Computational Methods in NeuroImage Analysis

Instructor: Moo K. Chung
mkchung@wisc.edu

Lecture 10
Logistic regression

November 12, 2010

Motivation/Data

Based on 8 elderly controls (EC) and 7 mild cognition impairment (MCI) subjects, we perform the logistic discriminant analysis (LDA) on cortical thickness and cortical surface area to see if these two measures can be used to discriminate MCI from EC.

Data is obtained from Sterling Johnson

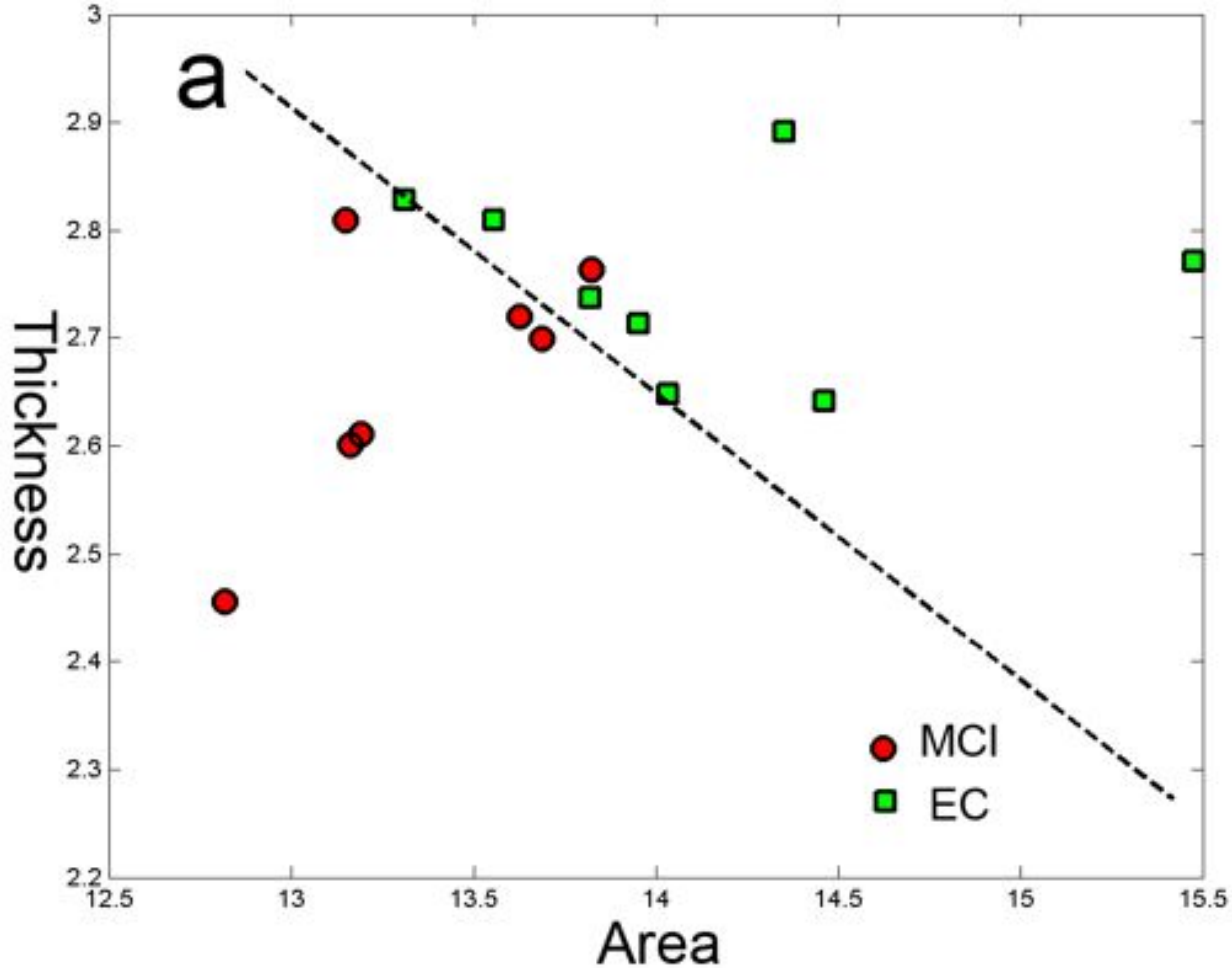
Previous work on AD and cortical thickness

- Cortical thickness has been shown to characterize cortical atrophy in AD patients quite well (Lerch et al., NeuroImage, 2005).
- Hypothesis on AD progression:

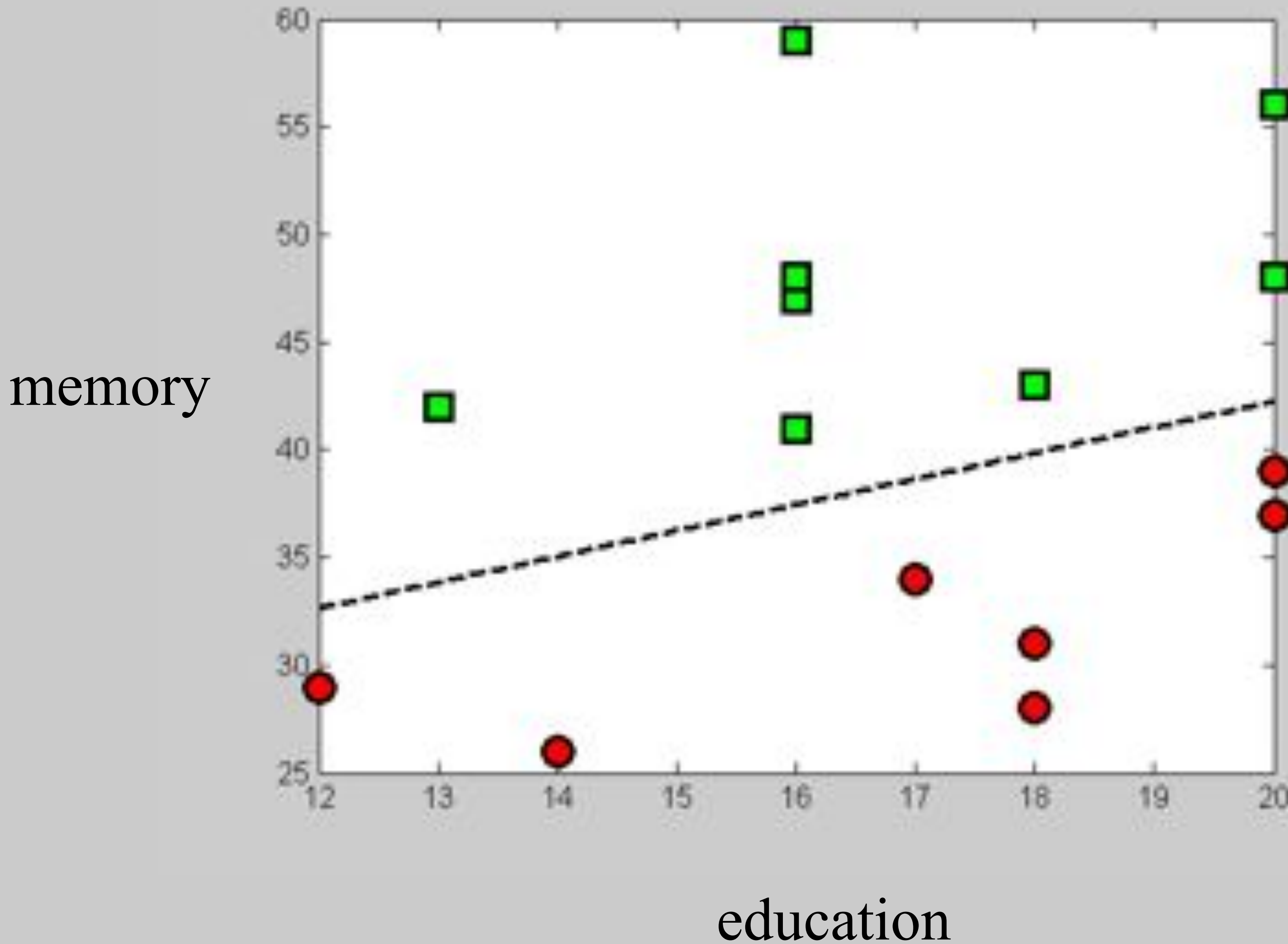
  NC → MCI → AD
- Question: It is unclear if cortical thickness will be an important biomarkers of discriminating EC vs. MCI.

	group	sex	age	education	memory	thickness	area
	0	2	63	16	47	2.7718	15.4768
	0	2	66	16	48	2.7132	13.953
	0	1	70	18	43	2.8094	13.5577
EC	0	1	75	20	48	2.7371	13.821
	0	2	75	16	59	2.6476	14.0353
	0	2	64	13	42	2.8913	14.354
	0	1	69	20	56	2.8284	13.3125
	0	1	81	16	41	2.6412	14.463
	1	2	75	12	29	2.7204	13.628
	1	1	62	20	39	2.6992	13.6875
MCI	1	1	68	17	34	2.7632	13.823
	1	2	77	14	26	2.6106	13.1918
	1	1	80	18	28	2.4556	12.8165
	1	1	78	20	37	2.6012	13.164
	1	2	64	18	31	2.8094	13.1511

Data: Group, Thickness, Area



Area is a discriminating variable.
EC has more area (folding).


memory
education
60
55
50
45
40
35
30
25
12
13
14
15
16
17
18
19
20

Suppose we have p regressors $X_1, \cdots, X_p$. These can be both imaging and nonimaging biomarkers such as local area, cortical thickness, gender, age and behavioral measures at a voxel. Let $x_{i1}, \cdots, x_{ip}$ denote the measurements for the i-th subject. Let the response variable Y_i be the clinical state of the i-th subject modeled as a Bernoulli random variable with parameter π_i. $Y_i = 1$ if the i-th subject is autistic with probability π_i while $Y_i = 0$ if the subject is normal with probability $1 - \pi_i$. π_i is the likelihood (probability) of a subject belong to the group 1, i.e. $\pi_i = P(Y_i = 1)$. For instance, Y_i can indicate the subject belongs to the elderly normal control or mild cognition impairment group respectively in an Alzheimer's disease study.

Now consider a general linear model

$$Y_i = \mathbf{x}_i'\beta + \epsilon_i, \tag{7.31}$$

where $\mathbf{x}_i' = (1, x_{i1}, \cdots, x_{ip})$ and $\beta' = (\beta_0, \cdots \beta_p)$. We may assume $\mathbb{E}\epsilon_i = 0$ and $\mathbb{V}\epsilon_j = \sigma^2$. In this case, (7.31) is no longer appropriate since

$$\mathbb{E}Y_j = \pi_i = \mathbf{x}_i'\beta$$

but $\mathbf{x}_i'\beta$ may not be in the range $[0, 1]$. This inconsistency is caused by trying to match the continuous variables $x_{i1}, \cdots, x_{ip}$ to the categorical variable Y_i directly. To address this problem, we introduce the *logistic regression function* g:

$$\pi_i = g(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i'\beta_i)}{1 + \exp(\mathbf{x}_i'\beta_i)}.$$

Then using the *logit function*, we can write this as

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \mathbf{x}_i'\beta_i.$$

The unknown parameters β are estimated via the maximum likelihood estimation (MLE). The likelihood function is

$$\begin{aligned} L(\beta | y_1, \cdots, y_n) &= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i'\beta_i)}{1 + \exp(\mathbf{x}_i'\beta_i)} \right]^{y_i} \prod_{i=1}^{n} \left[\frac{1}{1 + \exp(\mathbf{x}_i'\beta)} \right]^{1 - y_i}. \end{aligned}$$

The loglikelihood function is given by

$$\begin{aligned}\log L(\beta) &= \text{const.} + \sum_{i=1}^{n} y_i \log \pi_i + (1 - y_i)\log(1 - \pi_i) \\ &= \text{const.} + \sum_{i=1}^{n} y_i \mathbf{x}_i' \beta + \log(1 - \pi_i)\end{aligned}$$

and its maximum is obtained when

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - \pi_i) = 0.$$

In simplifying the expression, we used the following identities

$$\frac{\partial \pi_i}{\partial \beta_0} = \pi_i(1 - \pi_i)$$

and

$$\frac{\partial \pi_i}{\partial \beta_1} = x_i \pi_i (1 - \pi_i).$$

Since the logistic regression function π is in a complicated form, the maximum is obtained numerically. Define the *information matrix* to be

$$I(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial \beta' \partial \beta} = -\sum_{i=1}^{n} \pi_i (1 - \pi_i) \mathbf{x}_i \mathbf{x}_i'.$$

Then the Newton-Raphson algorithm is used to find the MLE in an iterative fashion. Starting with an arbitrary initial vector β^0, we estimate iteratively as

$$\beta^{j+1} = \beta^j + I(\beta^j)^{-1} \frac{\partial \log L(\beta)}{\partial \beta}(\beta^j).$$

Performance of a classifier

- Error rate = over all probability of making wrong classification

- Simplest way to estimate the error rate is to use the concept of cross-validation.

