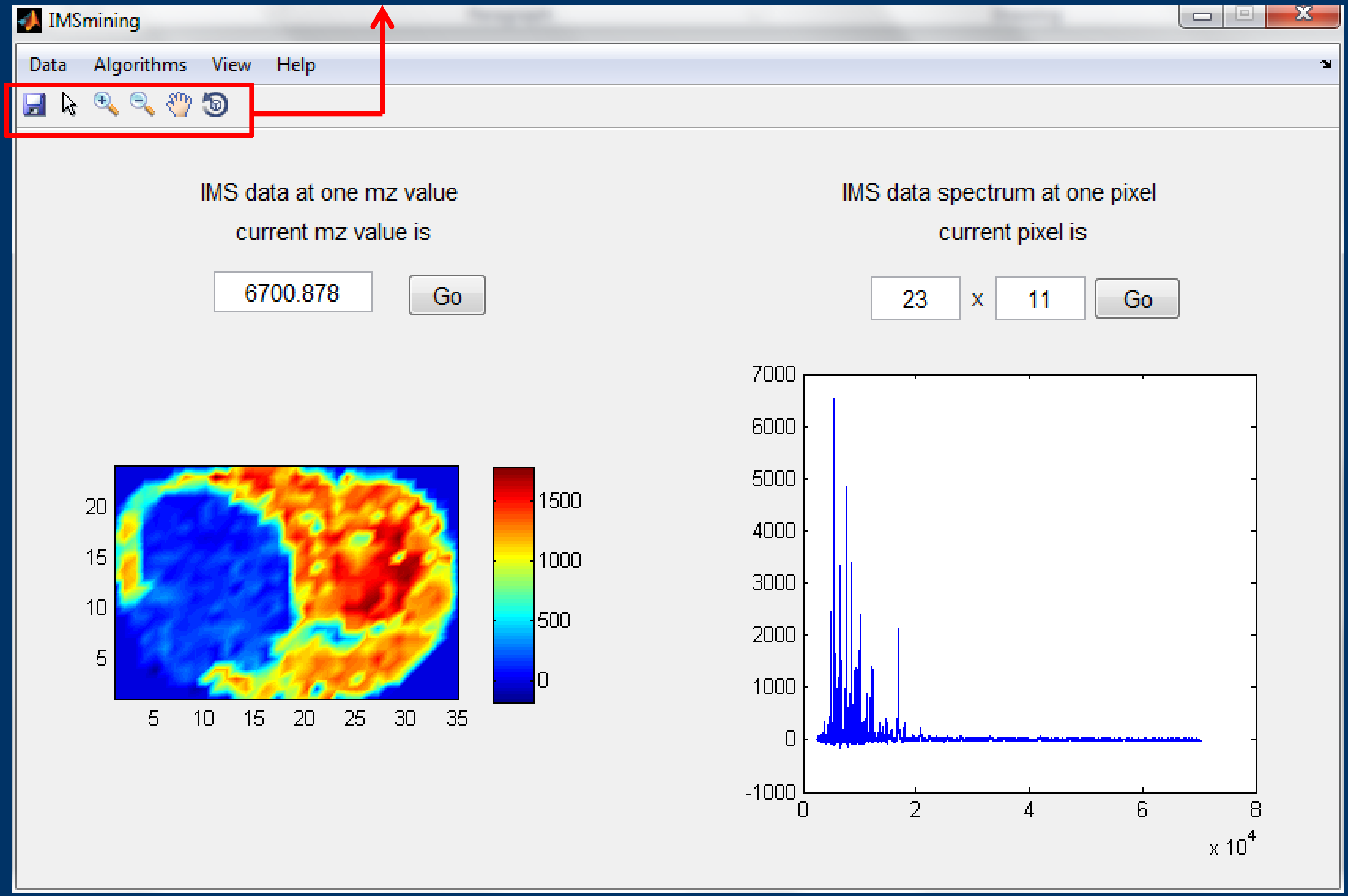
IMSmining

View the spectrum of a single pixel or the average spectrum of an area.



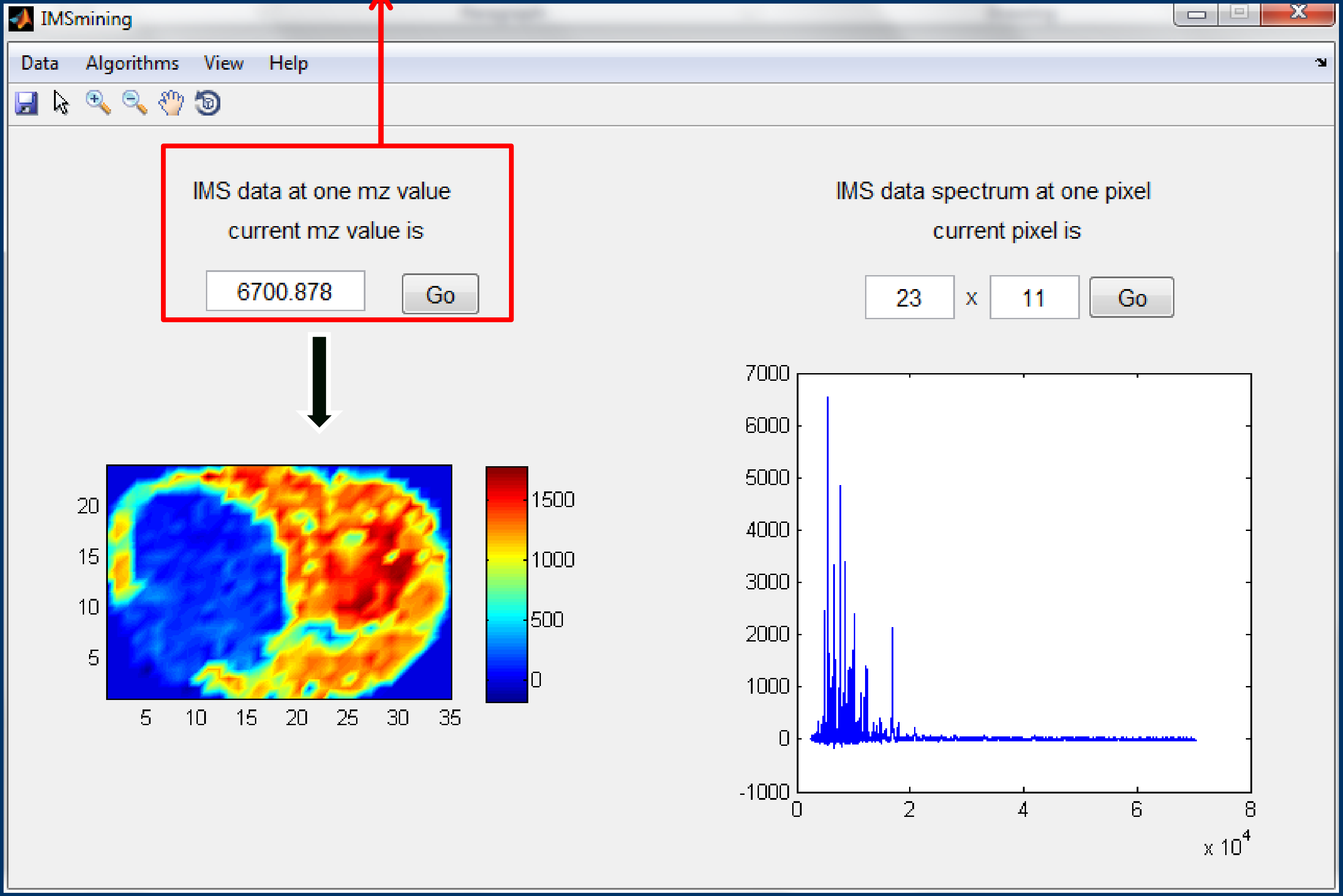
IMSmining

(Toolbar) Save, Zoom in, Zoom out, Drag, and Rotate.



IMSmining

View and intensity distribution matrix at a fixed m/z value.