- Leave one out cross-validation

Most statistical data analysis packages such as R and MATLAB has a built-in routine for estimating the parameters of the logistic regression.

The discriminant analysis resulting from the estimated logistic model is called the *logistic discrimination*. We classify the i-th subject according to a *classification rule*. The simplest rule is to assign the i-th subject as group 1 if we have

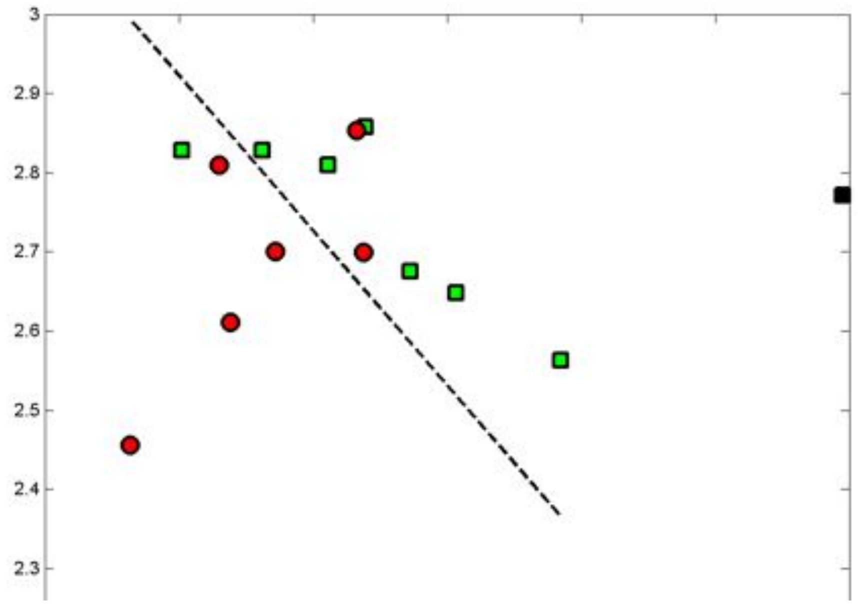
$$P(Y_i = 1) > P(Y_i = 0).$$

This statement is equivalent to $\pi_i > 1/2$. Depending on the bias and the error of the estimation, the value $1/2$ can be adjusted. For the fitted logistic model, we classify the i-th subject as group 1 if $\mathbf{x}_i'\beta_i > 0$ and as 0 if $\mathbf{x}_i'\beta_i < 0$. The plane $\mathbf{x}_i'\beta = 0$ is the *classification boundary* that separates the two groups. The performance of classification technique is measured by the *error rate* γ, the overall probability of misclassification.

The *cross-validation* is mainly used to estimate the error rate. This is done by randomly partitioning the data into the training and testing sets. In the *leave-one-out* scheme, the training set consists of $n-1$ subjects while the testing set consists of one subject. Suppose the i-th subject is taken as the test set. Then using the training set, we determine the logistic model. Using the estimated model, we test if the i-th subject is correctly classified. It is classified correctly, we let the classification error $e_{-1} = 0$ and $e_{-1} = 1$ otherwise. The leave-one-out error rate is given by

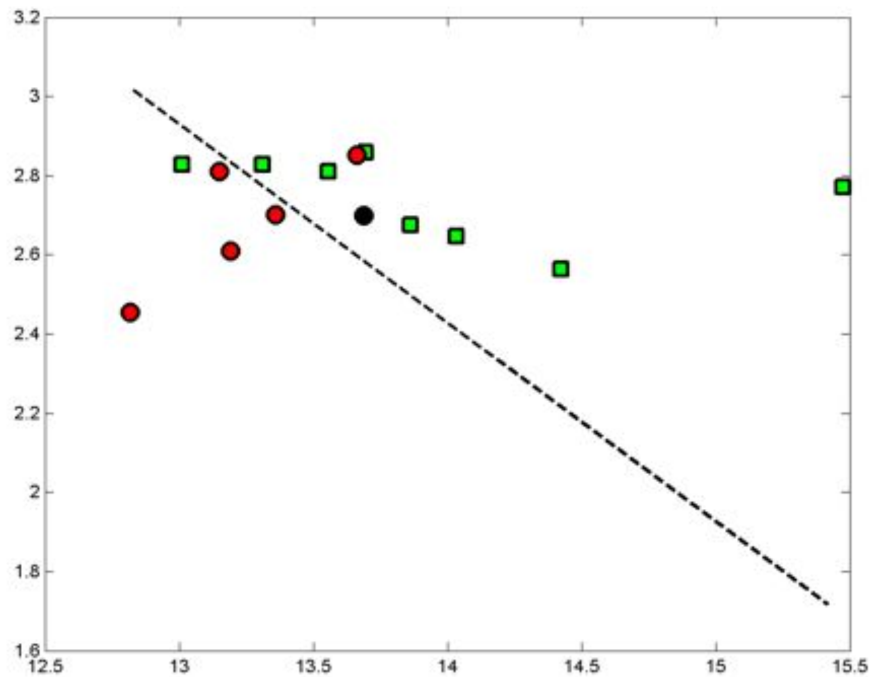
$$\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} e_{-i}.$$

The *discriminant power* is then given as $1 - \widehat{\gamma}$.



Correct decision

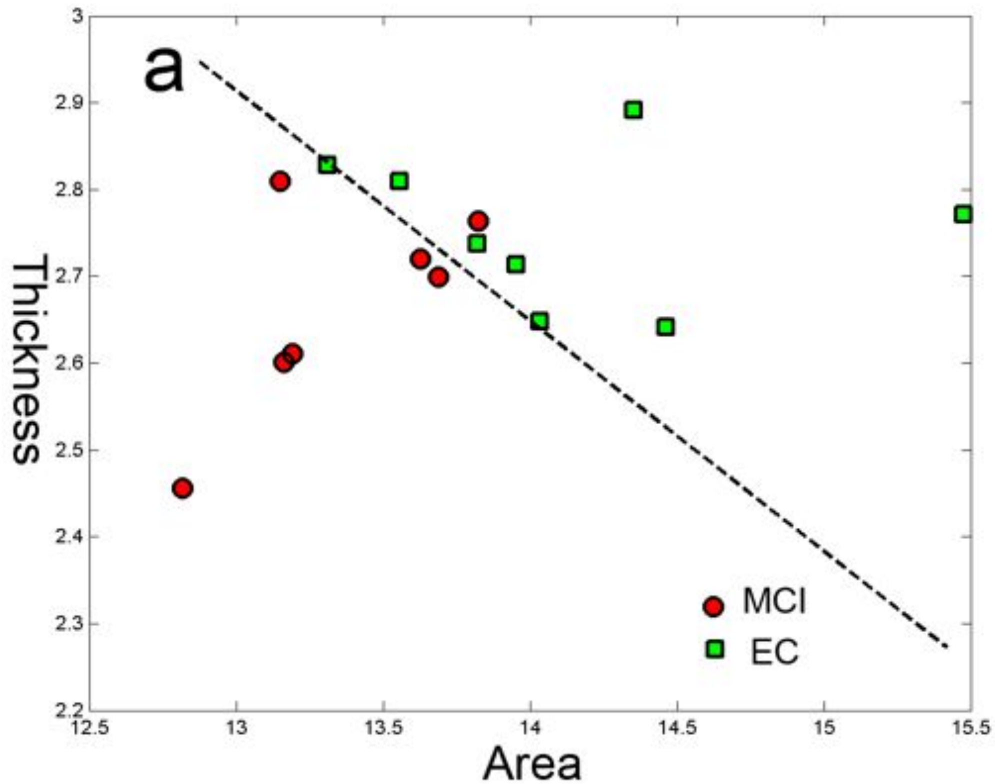
error= 0



Incorrect decision

error = 1

Significant variables



Error rate:

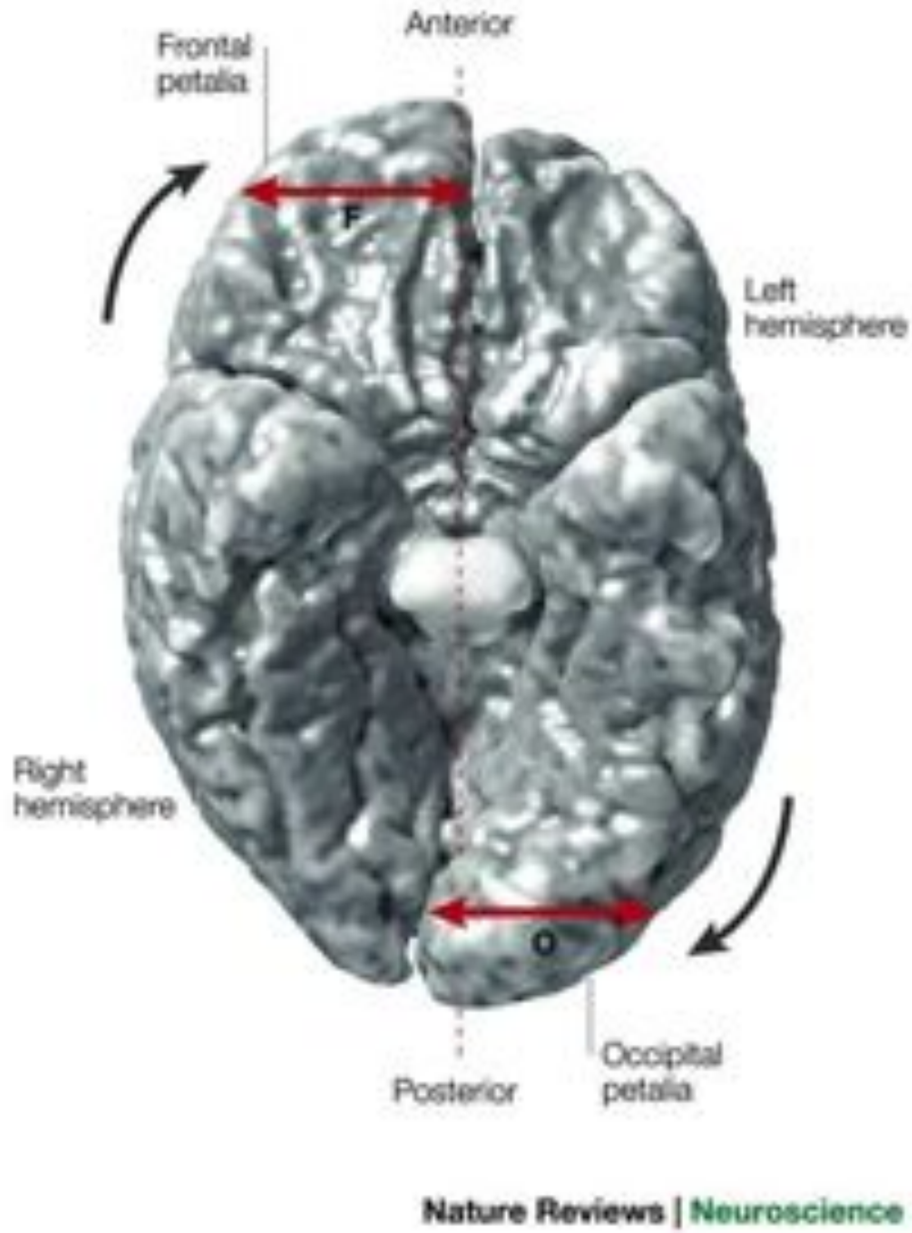
thickness --> 47%

thickness + area --> 20%   area --> 20%

MATLAB demonstration

Cortical Asymmetry Analysis

Yakovlevian torque



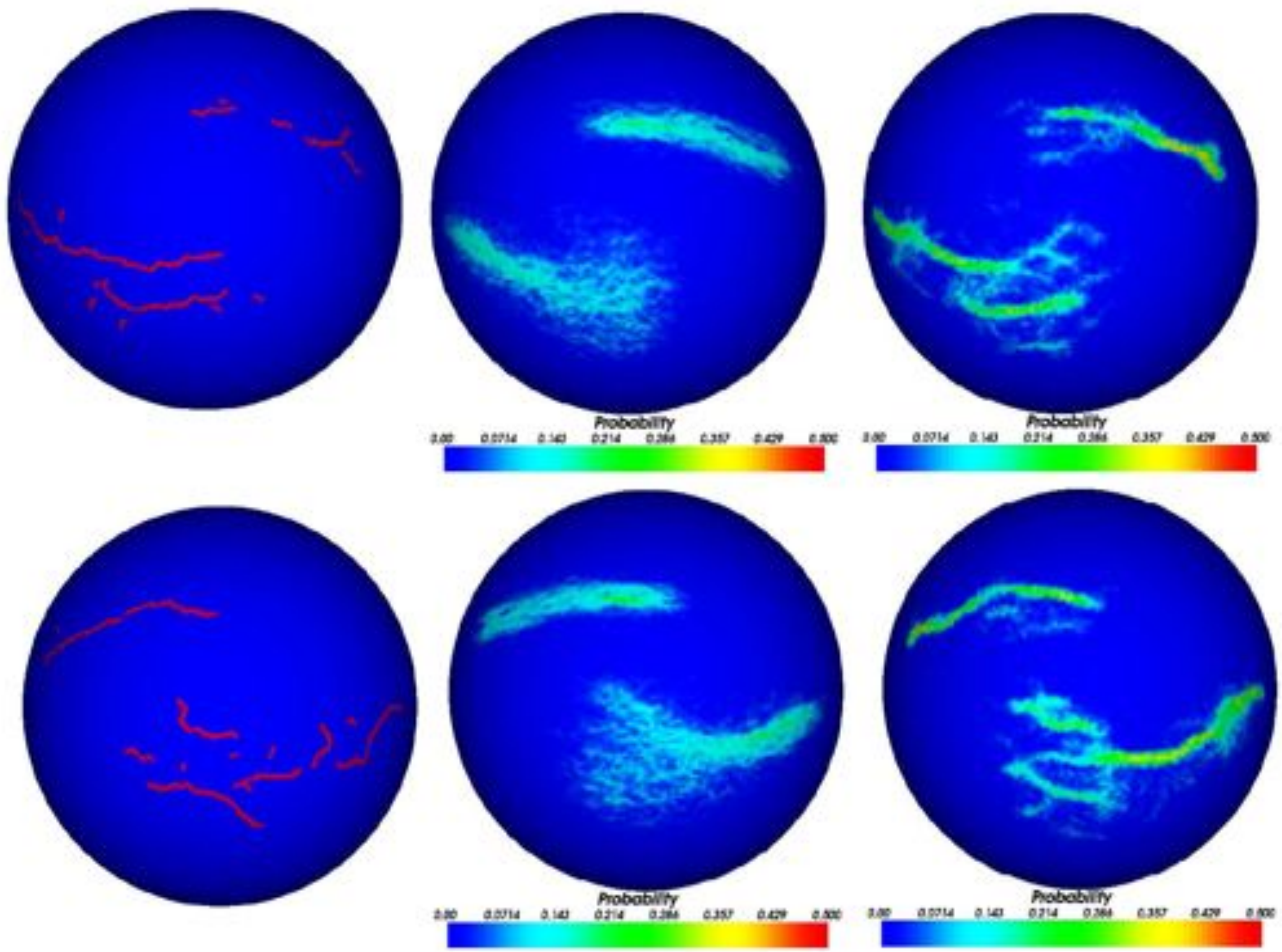
A twisting effect is also observed, known as Yakovlevian torque, in which structures surrounding the right Sylvian fissure are 'torqued forward' relative to their counterparts on the left.

Toga & Thompson 2003. Nature Reviews Neuroscience 4, 37-48

Sulcal pattern asymmetry on 149 subjects

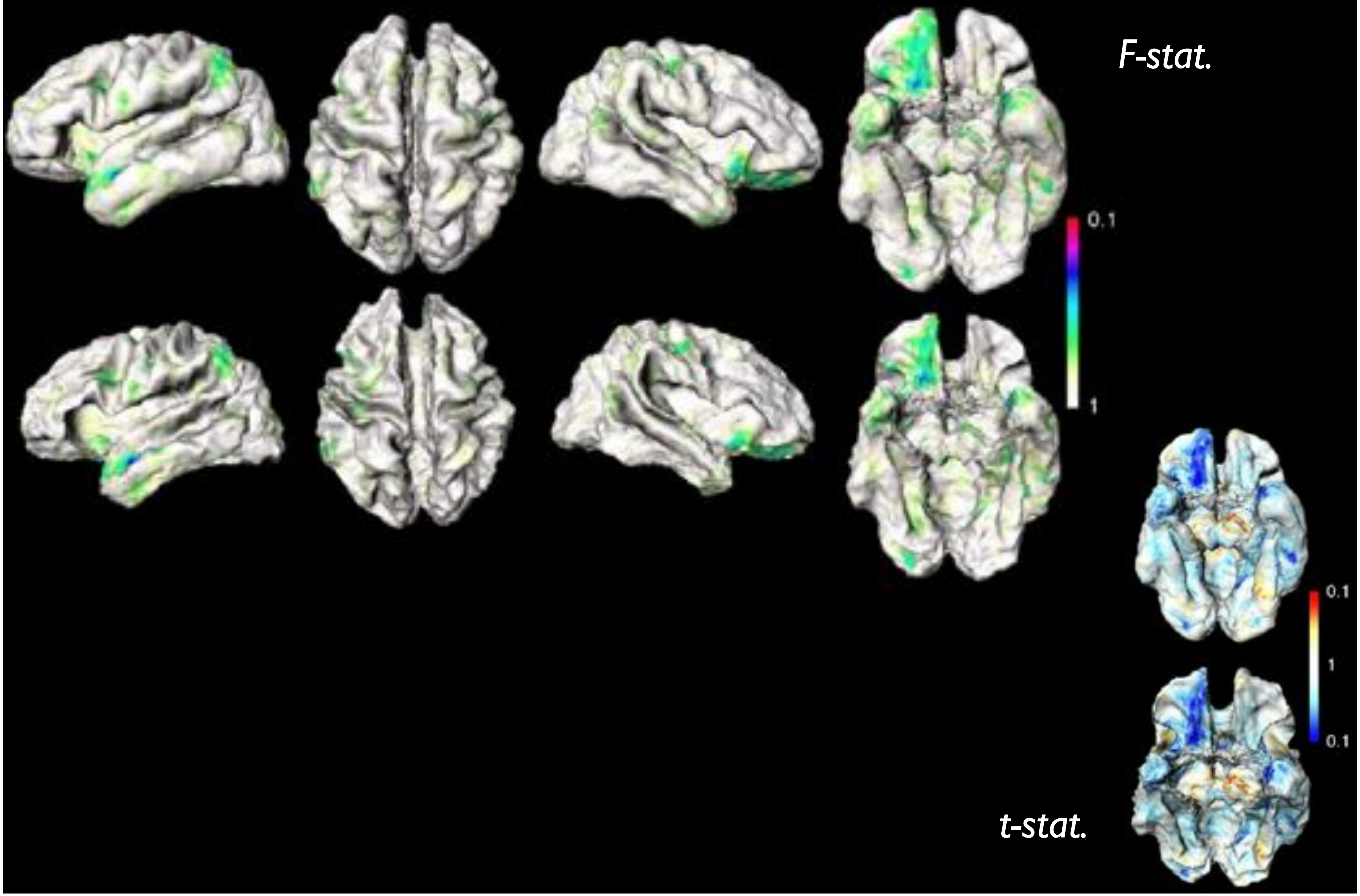
Left central & temporal sulci

Right central & temporal sulci

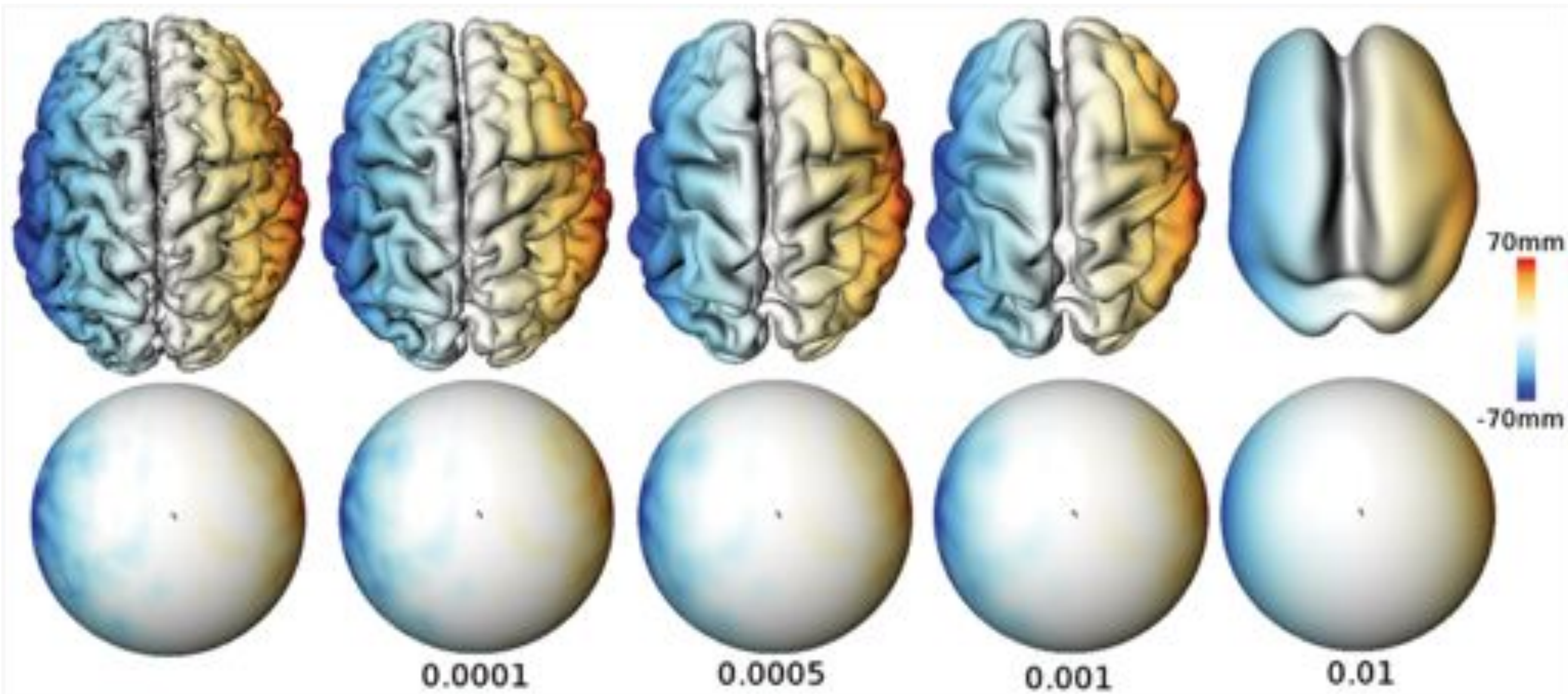


NeuroImage (2003)