IMS data at one mz value

current mz value is

6700.878

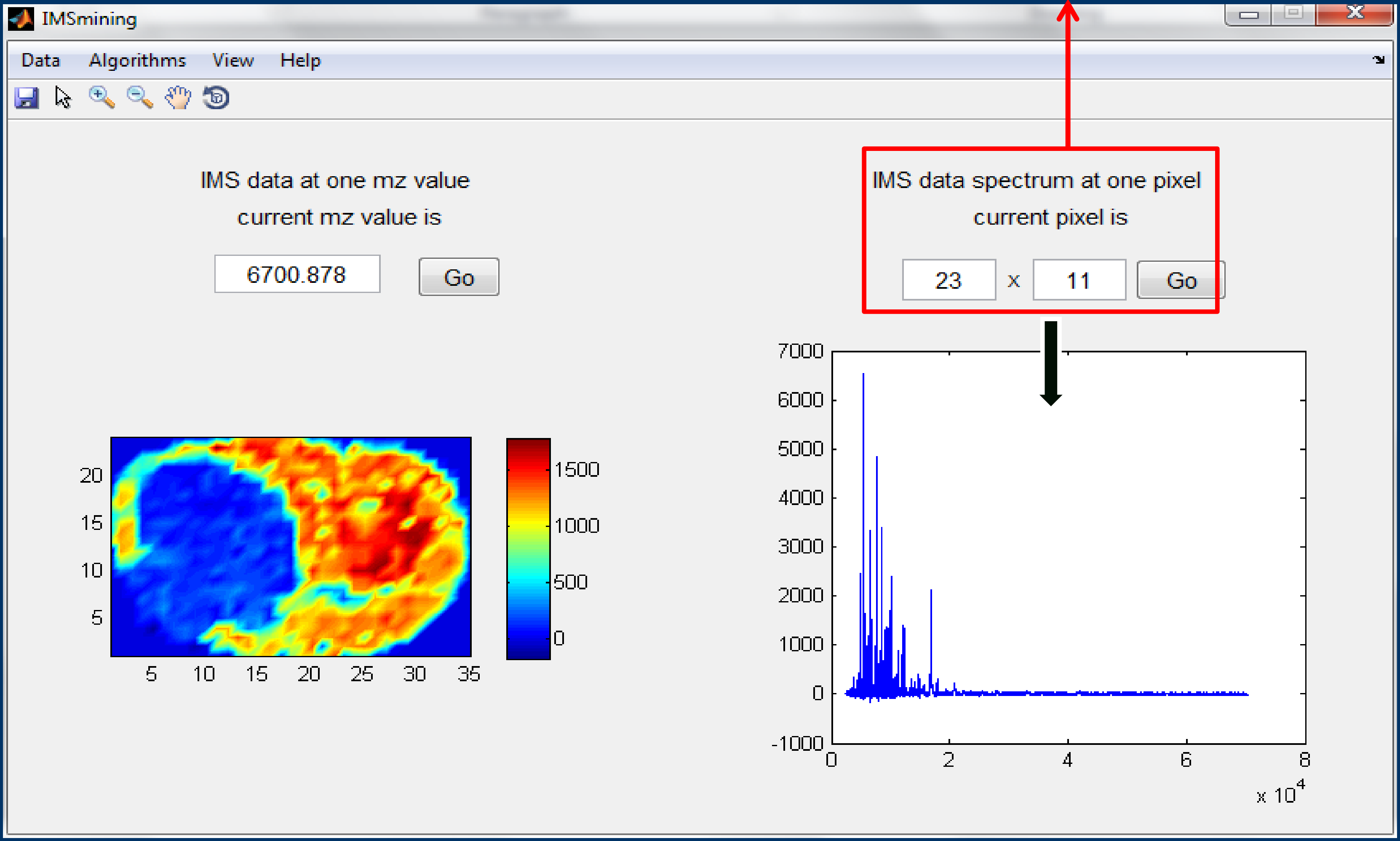
IMS data spectrum at one pixel

current pixel is

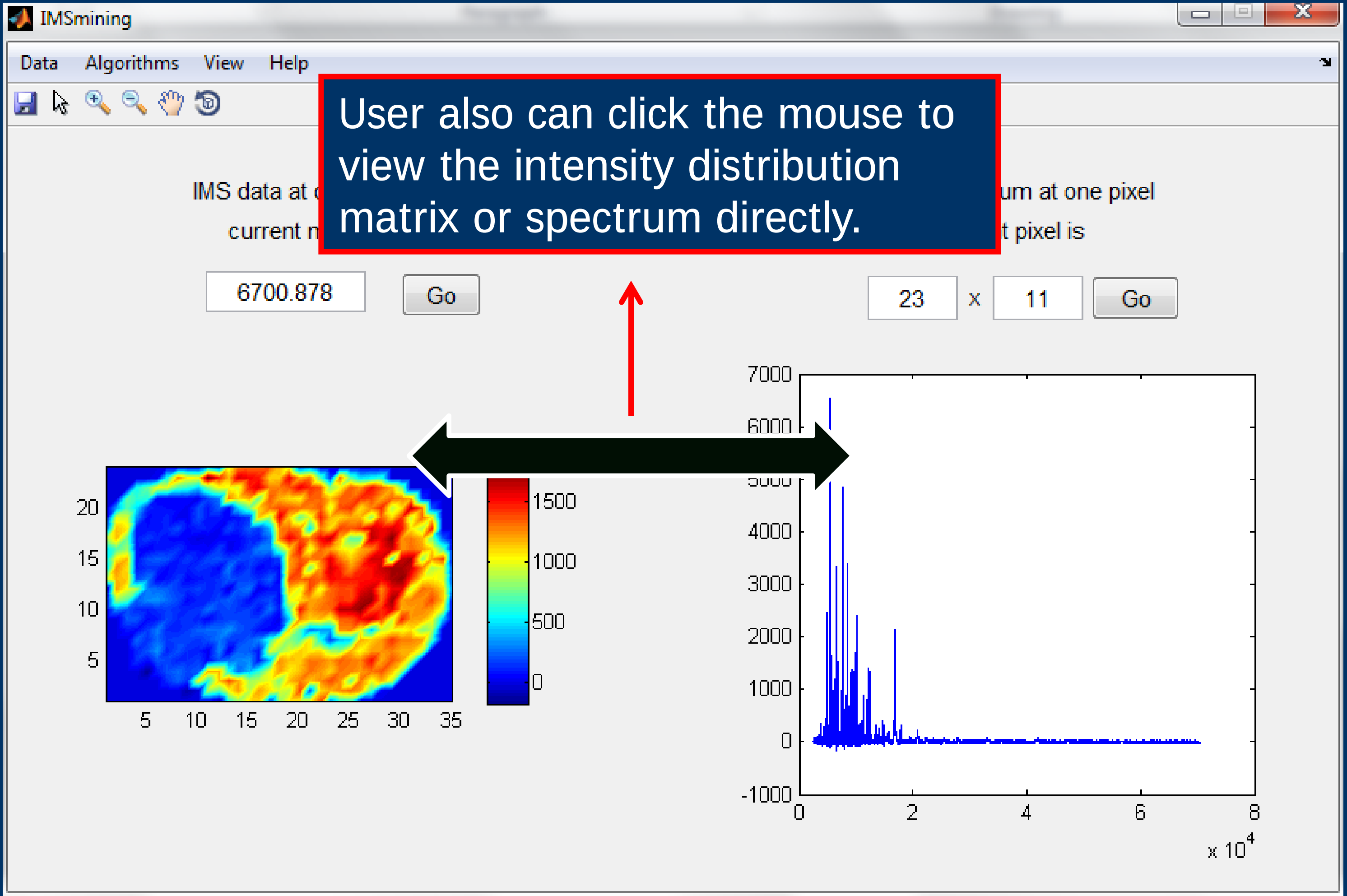
23 x 11

IMSmining

View the spectrum at a fixed pixel.

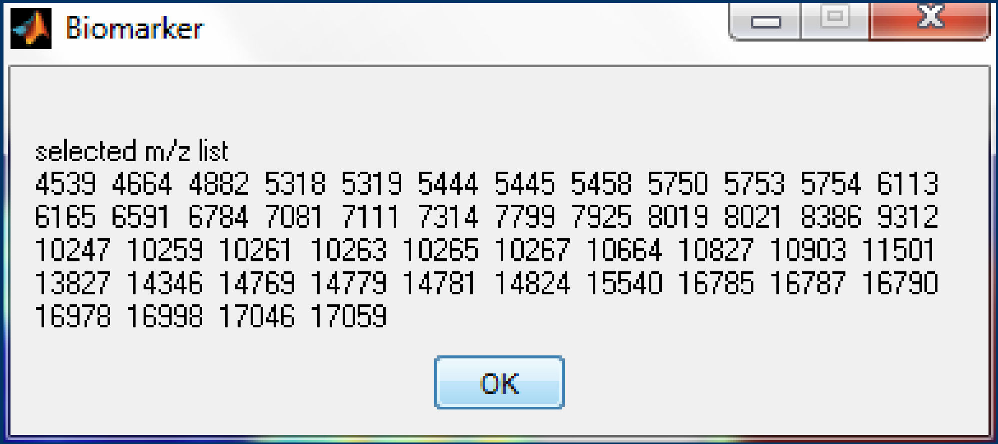


IMSmining

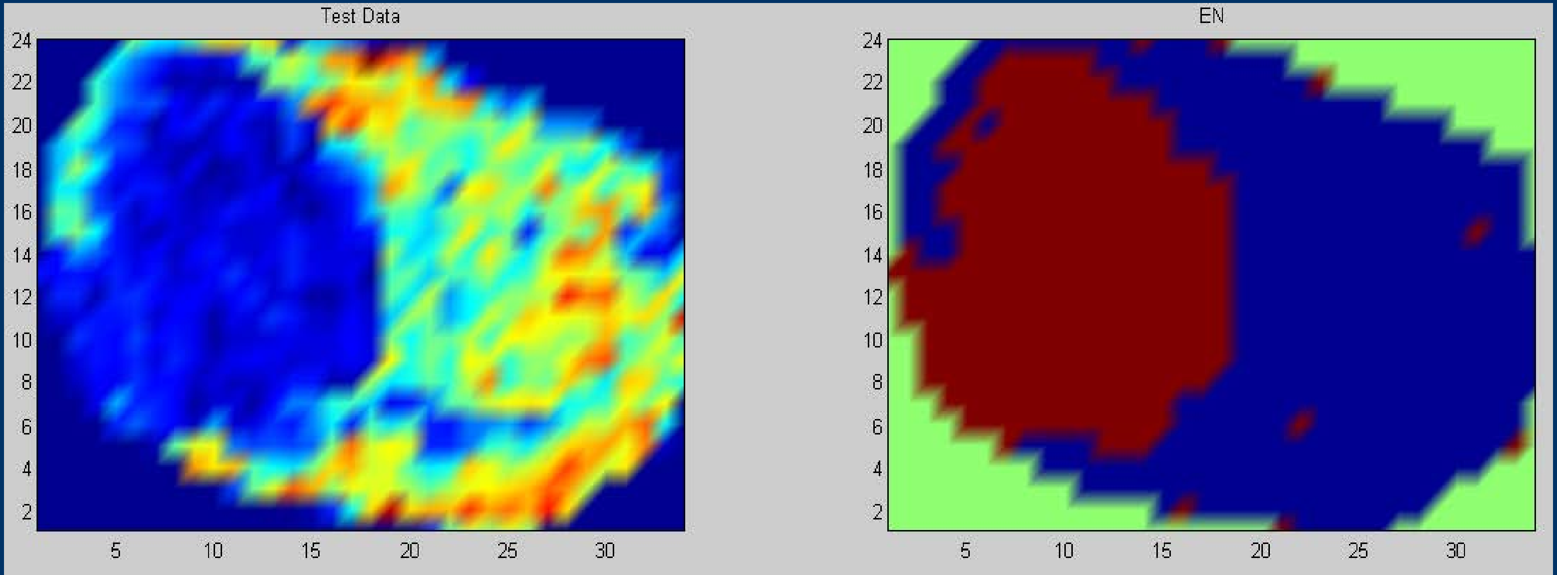


User also can click the mouse to view the intensity distribution matrix or spectrum directly.

EN & WEN



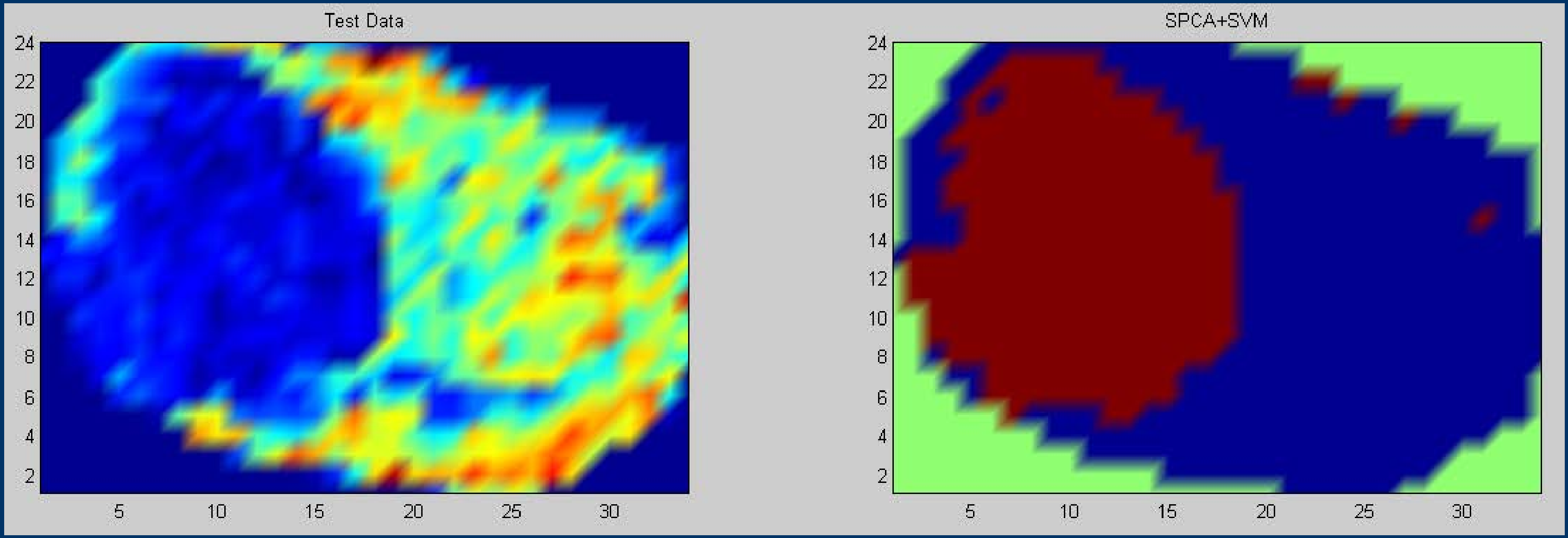
```
selected m/z list
4539  4664  4882  5318  5319  5444  5445  5458  5750  5753  5754  6113
6165  6591  6784  7081  7111  7314  7799  7925  8019  8021  8386  9312
10247  10259  10261  10263  10265  10267  10664  10827  10903  11501
13827  14346  14769  14779  14781  14824  15540  16785  16787  16790
16978  16998  17046  17059
```



SPCA+SVM/LDA

However, by using PCA, each principal component is a linear combination of all the original variables. SPCA uses a regression-style of PCA and adds a lasso term to produce sparse coefficients.

We also include SPCA for IMS processing.