Asymmetric pattern of abnormal cortical thickness in autism

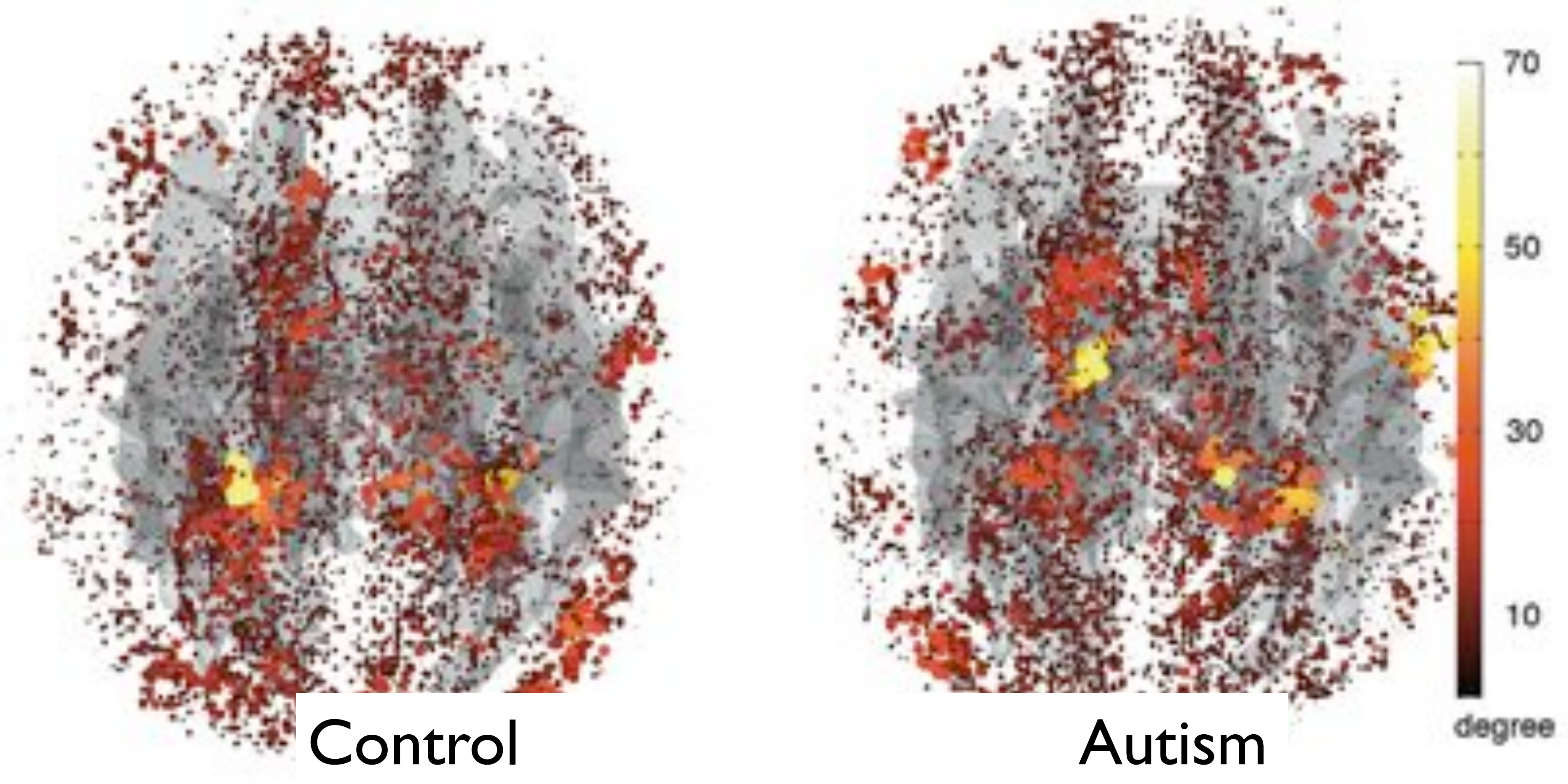


Asymmetry over multiscale



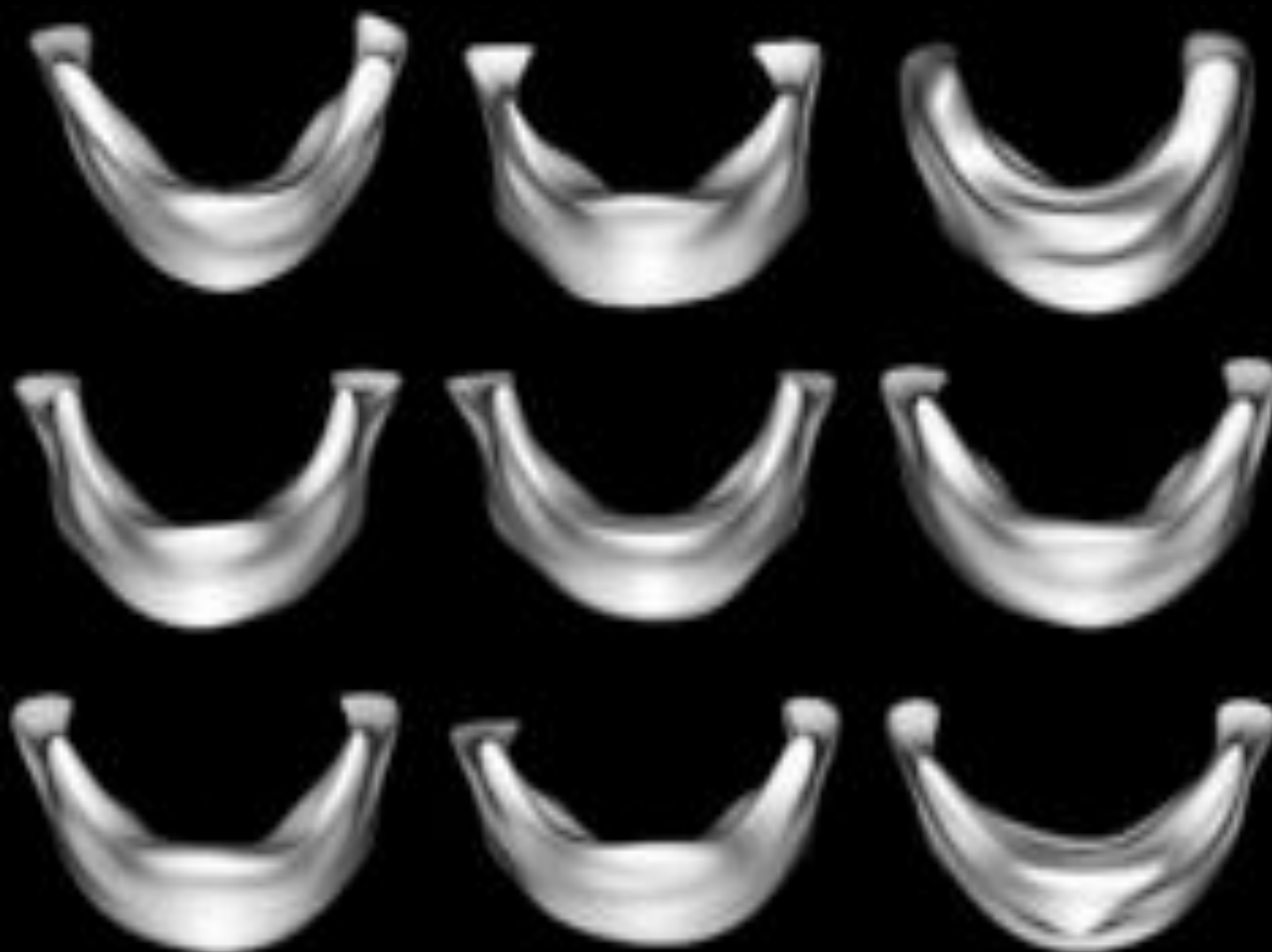
Color scale= x-coordinate

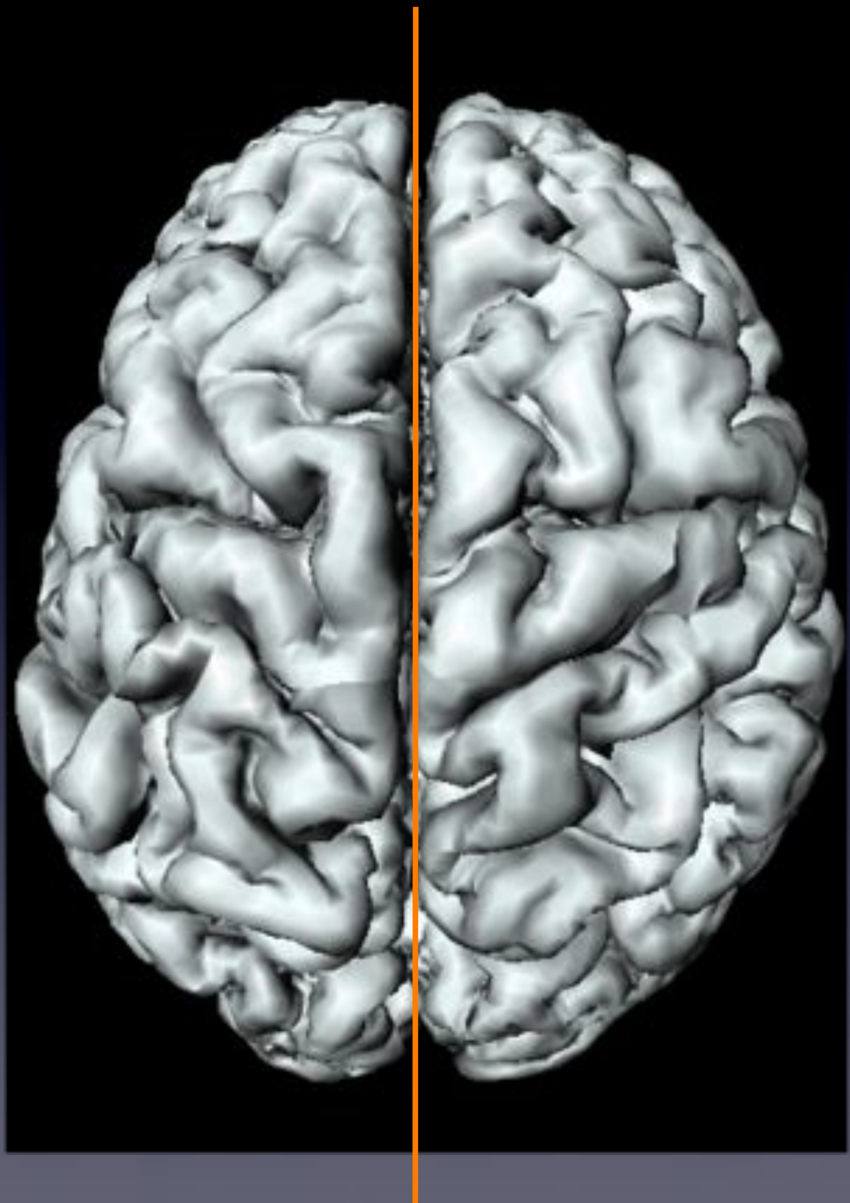
Asymmetric graph network:
Degree of nodes for all subjects



New research idea: network asymmetry analysis
You need to put the symmetry constraint in the graph modeling

Mandible left-right asymmetry





Hemispheres….
Are they symmetric ?
If not, what part is not symmetric ?

Previous 3D approach

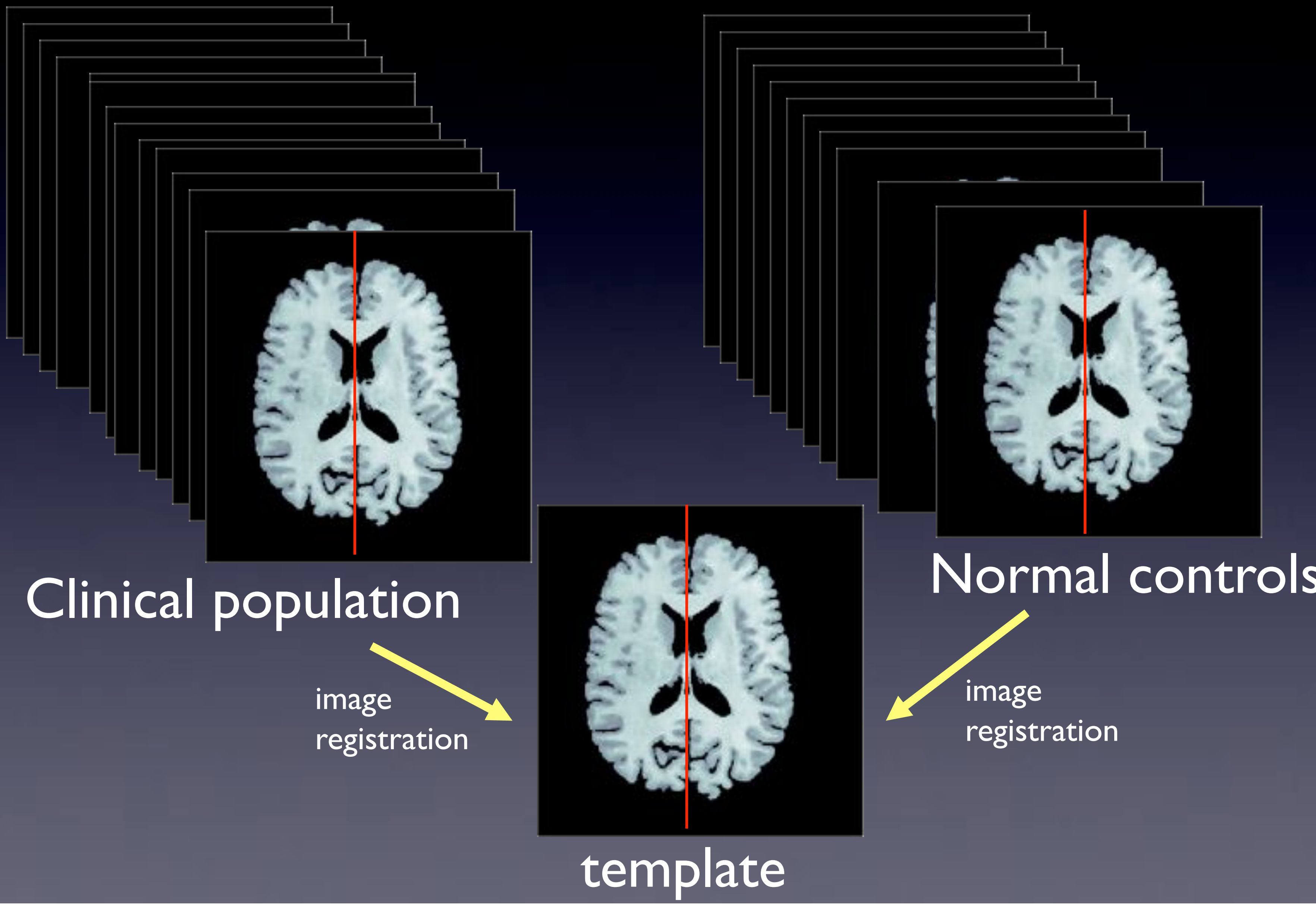
1. Image registration across subjects via a template

2. Image registration across hemispheres by registering the original MRI and its mirror reflection.

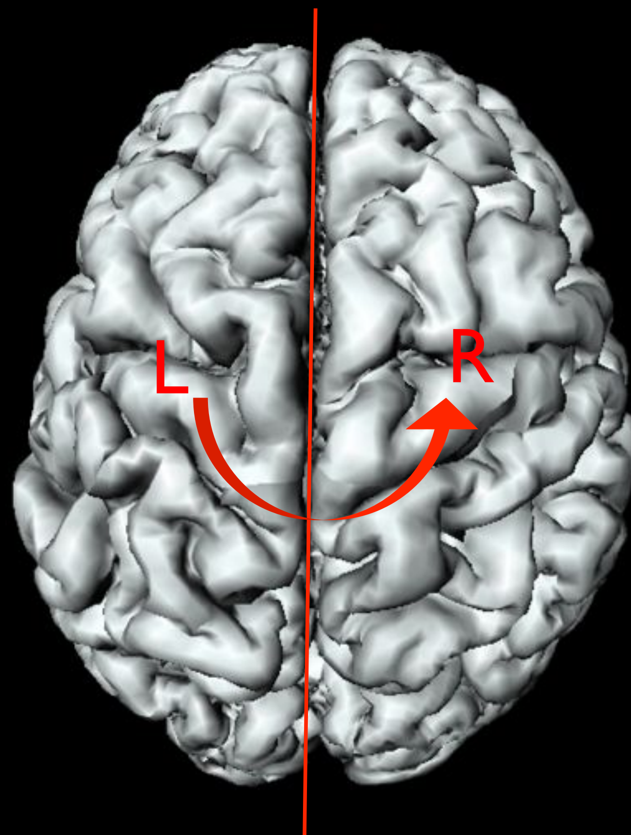
3. Construct asymmetry index at each voxel.

4. Feed the index into a statistical model.

Two population asymmetry analysis framework



Asymmetry Index



Localized asymmetry index (L-R)/(L+R)

Motivation: quantify abnormal brain structural asymmetry across hemispheres in a group of autistic subjects

Three issues with this well established 3D approach

1. 3D image registration can easily misalign sulcal pattern.

2. Mirror reflection and doing image registration is an additional computational burden.

3. The 3D approach does not work for 2D cortical surface data. New 2D framework is needed.

Related works in neuroanatomy

Surface model, parameterization → Surface registration → Surface data smoothing → Multiple comparison correction

Spherical harmonic descriptors
Guido Gerig
Martin Styner
Li Shen

PDE
Paul Thompson
Michael Miller

diffusion smoothing
(NeuroImage 2003
CVPR, 2003)
heat kernel smoothing
(NeuroImage, 2005)

Random field theory
Keith Worsley
Jonathan Taylor

Unified framework: Weighted Fourier Analysis (IEEE -TMI, 2007)

Three problems of spherical harmonic representation

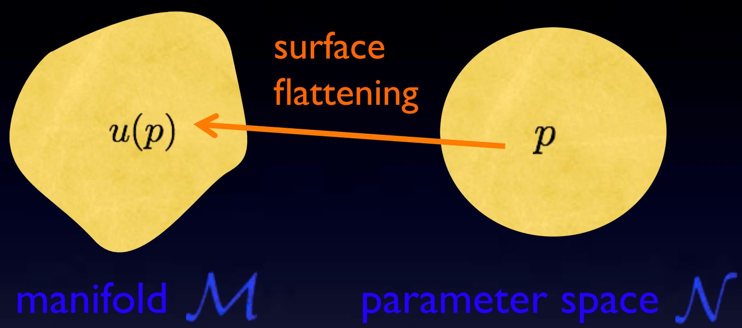
- Gibbs phenomenon (ringing artifacts)
- Computational bottleneck of solving large linear equations
- Slow convergence → Inefficient representation

(MICCAI 2008 workshop on mathematical foundations of computational anatomy)

↓

Weighted Fourier Analysis

Cortical manifold and function defined on the manifold



Anatomical manifold $\mathcal{M} \in \mathbb{R}^d$

Parameter space $\mathcal{N} \in \mathbb{R}^m$

Hilbert space $L^2(\mathcal{N})$ with inner product

$$\langle g_1, g_2 \rangle = \int_{\mathcal{N}} g_1(p) g_2(p) \mu(p)$$

Self-adjoint operator $\mathcal{L}$

$$\langle \mathcal{L} g_1, g_2 \rangle = \langle g_1, \mathcal{L} g_2 \rangle$$

Basis function

$$\mathcal{L}\psi_j = \lambda_j \psi_j$$

Weighted Fourier Series

function defined on
surface +
surface coordinates

t = scale
bandwidth
diffusion time

Self-adjoint PDE: $\boxed{\partial_t g + \mathcal{L} g = 0, g(p, t = 0) = f(p)}$

Analytic solution

Weighted Fourier Series $g(p,t) = \sum_{j=0}^{\infty} e^{-\lambda_j t} \langle f, \psi_j \rangle \psi_j(p)$

Isotropic Kernel Smoothing $= \int_{\mathcal{N}} K_t(p,q) f(q) \, d\mu(q)$

- For measurements $f(p_1), f(p_2), \cdots, f(p_n)$, $(n > 46,000)$, we set up normal equations:

$$f(p_i) = \sum_{l=0}^{k} \sum_{m=-l}^{l} \beta_{lm} Y_{lm}(p_i).$$

i-th mesh vertex

- Matrix form:

$$\underbrace{\begin{pmatrix} f(p_1) \\ f(p_2) \\ \vdots \\ f(p_n) \end{pmatrix}}_{\mathbf{F}} = \underbrace{\begin{pmatrix} Y_{00}(p_1) & Y_{1-1}(p_1) & Y_{10}(p_1) & \cdots & Y_{kk}(p_1) \\ Y_{00}(p_2) & Y_{1-1}(p_2) & Y_{10}(p_2) & \cdots & Y_{kk}(p_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{00}(p_n) & Y_{1-1}(p_n) & Y_{10}(p_n) & \cdots & Y_{kk}(p_n) \end{pmatrix}}_{\mathbf{Y}} \underbrace{\begin{pmatrix} \beta_{00} \\ \beta_{1-1} \\ \vdots \\ \beta_{kk} \end{pmatrix}}_{\beta}$$

40962 x 7000

Estimation: $\widehat{\beta} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{F}.$

Iterative residual fitting (IRF) algorithm

Scalable approach to solving a huge linear equation

Step 1. measurements $f(p_1), \cdots, f(p_n)$

Step 2. Set initial degree=0 $k = 0$

Step 3. Solve $f(p_i) = \sum_{m=-k}^{k} \beta_{km} Y_{km}(p_i)$

Project data into a finite subspace

Iterate

Step 3.5. $f \leftarrow f - \hat{f}$

Once low frequency parts are estimated, we throw them away

Step 4. Set degree $k \leftarrow k + 1$

MATLAB code available at
http://www.stat.wisc.edu/~mchung/

IEEE-TMI 2007

Weighted Fourier Analysis

Shape Asymmetry

Surface registration via WFS

Given two *l*-th degree WFS surfaces v_{i1}, v_{i2}

find the displacement d_i that minimizes the discrepancy between two surfaces:

$$v_{i2} - v_{i1} = \arg \min_{d_i \in \mathcal{H}_l} \int_{\mathcal{M}} [v_{i1} + d_i(v_{i1}) - v_{i2}]^2 \, d\mu(p).$$

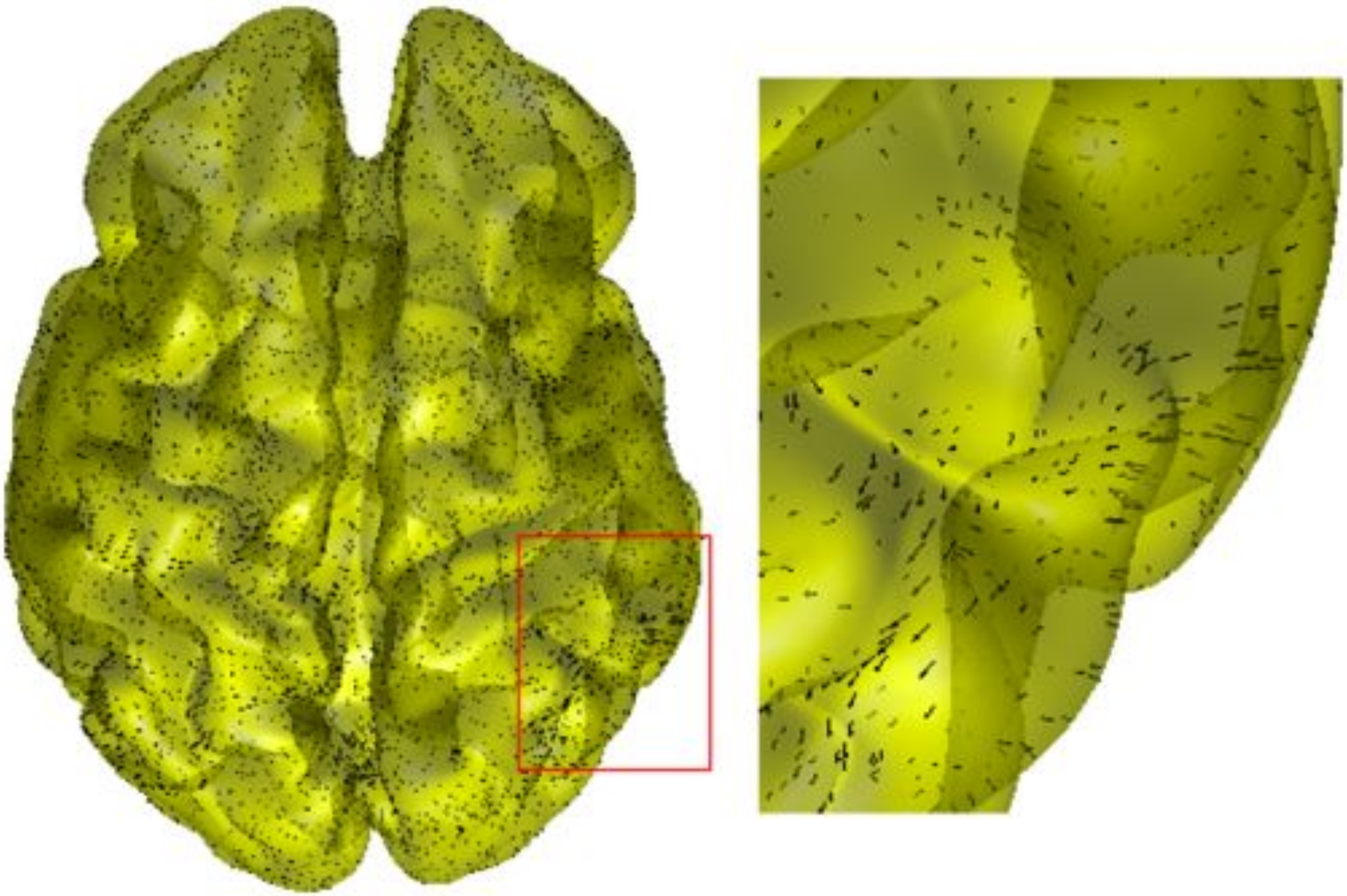
$\mathcal{H}_l$: subspace spanned by up to *l*-th degree spherical harmonics