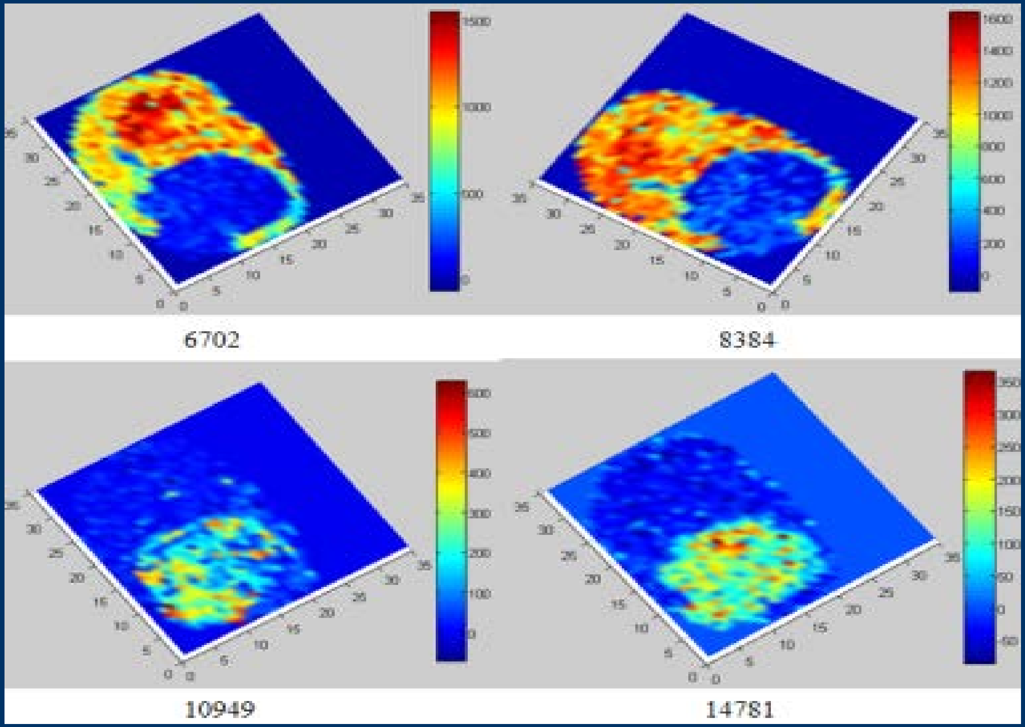
Results

Methods	Accuracy	Sensitivity	Time(s)
PCA+SVM	84.88%	88.23%	16.49
EN/WEN	94.02%	90.04%	29.56
SPCA+SVM	95.43%	93.75%	14.67

- PCA+SVM/LDA is the worst one.
- EN and WEN perform better and are almost the same based on this dataset.
- SPCA+SVM/LDA is the best one considering results and time consuming.

Help in Discovery of New Biomarkers:

EN4IMS/WEN models found more important biomarkers compared to PCA method . The peaks (*m/z* 6700, 8380, 10952, 14788) known as cancer biomarkers are on EN4IMS/WEN list but not on PCA list



6702

8384

10949

14781

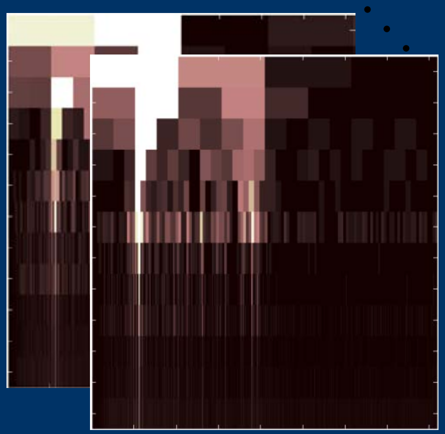
Spatial intensity distribution of 4 selected biomarkers. The x- and y- dimensions are the spatial dimensions and intensities of the selected m/z values are represented by the color

IMSmining also includes MRA for IMS Data Biomarker Selection and Classification

- Basic Idea and Procedure:

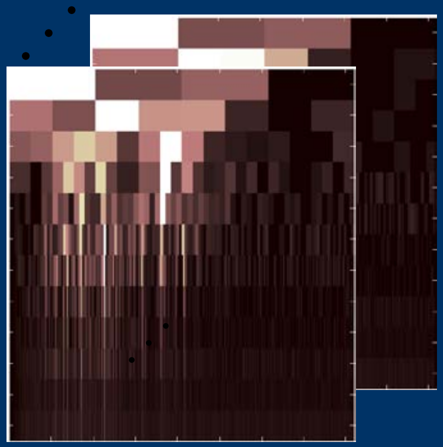
1. Apply wavelet transform to IMS data
2. The idea of wavelet pyramid method for image matching was then applied for biomarker selection, in which Jaccard similarity was used to measure the similarity of wavelet coefficients and determine feature vectors to serve as the biomarkers.
3. Last, the Naive Bayes classifier was used for classification based on feature vectors in terms of wavelet coefficients.

Biomarker Selection using Wavelets

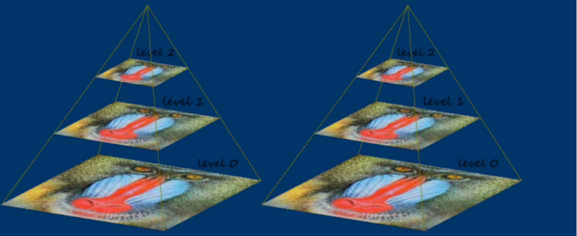


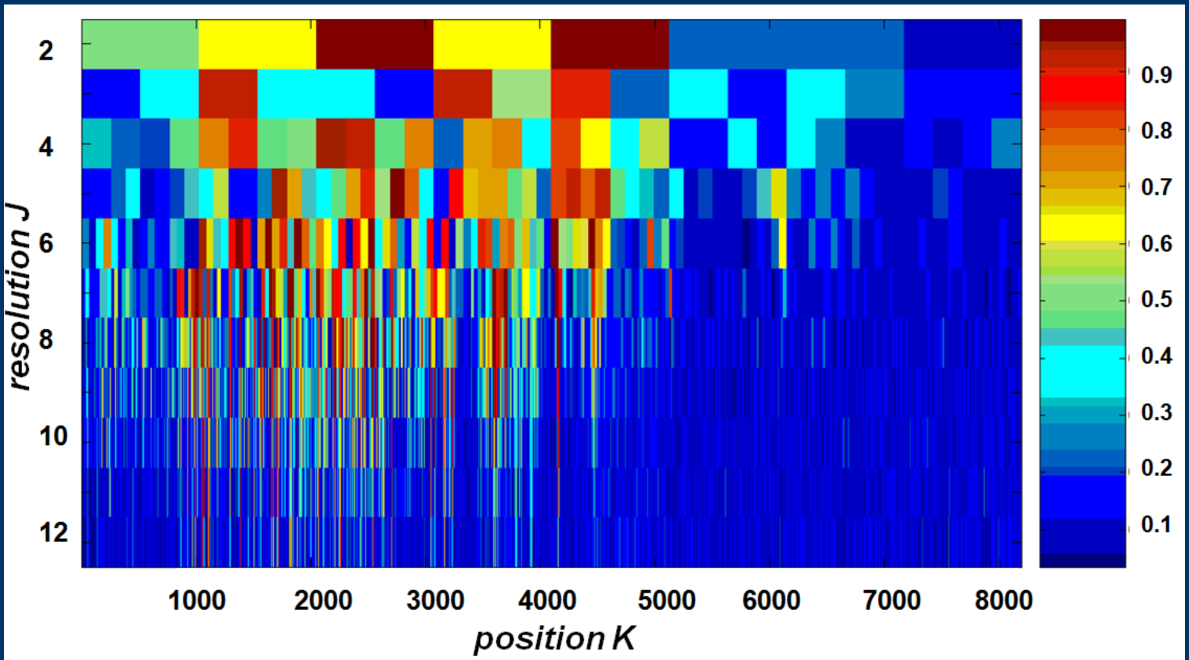
Cancer wavelet coefficients group

Statistically compare using Jaccard similarity



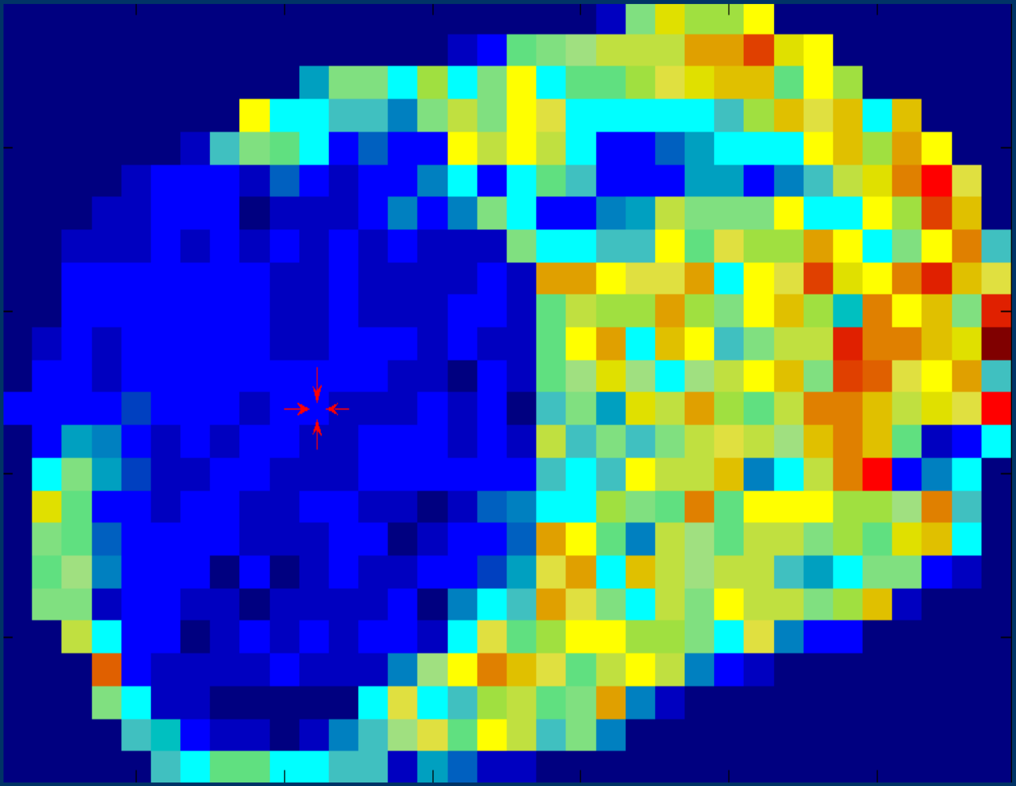
Non-cancer wavelet coefficients group



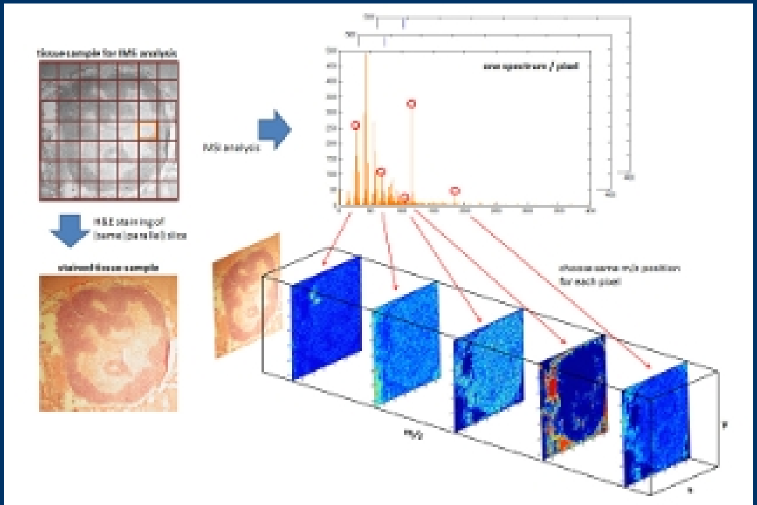


Spatial Information Consideration

The state (cancer or non-cancer) of a pixel is highly determined by the configuration of its neighboring system.

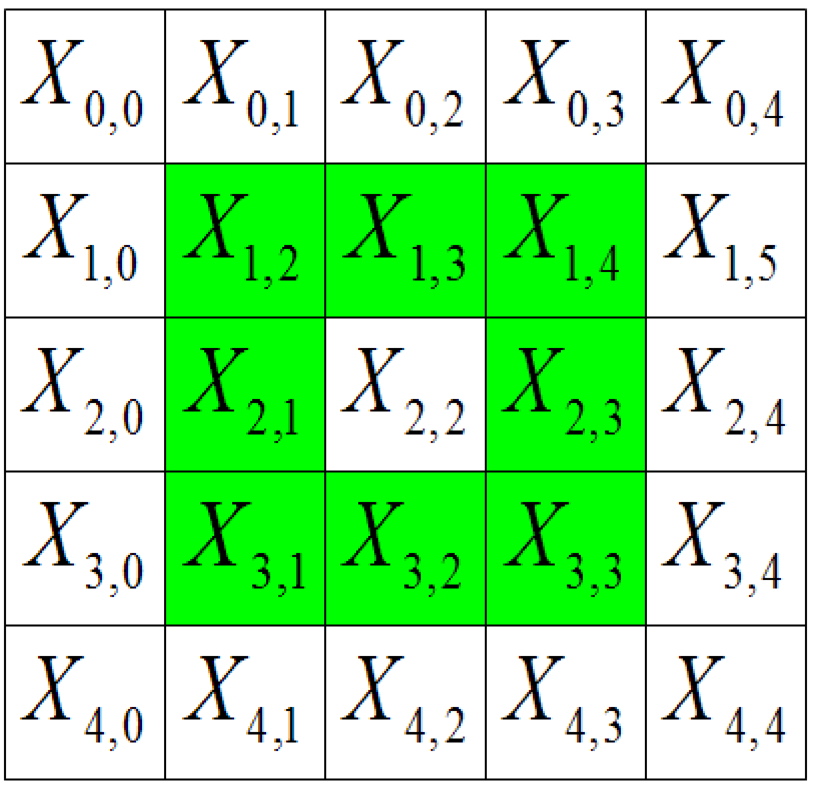


To fully utilize IMS data, it is desirable to not only identify the peaks of the spectrum within individual pixels but also to study correlation and distribution using the spatial information for the entire image cube.


tissue sample for IMS analysis
IMS analysis
one spectrum / pixel
H&E staining of (same/parallel) slice
stained tissue sample
choose same m/z position for each pixel
m/z
y
x

Markov Random Fields

Local property (the Markovianity)



$X_{0,0}$	$X_{0,1}$	$X_{0,2}$	$X_{0,3}$	$X_{0,4}$
$X_{1,0}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$
$X_{2,0}$	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$
$X_{3,0}$	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$
$X_{4,0}$	$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	$X_{4,4}$

Define

S - set of lattice points

s - a lattice point, $s \in S$

X_s - the value of X at s

∂s - the neighboring points of s

$$p(x_s | x_r \text{ for } r \neq s) = p(x_s | x_{\partial r})$$

MCMC-MRF Algorithm for IMS Data Processing

- Lu Xiong and Don Hong, An MCMC-MRF Algorithm for Incorporating Spatial Information in IMS Data Processing, submitted manuscript, 2014.

MS Data Preprocessing Papers

- S. Chen, D. Hong, and Y. Shyr, Wavelet-Based Procedures for Proteomic MS Data Processing, *Computational Statistics and Data Analysis*, **52** (2007), 211-220.
- D. Hong and Y. Shyr, Mathematical Framework and Wavelets Applications in Proteomics for Cancer Study, In: Handbook of Cancer Models With Applications, (Wai-Yuan Tan and Leonid Hannin Eds.), pp. 471-499, World Scientific Publication, Singapore, 2008. ISBN: 981-277-947-7.
- S. Chen, M. Li, D. Hong, D. Billheimer, HM. Li, BG. Xu, and Y. Shyr, A Novel Comprehensive Wave-form MS Data Processing Method, *Bioinformatics*, Vol. 25, no. 6, 2009, 808-814.

IMS Cancer Data Processing Papers

- EN4IMS

  *FQ. Zhang and **D. Hong**, Elastic net-based framework for imaging mass spectrometry data biomarker selection and classification , Statistics in Medicine , 30 (2011), 753-768.*

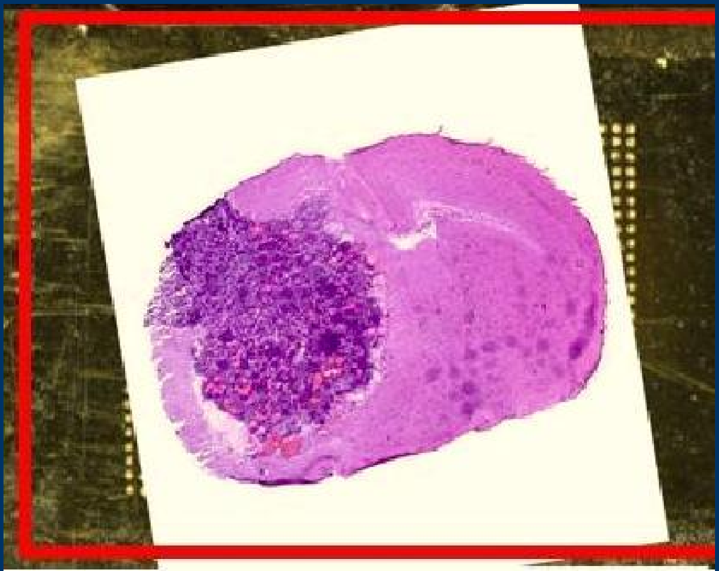
- WEN

  ***D. Hong** and FQ. Zhang, Weighted Elastic Net Model for Mass Spectrometry Imaging Processing, Math. Model. Nat. Phenom. ,Vol. 5, No. 3, 2010, pp. 115-133.*

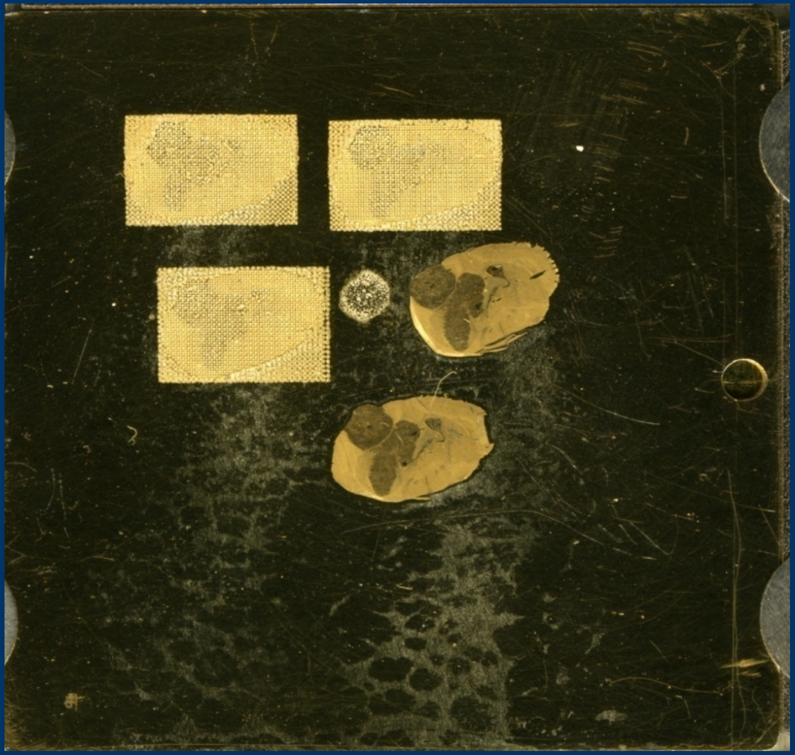
- MRA Based method

  L. Xiong and D. Hong, Multi-Resolution Analysis Method for IMS Proteomic Data Biomarker Selection and Classification, British Journal of Mathematics & Computer Science, 5(1): 65-81, 2015. http://www.sciencedomain.org/issue.php?iid=707&id=6