$v_{i1} + d_i(v_{i1})$: deformation of coordinates v_{i1}

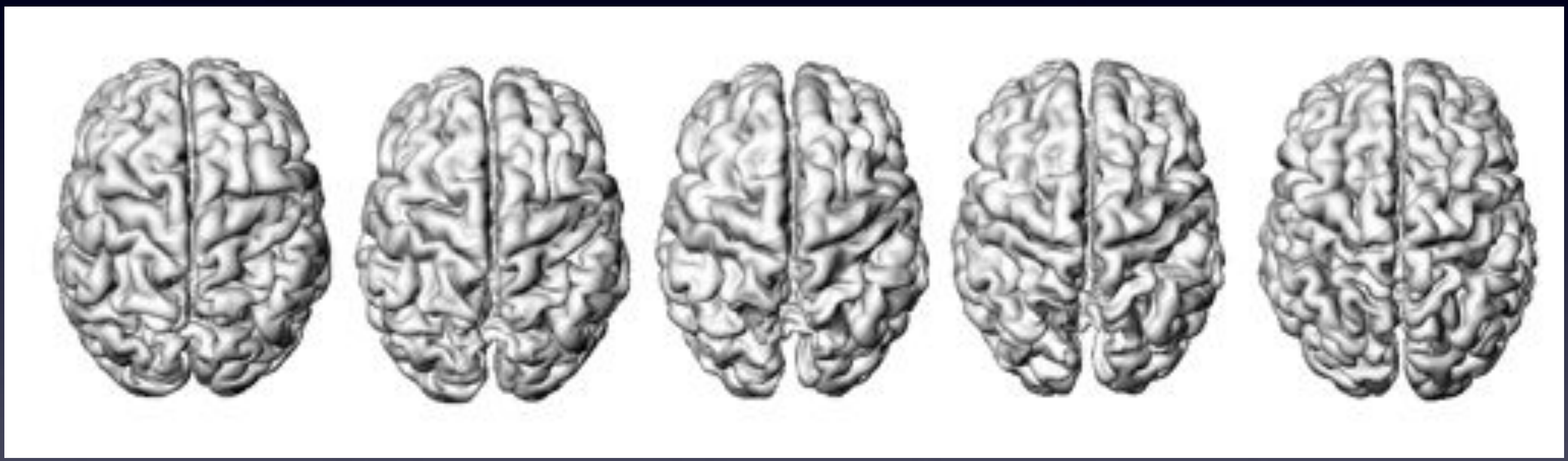
Consequence: For fixed (θ, φ) ,

$v_{i1}(\theta, \varphi)$ corresponds to $v_{i2}(\theta, \varphi)$

Subsampled surface displacement vector fields



Example of surface registration



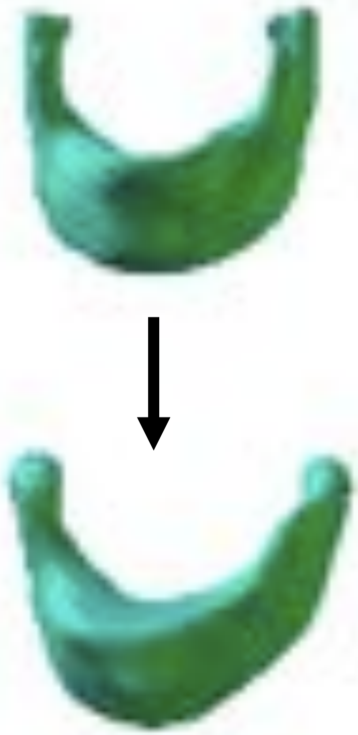
subject 1 subject 2

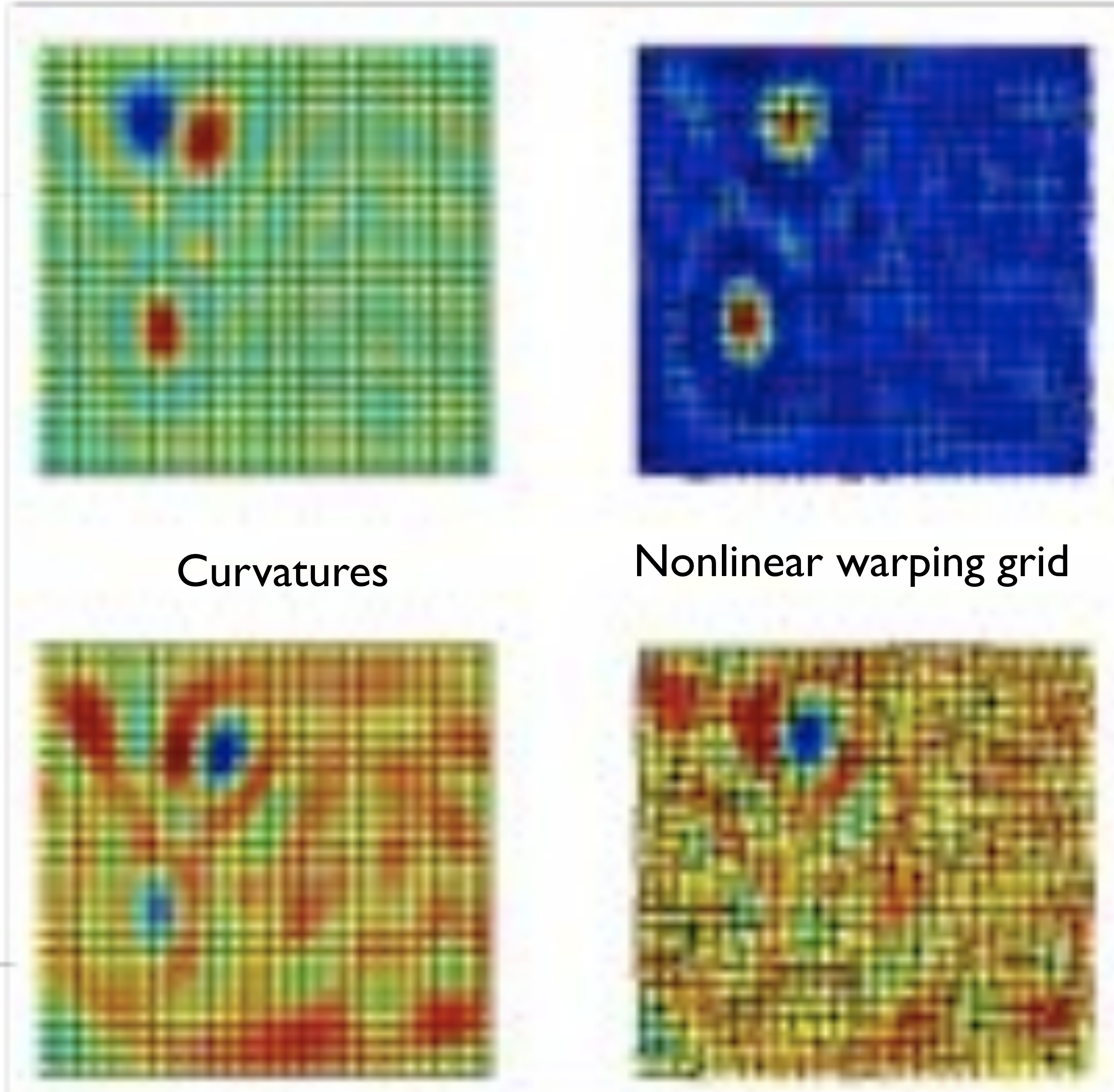
$\alpha = 0$

$$v_{i1} + \alpha d_i(v_{i1})$$

$\alpha = 1$

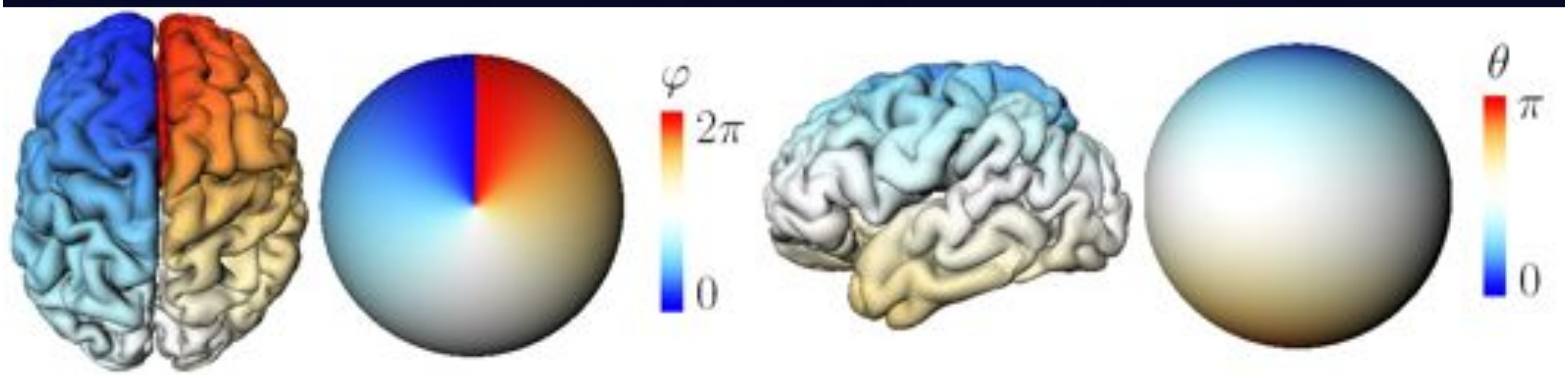
Nonlinear surface registration via curvature matching