Experiment Data



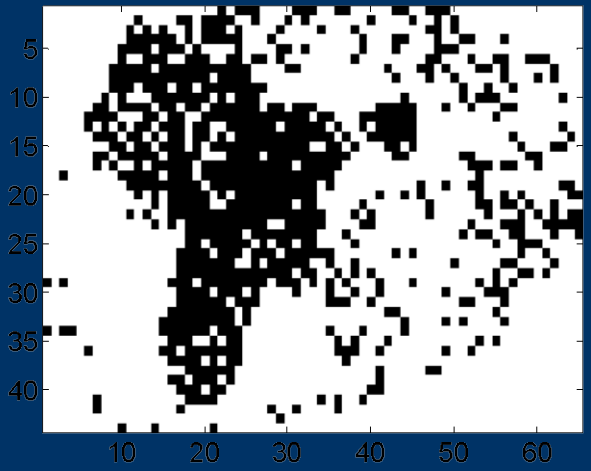
→

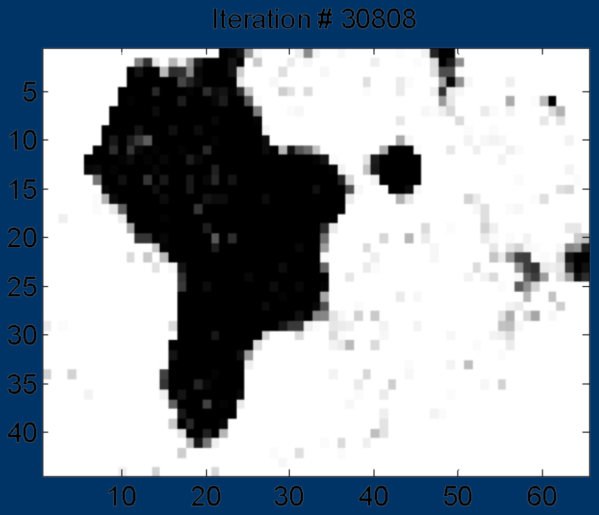


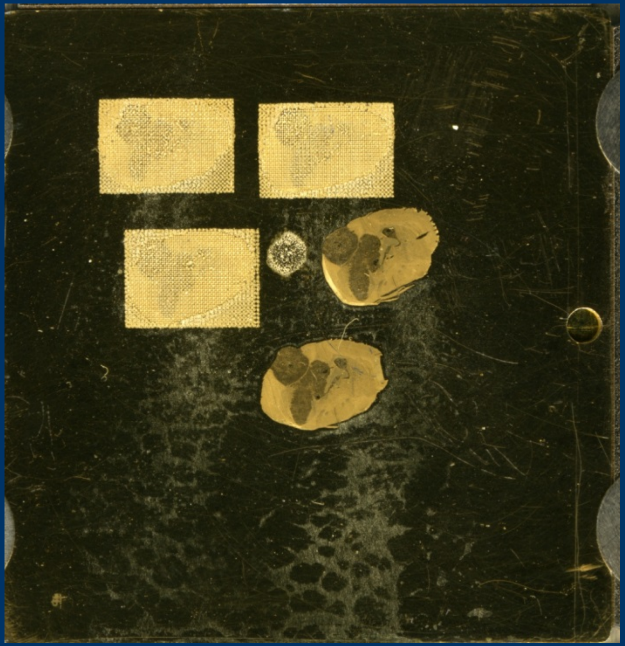
Training data

Test data

Result


5
10
15
20
25
30
35
40
10
20
30
40
50
60


Iteration # 30808
5
10
15
20
25
30
35
40
10
20
30
40
50
60



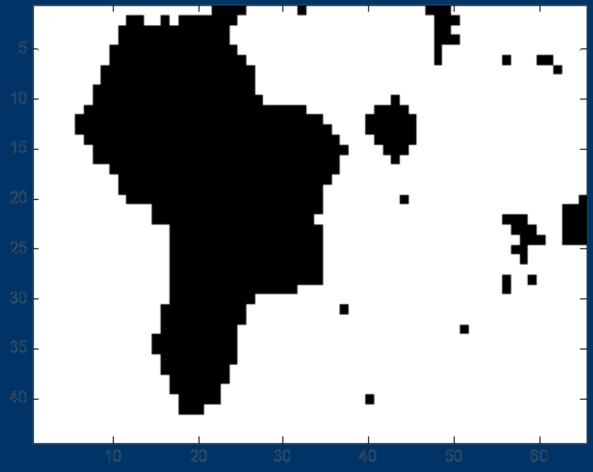
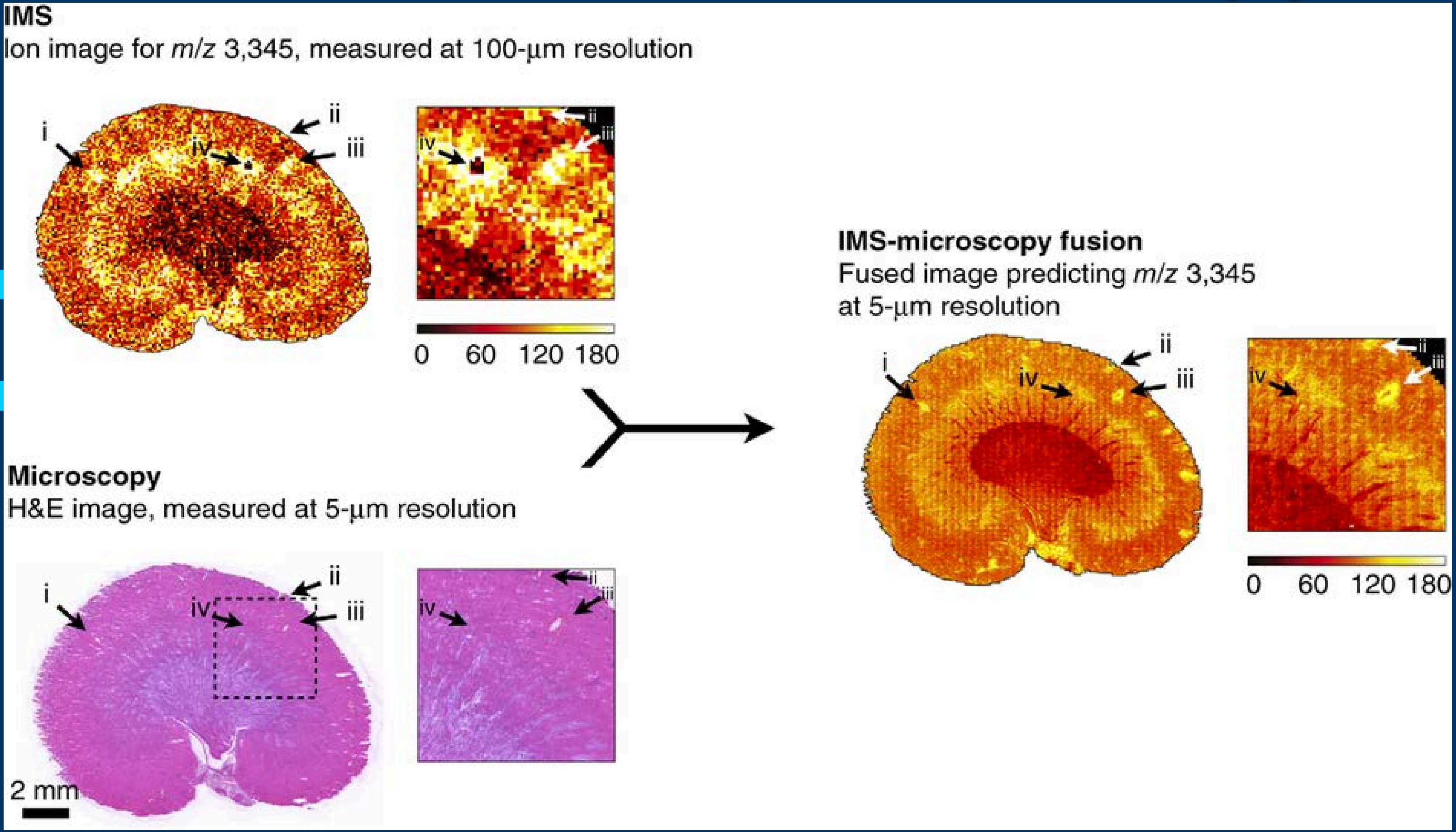

5
10
15
20
25
30
35
40
10
20
30
40
50
60

Image fusion of mass spectrometry and microscopy

- Raf Van de Plas, Junhai Yang, Jeffrey Spraggins & Richard M Caprioli, Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping, *Nature Methods* 12, 366–372 (2015) doi:10.1038/nmeth.3296

Fusion IMS & Microscopy

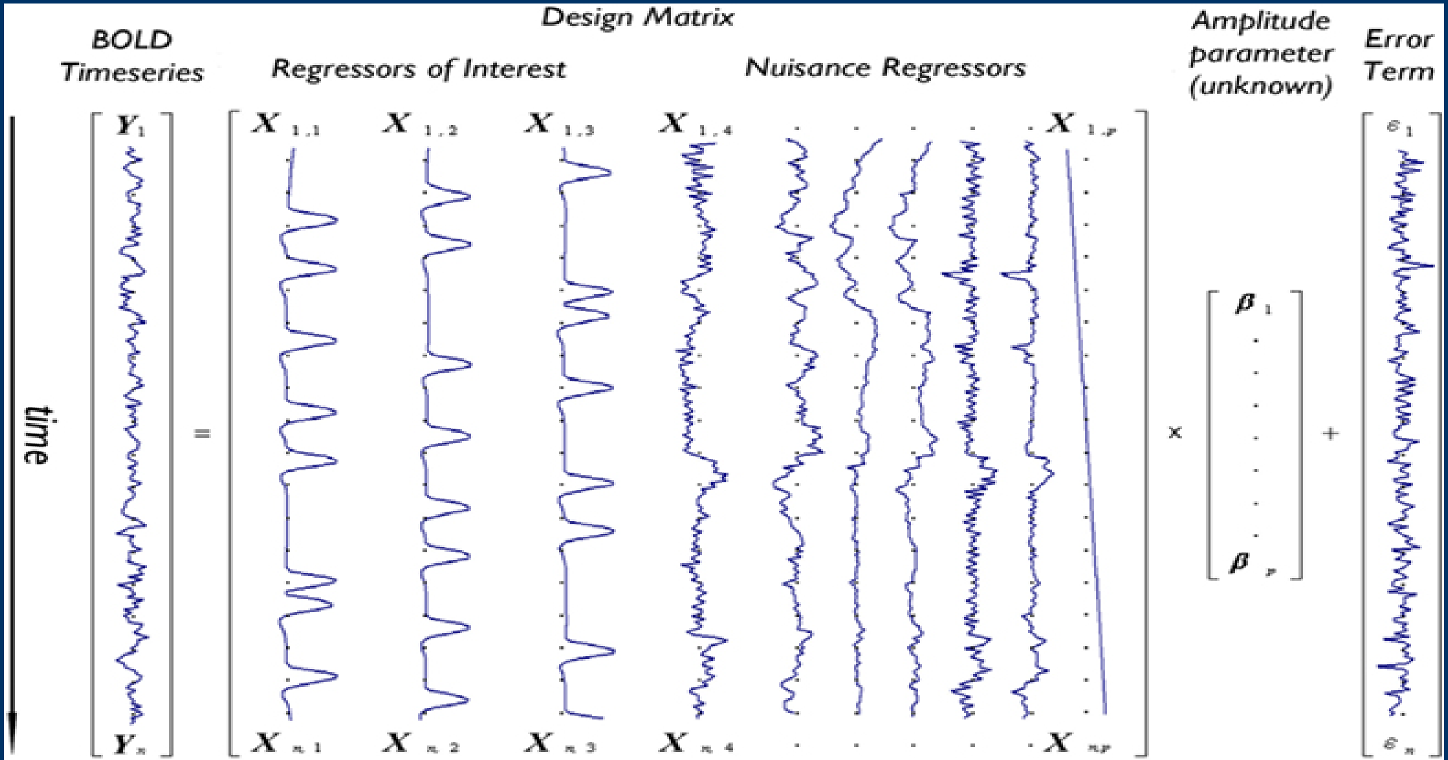

IMS
Ion image for m/z 3,345, measured at 100-µm resolution
i
ii
iii
iv
0 60 120 180
Microscopy
H&E image, measured at 5-µm resolution
2 mm
IMS-microscopy fusion
Fused image predicting m/z 3,345
at 5-µm resolution

- **From** Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping, Raf Van de Plas, Junhai Yang, Jeffrey Spraggins & Richard M Caprioli *Nature Methods* **12**, 366–372 (2015) doi:10.1038/nmeth.3296

II. fMRI Data Processing

- The aims of fMRI analysis include: localizing regions of the brain activated by a certain task; determining distributed networks that correspond to brain function; and making predictions about psychological or disease states.
- Also, modeling brain connectivity, visualizing and localizing activation, and analyzing region of interest

Statistical Analysis of fMRI Data

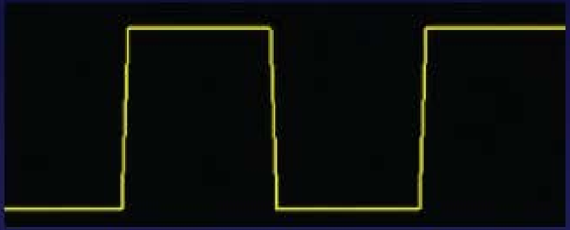


fMRI Data Processing

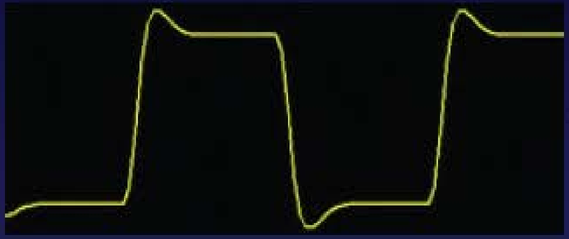
- Two popular models to process the fMRI data are General Linear Model (GLM) and Independent Component Analysis (ICA). Both methods need to incorporate adaptivity for fMRI study.
- The fMRI design that the BOLD response behaved as a linear time invariant system is crucial for simplifying the analysis by allowing the using of GLM.
- Since 2000, new approaches that analyzes patterns of activity rather than the response at voxels are developed including using pattern information analysis and machine learning.