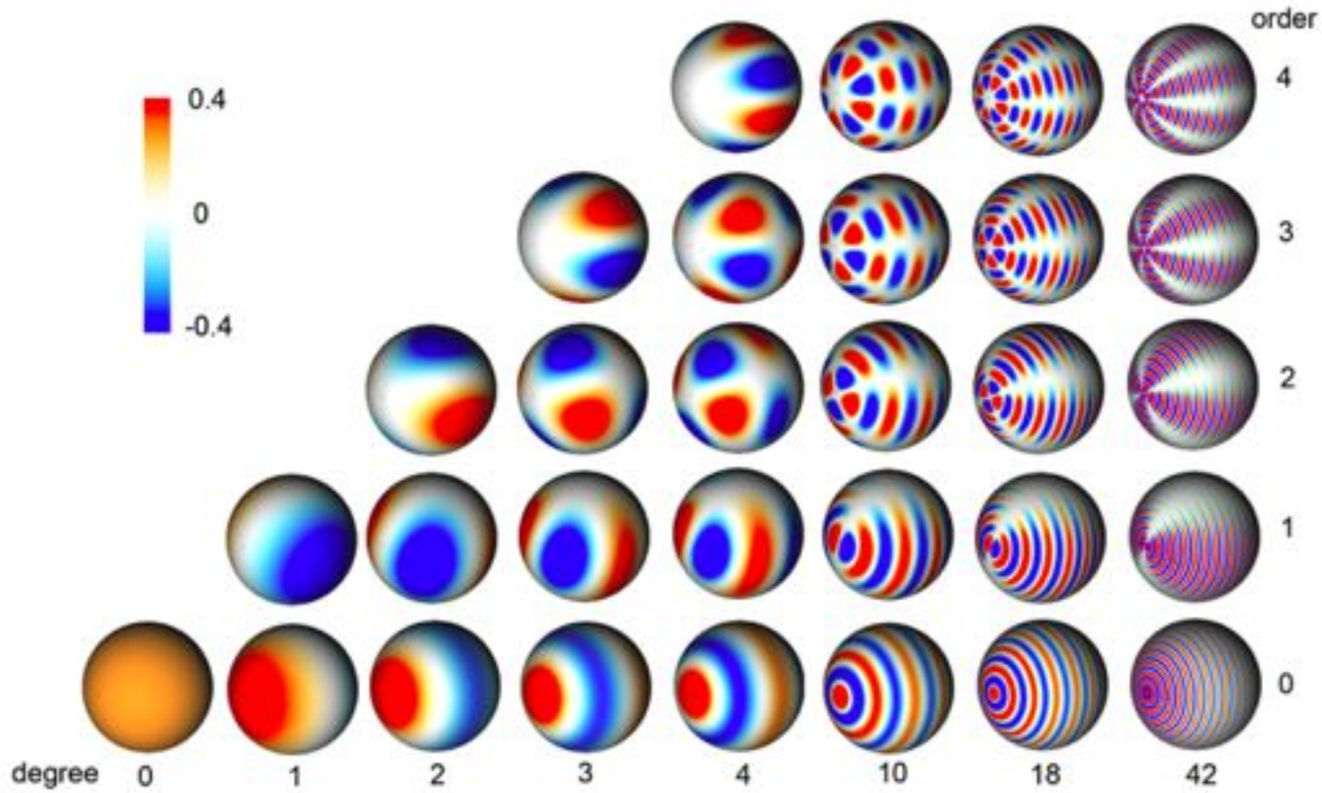


Shubing Wang 2008 PhD thesis

Decoding cortical surface asymmetry



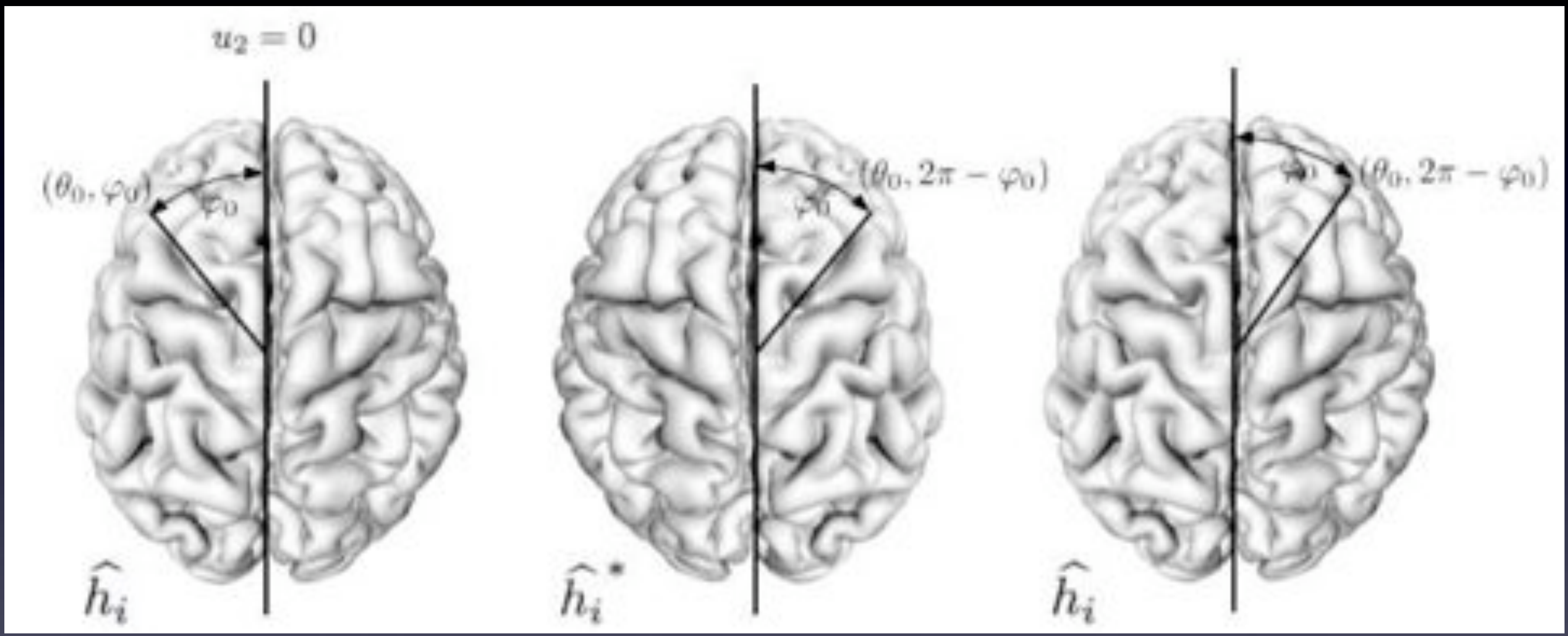
Spherical harmonic of degree *l* and order *m*



Spherical harmonics can be decomposed into symmetric & asymmetric components

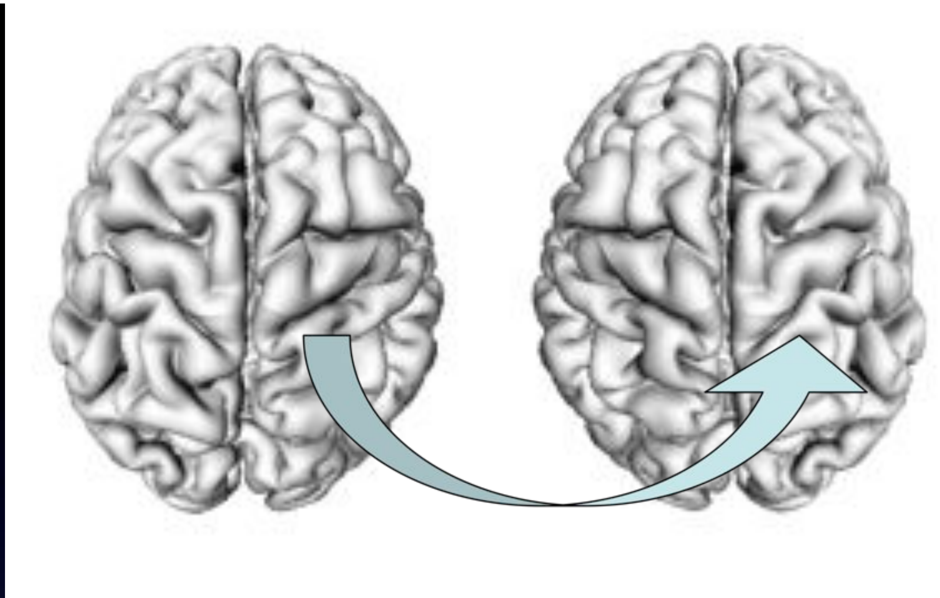
Cortical asymmetry analysis

Establishing hemispheric correspondence algebraically



Mirror reflection: It is done algebraically on WFS

Surface registration



Establishing hemispheric correspondence

What is invariant under mirror reflection ?

$$\widehat{g}(\theta,\varphi)=\sum_{l=0}^{k}\sum_{m=-l}^{l}e^{-l(l+1)\sigma}\langle f,Y_{lm}\rangle Y_{lm}(\theta,\varphi).$$

$\updownarrow$ Mirror reflection

$$\widehat{g}(\theta,2\pi-\varphi) = \sum_{l=0}^{k}\sum_{m=-l}^{-1}e^{-l(l+1)\sigma}\langle f,Y_{lm}\rangle Y_{lm}(\theta,\varphi)$$

Symmetric Part

$$-\sum_{l=0}^{k}\sum_{m=0}^{l}e^{-l(l+1)\sigma}\langle f,Y_{lm}\rangle Y_{lm}(\theta,\varphi)$$

Antisymmetric part

Shape decomposition into symmetric and asymmetric parts

$$S(\theta,\varphi) = \frac{1}{2}\Big[\widehat{g}(\theta,\varphi) + \widehat{g}(\theta,2\pi-\varphi)\Big] = \sum_{l=0}^{k}\sum_{m=-l}^{-1} e^{-l(l+1)\sigma}\langle f, Y_{lm}\rangle Y_{lm}(\theta,\varphi)$$

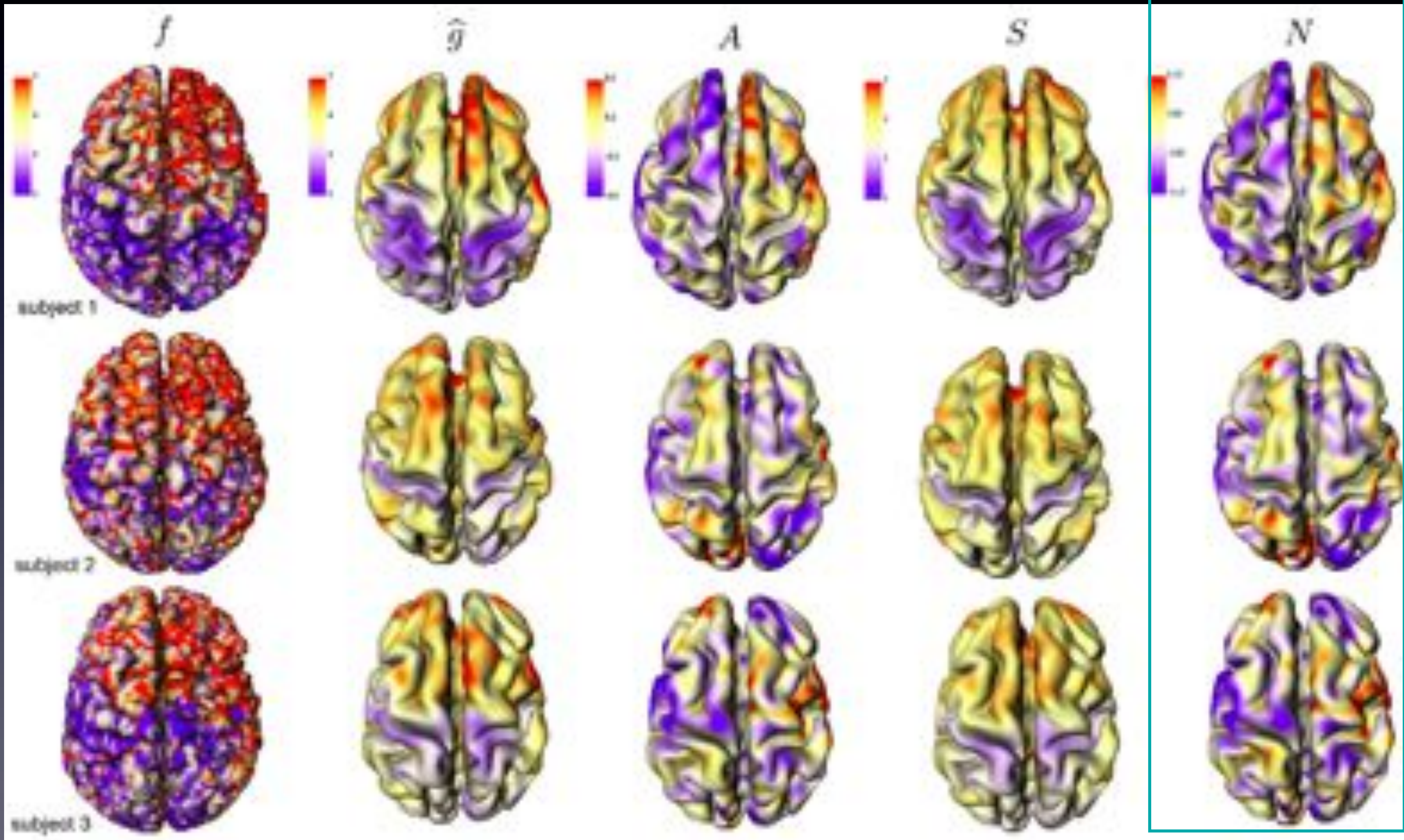
$$A(\theta,\varphi) = \frac{1}{2}\Big[\widehat{g}(\theta,\varphi) - \widehat{g}(\theta,2\pi-\varphi)\Big] = \sum_{l=0}^{k}\sum_{m=0}^{l} e^{-l(l+1)\sigma}\langle f, Y_{lm}\rangle Y_{lm}(\theta,\varphi)$$

(L-R)/(L+R) $\downarrow$ Normalized asymmetry index

$$N(\theta,\varphi) = \frac{\widehat{g}(\theta,\varphi) - \widehat{g}(\theta,2\pi-\varphi)}{\widehat{g}(\theta,\varphi) + \widehat{g}(\theta,2\pi-\varphi)} = \frac{\sum_{l=1}^{k}\sum_{m=-l}^{-1} e^{-1(l+1)\sigma}\langle f, Y_{lm}\rangle Y_{lm}(\theta,\varphi)}{\sum_{l=0}^{k}\sum_{m=0}^{l} e^{-l(l+1)\sigma}\langle f, Y_{lm}\rangle Y_{lm}(\theta,\varphi)}$$

Asymmetry Index on Cortical Thickness

Cortical thickness → Weighted SPHARM → Asymmetry index → Symmetry index → Normalized asymmetry index