General Linear Model

1. Model (1 or more Regressors)

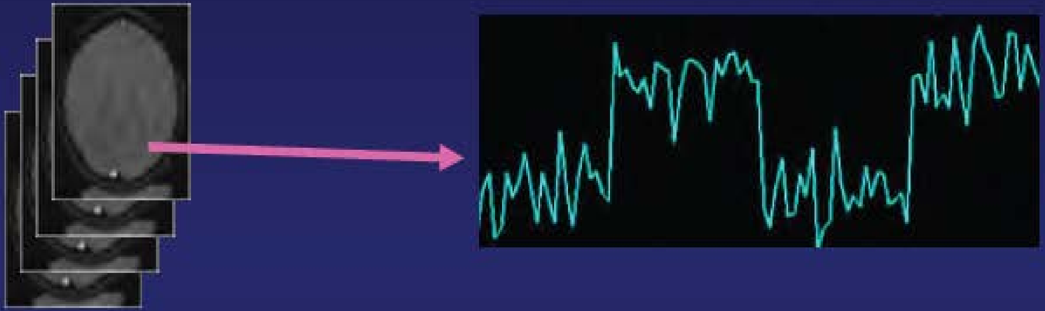


or



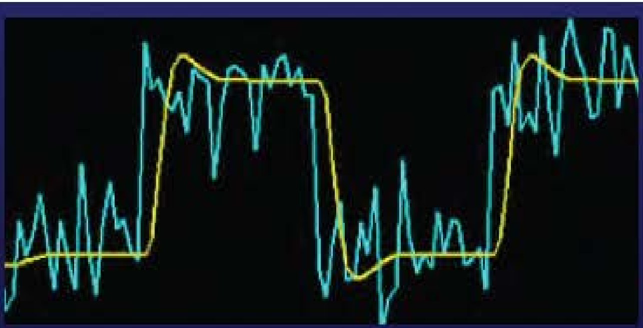
$x_i(j)$

2. Data

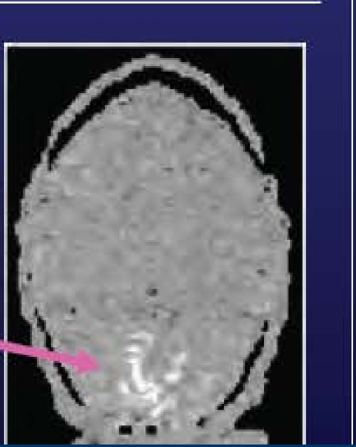


$y(j)$

3. Fitting the Model to the Data at each voxel

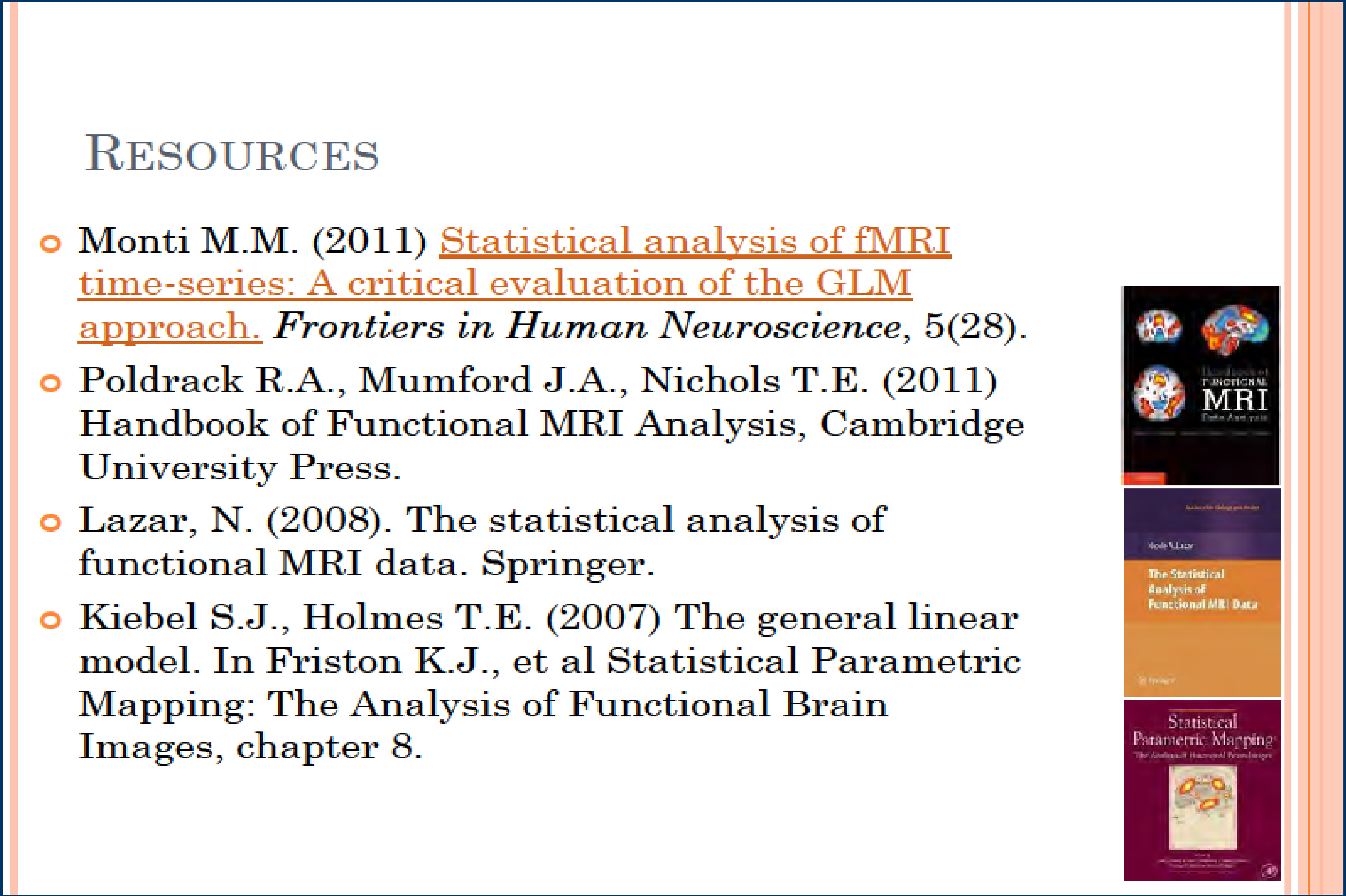


Regression Results



$$y(j) = \hat{\beta}_0 + \sum^{M} \hat{\beta}_i x_i(j) + e(j)$$

GLM



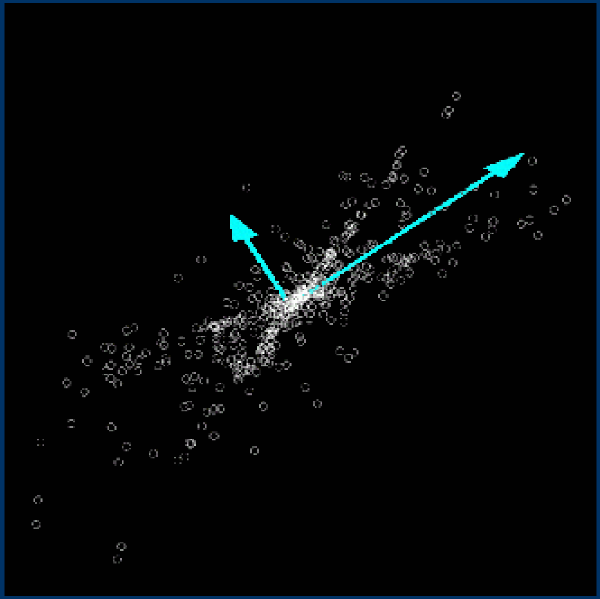
Resources

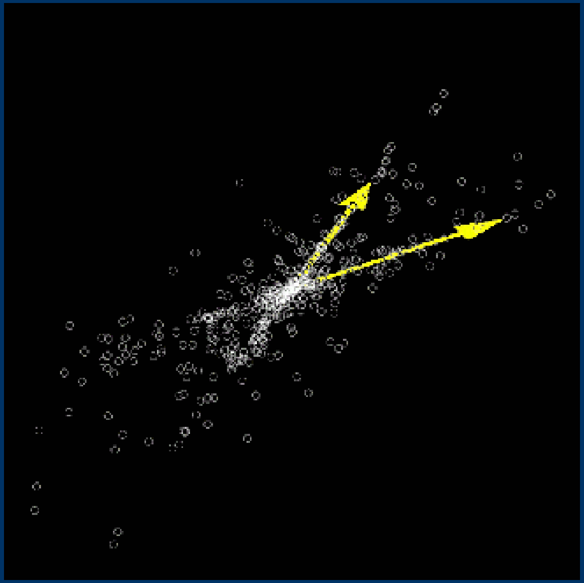
- Monti M.M. (2011) Statistical analysis of fMRI time-series: A critical evaluation of the GLM approach. *Frontiers in Human Neuroscience*, 5(28).
- Poldrack R.A., Mumford J.A., Nichols T.E. (2011) Handbook of Functional MRI Analysis, Cambridge University Press.
- Lazar, N. (2008). The statistical analysis of functional MRI data. Springer.
- Kiebel S.J., Holmes T.E. (2007) The general linear model. In Friston K.J., et al Statistical Parametric Mapping: The Analysis of Functional Brain Images, chapter 8.

- Context determines the meaning and interpretation of the word “network” in brain imaging analysis.
- GLM and seed-based methods define a network as a subset of voxels whose time series are significantly correlated with a reference signal.
- ICA defines a network as a subset of voxels whose time series are significantly correlated with the estimated ICA time course

- GLM usually needs to know the design patterns of the experiments. It is simple to extend the application of GLM from single subject analysis to the group analysis. However, for some specific design experiment, we merely know the patterns.
- We can use ICA to separate the independent patterns and the spatial maps from the BOLD signals. Temporal concatenation is the most popular way to organize the group data.
- We can obtain the group specific temporal responses and common spatial maps via the tensor model.

PCA and ICA

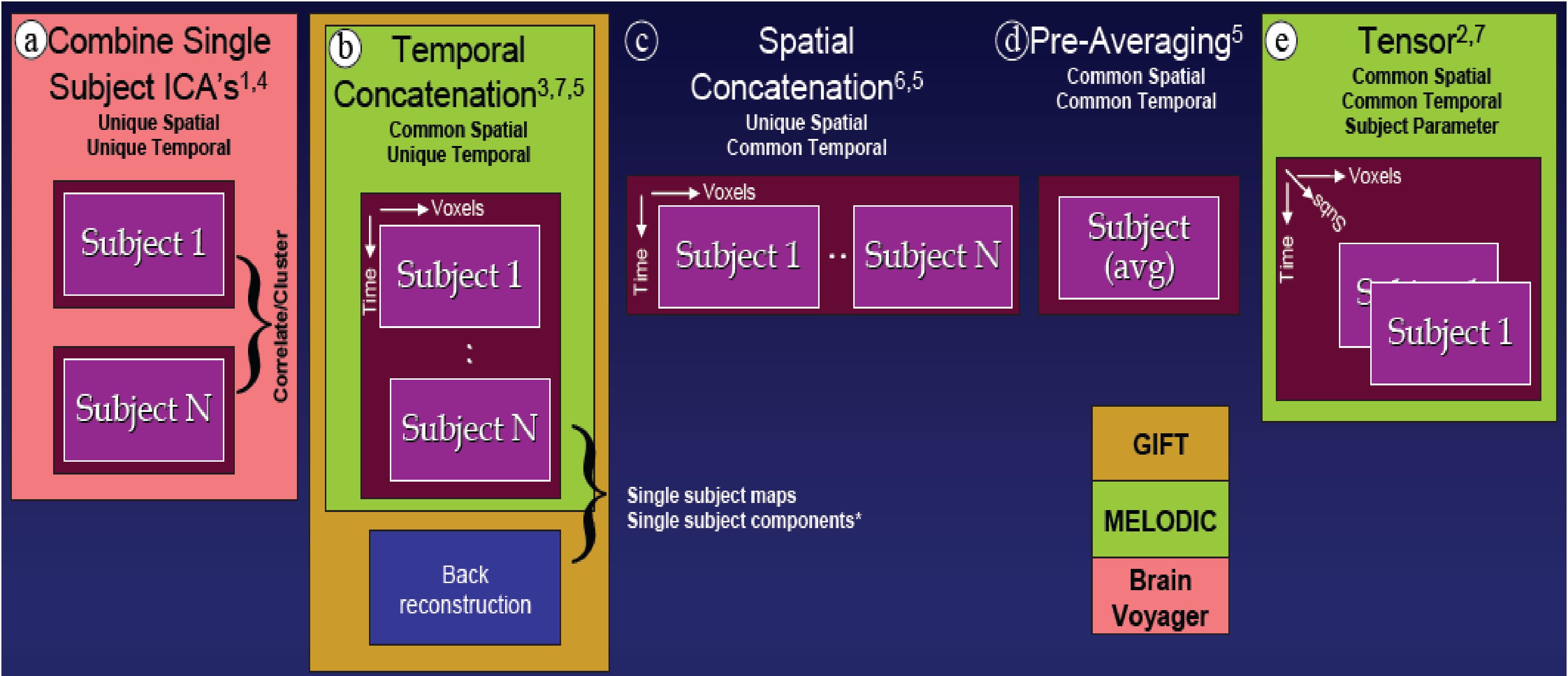




PCA finds directions of maximal variance

ICA finds directions which maximize independence

Group ICA



1) Calhoun VD, Adali T, McGinty V, Pekar JJ, Watson T, Pearlson GD. (2001): fMRI Activation In A Visual-Perception Task: Network Of Areas Detected Using The General Linear Model And Independent Component Analysis. NeuroImage 14(5):1080-1088.

2) Beckmann CF, Smith SM. (2005): Tensorial extensions of independent component analysis for multisubject FMRI analysis. NeuroImage 25(1):294-311.

3) Calhoun VD, Adali T, Pearlson GD, Pekar JJ. (2001): A Method for Making Group Inferences from Functional MRI Data Using Independent Component Analysis. Hum.Brain Map. 14(3):140-151.

4) Esposito F, Scarabino T, Hyvarinen A, Himberg J, Formisano E, Comani S, Tedeschi G, Goebel R, Seifritz E, Di SF. (2005): Independent component analysis of fMRI group studies by self-organizing clustering. Neuroimage. 25(1):193-205.

5) Schmithorst VJ, Holland SK. (2004): Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data. J.Magn Reson.Imaging 19(3):365-368.

6) Svensen M, Kruggel F, Benali H. (2002): ICA of fMRI Group Study Data. NeuroImage 16:551-563.

7) Guo Y, Giuseppe P. (In Press): A unified framework for group independent component analysis for multi-subject fMRI data. NeuroImage.

- V. D. Calhoun, J. Liu, and T. Adali, "A Review of Group ICA for fMRI Data and ICA for Joint

ICA for fMRI

- Let $i = 1, \cdots, N$ represent N subjects, $t = 1, \cdots, T$ the T time points, and $v = 1, \cdots, V$ the V voxels. Use X to represent the temporal concatenation of fMRI data of N subjects (groups). The group ICA model can be expressed as:
  - $X = MS + E,$
- where each column of S is a $q \times 1$ vector containing q statistically independent non-Gaussian spatial source signals at the voxel v, M is a $TN \times q$ matrix that contains q independent spatial source signals of the observed multi-subject fMRI images, each column of E is a $TN \times 1$ vector representing the N subjects' noise at voxel v following a multivariate normal distribution $MVN(0, IN \otimes \Sigma v)$ where Σv is a $T \times T$ error covariance.

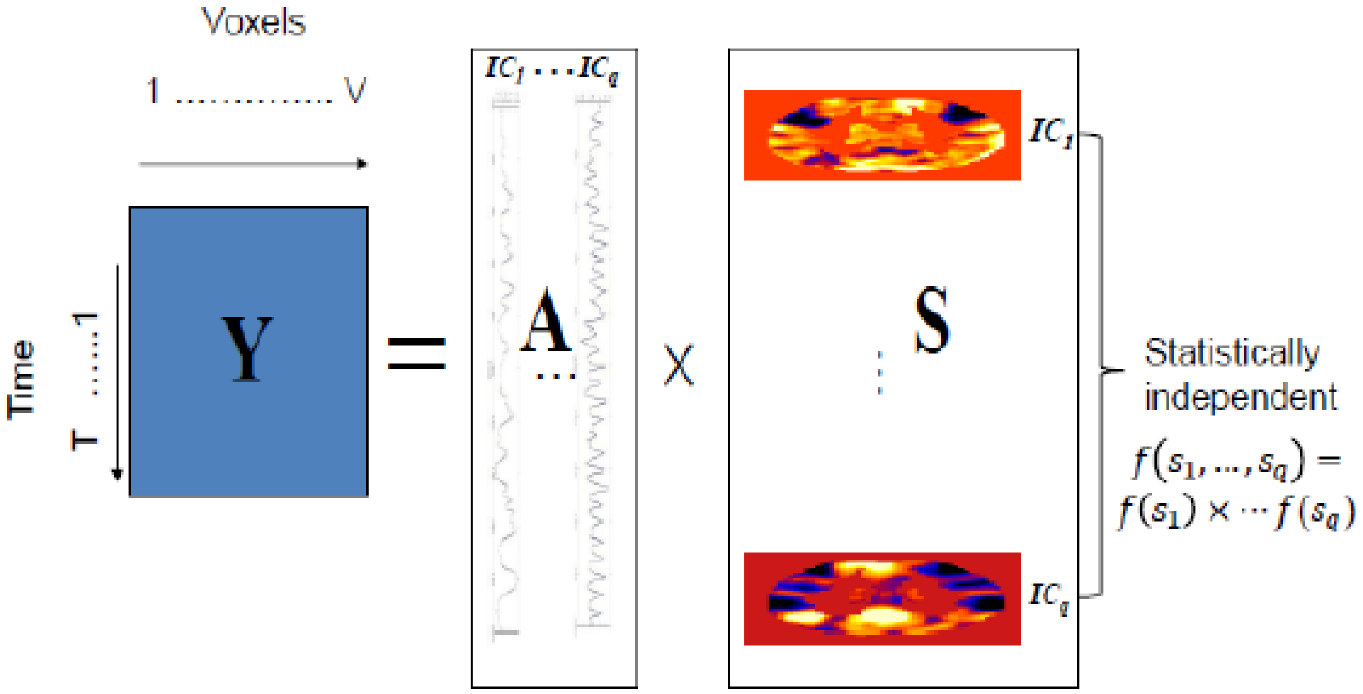

Voxels
1 V
Time
T 1
Y
=
$IC_1 \cdots IC_q$
A
...
X
S
IC_1
IC_q
Statistically independent
$f(s_1, \dots, s_q) = f(s_1) \times \cdots f(s_q)$

Figure: ICA explanation.

Probabilistic Group ICA for fMRI

- To be better to investigate differentiations between groups of subjects in terms of a common time courses with subject loading coefficients under the ICA spatial map, following a similar procedure of Probabilistic PCA, JS Liang at MTSU COMS PhD program proposed a so-called multi-group Probabilistic tensorial ICA algorithm for fMRI data analysis to improve computing efficiency and accuracy of the group ICA algorithm.

Multi-Group PTICA for fMRI

- Algorithm Outline:
- 1. Modify the original data X time course to be normalized to zero mean and unit variance.
- 2. Reduce the dimension of the data at subject level using PCA. Use the appropriate matrix H, we have
- $HX = HMS + HE$. This can be deemed as a new equation of the same format
    - $X = MS + E$
- 3. Use Laplace approximation or minimum description length(MDL) to estimate the number of the latent independent components to extract reasonable features from the mixed data in the following steps.

Algorithm Outline (cont.):

- 4. Use PICA approach for the decomposition of data X into a mixing matrix M and spatial maps S.
- $M_{ML} = U_{q}(\Lambda_{q} - \sigma^{2} * I_{q})^{1/2} Q^{t}$
- $S_{ML} = (A^{t} A)^{-1} A^{t} X$
- $\sigma^{2}_{ML} = [1/(NT - q)] \Sigma_{l=q+1}^{NT} \lambda_{l}$
- where $X = U(N\Lambda)^{1/2} V$. U_{q} and Λ_{q} contain the first q eigenvectors and eigenvalues of U and Λ, and Q denotes a $q \times q$ orthogonal rotation matrix.
- 5. For each group, use rank-1 approximation via SVD to
- decompose the mixing matrix M to $C \otimes A$, where $\otimes$ is
- Khatri-Rao Product. Therefore, for each group, we try to find the common time course A and the corresponding subject loading C .

- 6. Iterate the above steps until convergence.
- 7. Use a reconstruction method to rebuild the time courses *A* and spatial maps *S*.
- 8. Apply some inference methods like modified *z* maps to detect the activation maps.
- 9. Compare the time courses between the groups to find the potential ICs.

Suppose we have two groups: the contrast group and the group with Autism. Two groups have two time courses of the same independent spatial map. Calculate the correlate coefficient r between them. If r is quite small, we intend to believe this IC is associated with the illness.

Applications

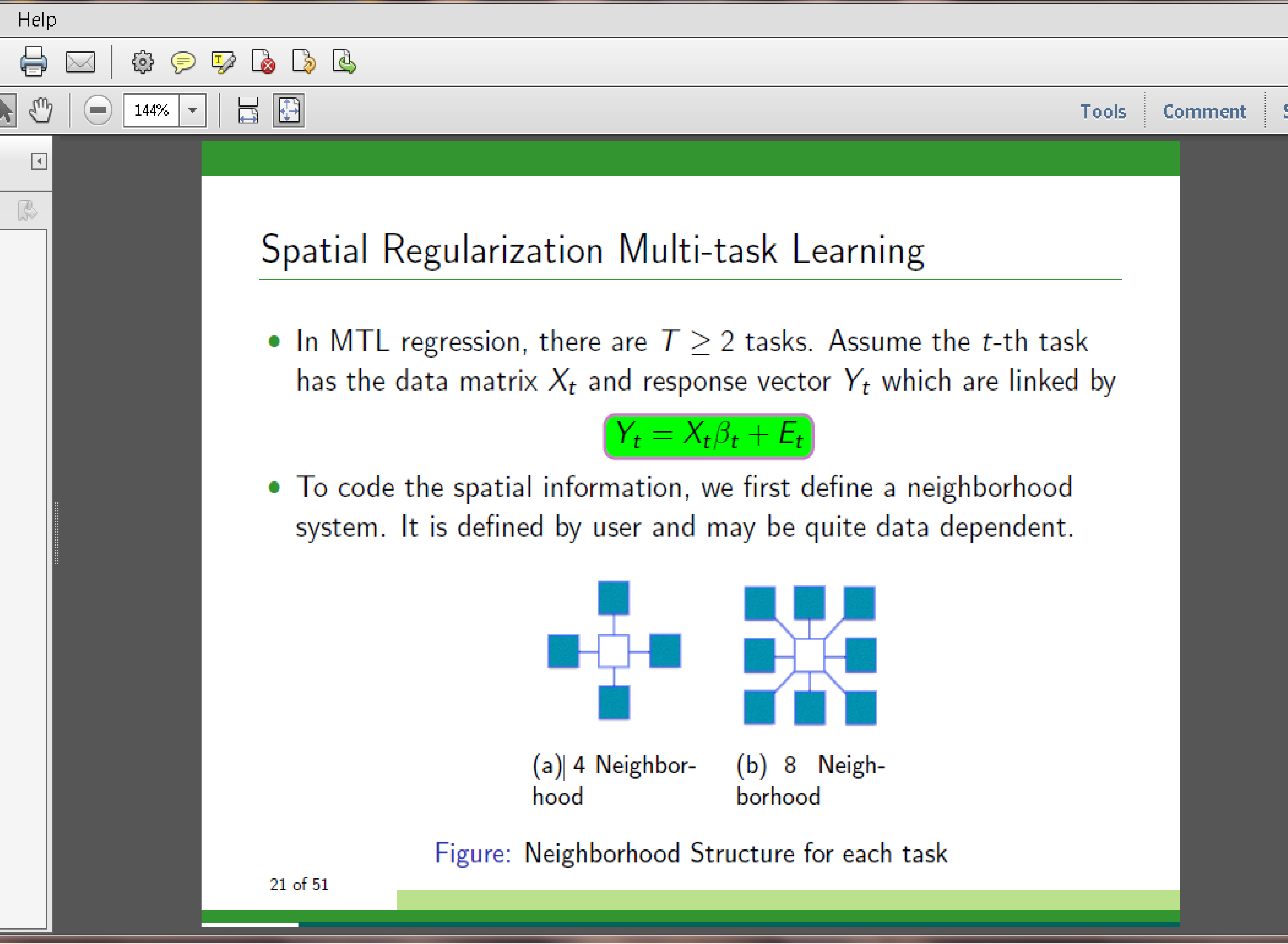
- Autism data analysis: not only compare patient group with normal group, the corresponding model can be used to study and compare groups based on severity scores to seek possible new biomarkers
- Alzheimer Disease data study

Multi-task Learning for fMRI Data Processing

- Ph.D. student, Xin Yang (co-supervising with Qiang Wu), applies multitask learning scheme for fMRI data analysis.