Discriminant power approah

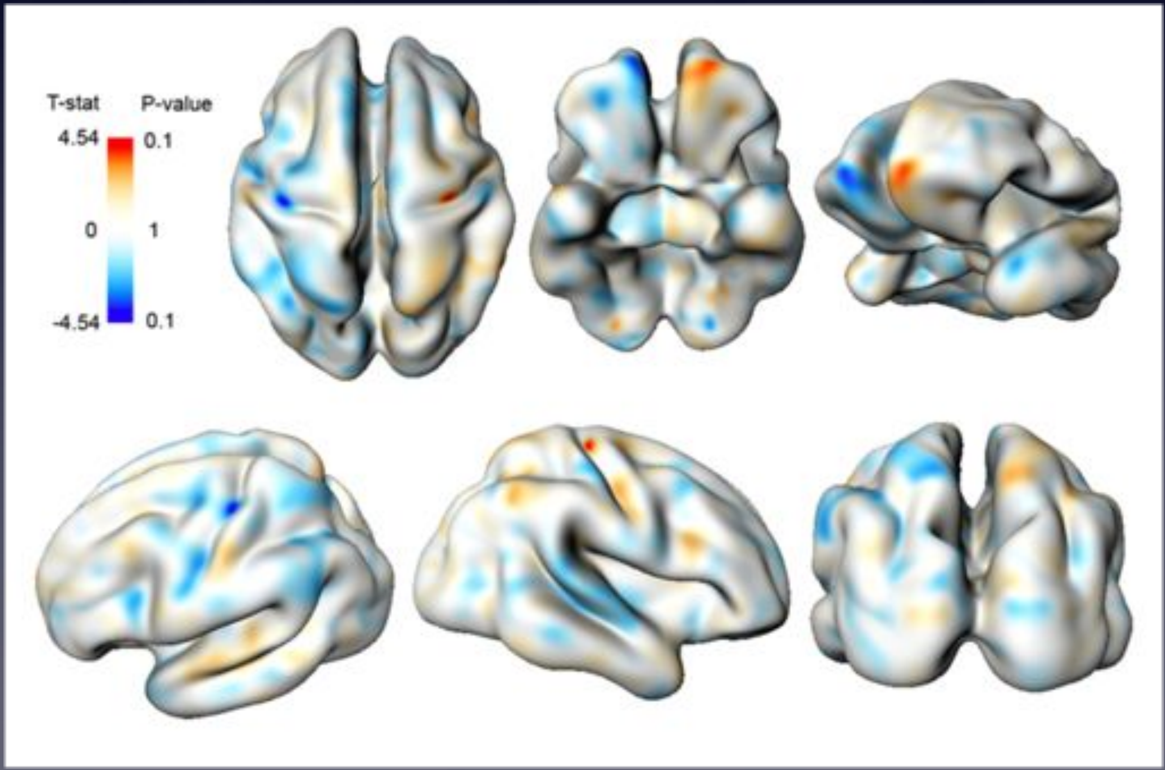
Statistical Parametric Map

multiple comparison correction via the random field theory (Worsley et al. 1995) → not so trivial

$$P(\sup_{p \in \partial\Omega} Z(p) > h) \approx \sum_{d=0}^{2} \phi_d(\partial\Omega)\rho_d(h)$$

> 40000 correlated hypotheses

Very involving mathematical derivation



T-stat resulting showing group difference between autism and control

T random field on manifolds

$$P\Big(\max_{\mathbf{x}\in\partial\Omega_{atlas}} T(\mathbf{x}) \geq y\Big) \approx 2\rho_0(y) + \|\partial\Omega_{atlas}\|\rho_2(y)$$

Euler characteristic density

$$\rho_0(y) = \int_y^\infty \frac{\Gamma(\frac{n}{2})}{((n-1)\pi)^{1/2}\Gamma(\frac{n-1}{2})}\Big(1+\frac{y^2}{n-1}\Big)^{-n/2} dy,$$

$$\rho_2(y) = \frac{1}{\mathrm{FWHM}^2}\frac{4\ln 2}{(2\pi)^{3/2}}\frac{\Gamma(\frac{n}{2})}{(\frac{n-1}{2})^{1/2}\Gamma(\frac{n-1}{2})}y\Big(1+\frac{y^2}{n-1}\Big)^{-(n-2)/2}$$

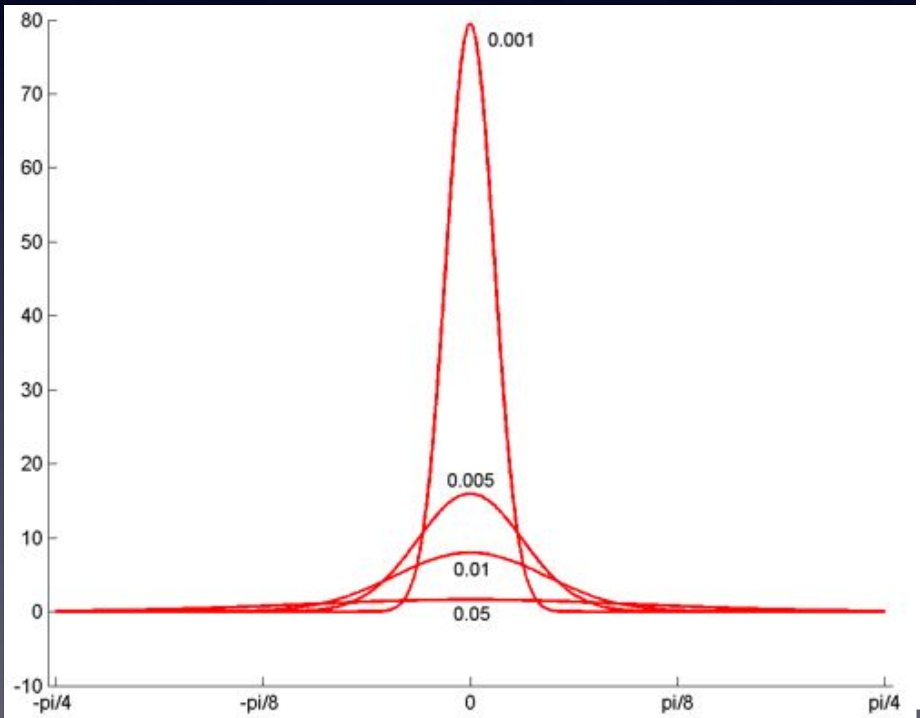
Worsley (1995, NeuroImage)

FWHM of smoothing kernel or residual field

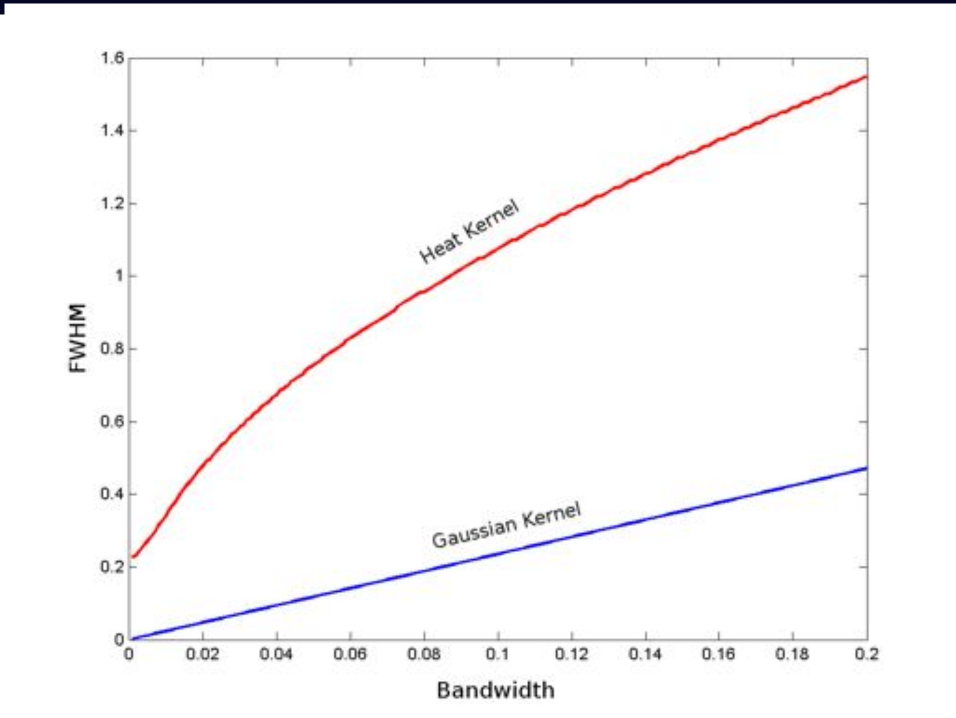
WFS is related to heat kernel smoothing

WFS $$g(p,t) = \sum_{j=0}^{\infty} e^{-\lambda_j t} \langle f, \psi_j \rangle \psi_j(p)$$

Heat kernel smoothing $$= \int_{\mathcal{N}} K_t(p,q) f(q) \, d\mu(q)$$



Shape of heat kernel



Numerical computation

Hypothesis & P-value free approach

Discriminant Power Map

Logistic model

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 n_i$$

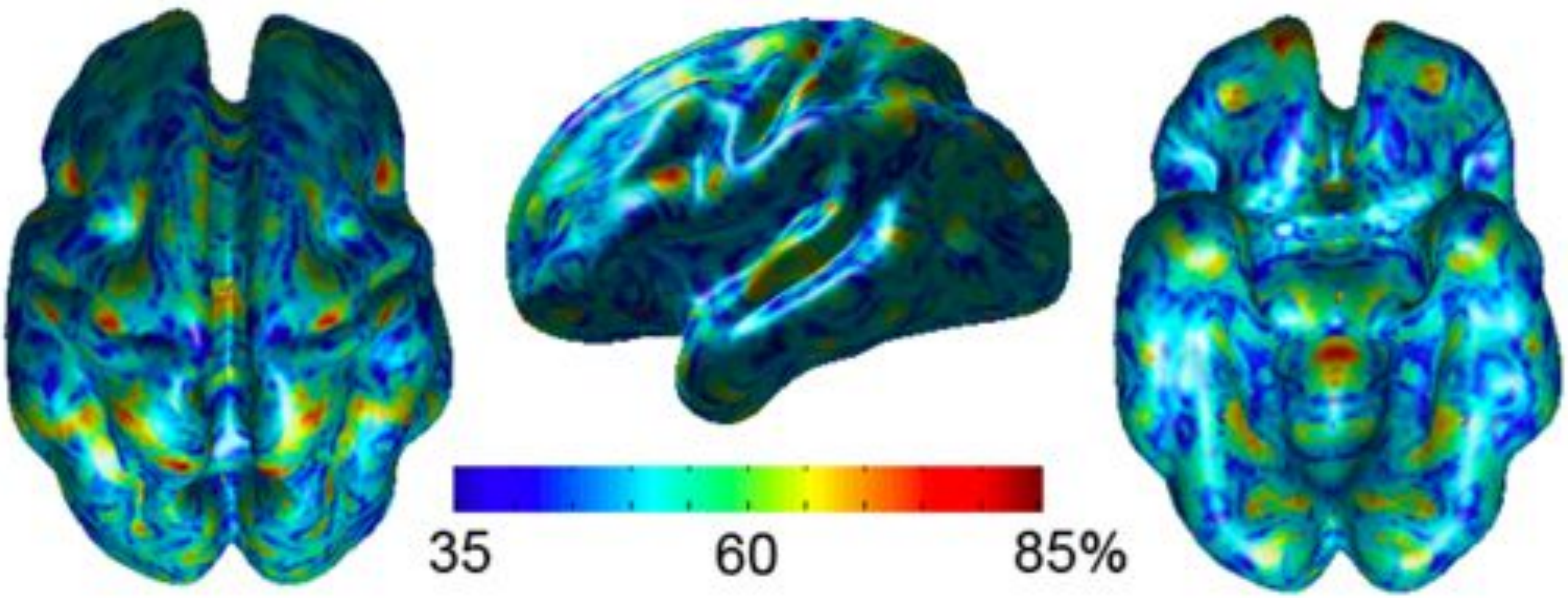
Probability of autism

Asymmetry index

Classification rule: $\pi_i > \frac{1}{2}$ $\xrightarrow{\text{Leave-one-out cross-validation}}$ Classification error rate

Discriminant Power Map

= 1 - error rate



Avoid the traditional hypothesis driven approach

No need to compute P-value → No need for random field theory

Lecture 11 Topics

Numerical optimization
Compressed sensing
Covariance matrix estimation

Read

figueiredo2007.compressed.sensing.pdf
lee.2011.TMI.pdf
peng.2009.jasa….