- The idea is to find a better pattern representation by using multi-task learning technique defined on voxels.

- In the subsequent slides, index t is for the task labels.



Spatial Regularization Multi-task Learning

- In MTL regression, there are $T \geq 2$ tasks. Assume the t-th task has the data matrix X_t and response vector Y_t which are linked by

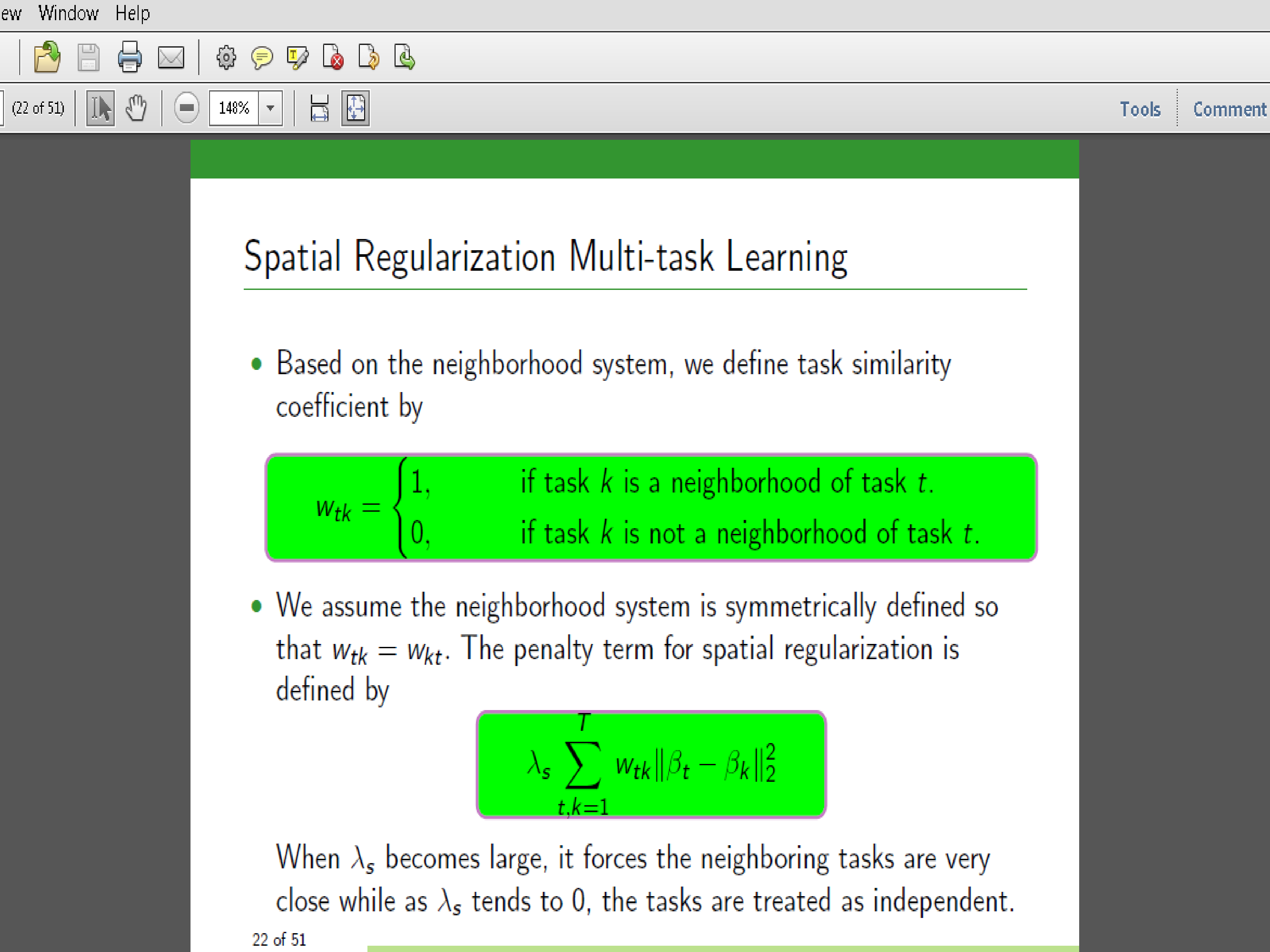
$$\boxed{Y_t = X_t\beta_t + E_t}$$

- To code the spatial information, we first define a neighborhood system. It is defined by user and may be quite data dependent.

(a) 4 Neighborhood

(b) 8 Neighborhood

Figure: Neighborhood Structure for each task



Spatial Regularization Multi-task Learning

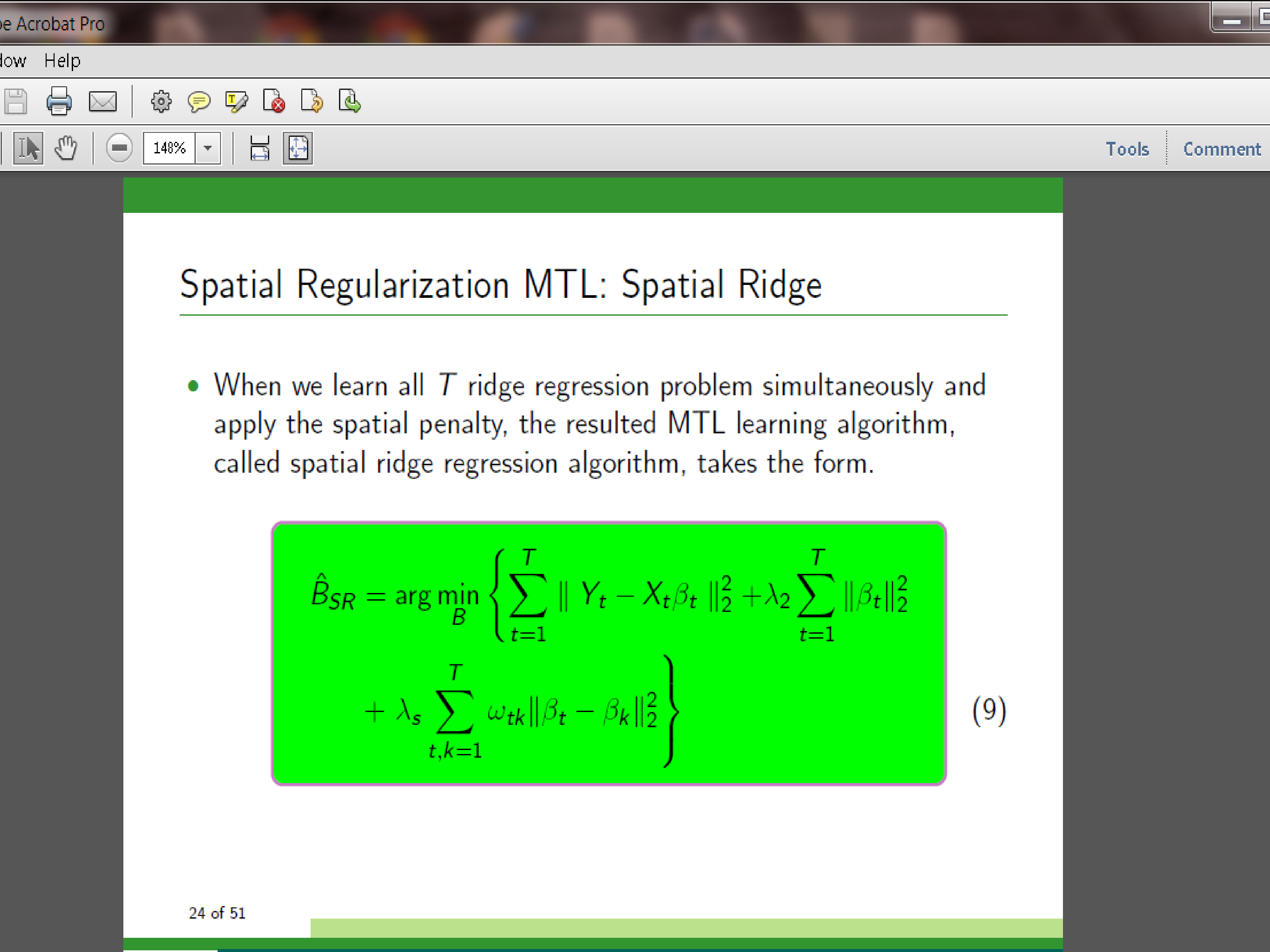
- Based on the neighborhood system, we define task similarity coefficient by

$$w_{tk} = \begin{cases} 1, & \text{if task } k \text{ is a neighborhood of task } t. \\ 0, & \text{if task } k \text{ is not a neighborhood of task } t. \end{cases}$$

- We assume the neighborhood system is symmetrically defined so that $w_{tk} = w_{kt}$. The penalty term for spatial regularization is defined by

$$\lambda_s \sum_{t,k=1}^{T} w_{tk} \|\beta_t - \beta_k\|_2^2$$

When λ_s becomes large, it forces the neighboring tasks are very close while as λ_s tends to 0, the tasks are treated as independent.



Spatial Regularization MTL: Spatial Ridge

- When we learn all T ridge regression problem simultaneously and apply the spatial penalty, the resulted MTL learning algorithm, called spatial ridge regression algorithm, takes the form.

$$\boxed{\begin{aligned}\hat{B}_{SR} = \arg\min_{B} \Bigg\{ &\sum_{t=1}^{T} \| Y_t - X_t\beta_t \|_2^2 + \lambda_2 \sum_{t=1}^{T} \|\beta_t\|_2^2 \\ &+ \lambda_s \sum_{t,k=1}^{T} \omega_{tk}\|\beta_t - \beta_k\|_2^2 \Bigg\}\end{aligned}} \qquad (9)$$

Multi-task Learning for fMRI

- From the simulation and real data results, we see that MTL algorithms achieve much better results than STL algorithms.
- Regularized MTL algorithm considers all task’s average instead of tasks with similar features.

Acknowledgement

- Mass Spectrometry Research Center at Vanderbilt University School of Medicine.
- Institute of Imaging Science, Vanderbilt University
- Department of Child & Adolescent Psychiatry, Vanderbilt University
- Institute of Images and Pattern Recognition, North China University of Technology

Journal of Health & Medical Informatics

- The Journal of Health & Medical Informatics (JHMI) is a scholarly journal that integrates the information and communication related to healthcare, medicine and their applications in medical information systems.

  Don Hong, Editor-in-Chief,

  Middle Tennessee State University, Murfreesboro, TN

  David Randall, Co-Editor-in-Chief,

  the Consumer Driven Health Care Institute, Washington DC, USA

- Submit manuscript at http://editorialmanager.com/omicsgroup/ or send as an e-mail attachment to the Editorial Office at editor.jhmi@omicsonline.org

Thank you !!!