Topic Modeling of Multimodal Data: Latent Dirichlet Allocation of functional MRI, Structural MRI, and Phenotypic Measures of ADHD

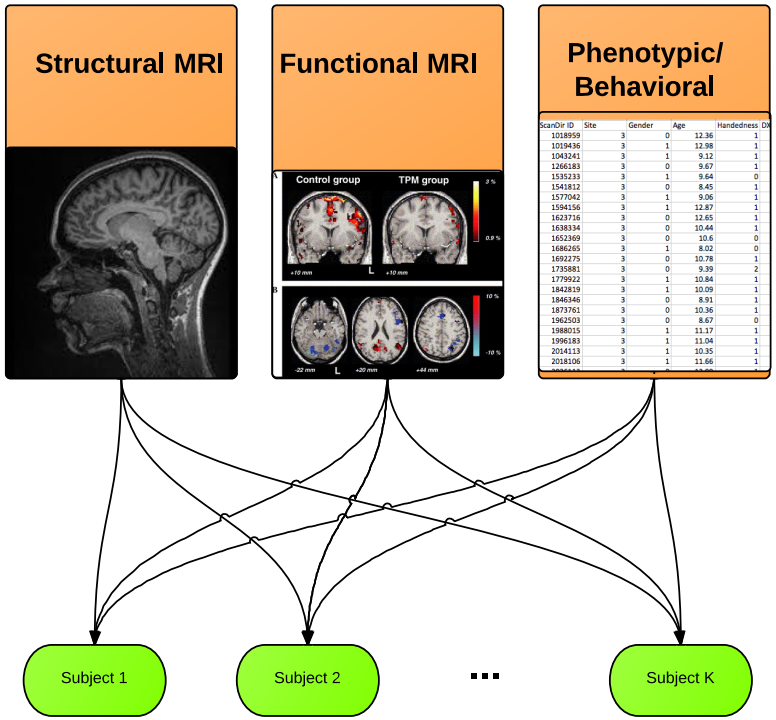
Ariana Anderson, Pamela K. Douglas, Mark S. Cohen

University of California, Los Angeles
Departments of Psychiatry and Behavioral Sciences
Laboratory of Integrative Neuroimaging Technology

March 6, 2013

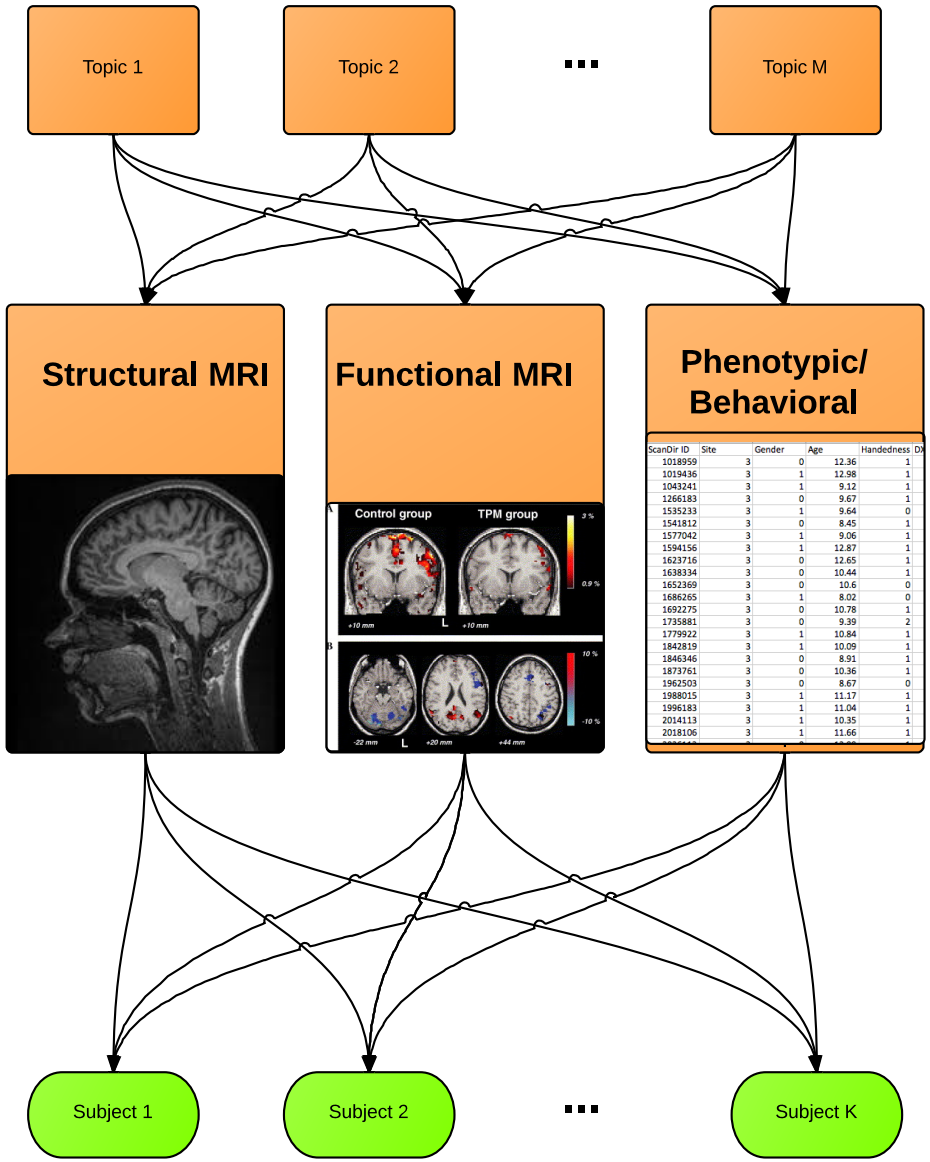
Topic Modeling of Multimodal Data: Motivation

- When data are from different modalities and are from inherently different spaces, how do we meaningfully group features to identify any similar dimensions, or common information, that they describe?



Topic Modeling of Multimodal Data: Motivation

- When data are from different modalities and are from inherently different spaces, how do we meaningfully group features to identify any similar dimensions, or common information, that they describe?
- We present Latent Dirichlet Allocation (LDA), or topic modeling, as a method of categorizing, dimension reducing, and interpreting multimodal data.
- As opposed to traditional *supervised* machine-learning (ML) models where features are used to classify certain states (ADHD vs. healthy controls), we use an *unsupervised* Bayesian generative model to propose clusters, or topics, of multimodal features.


Topic 1
Topic 2
...
Topic M
Structural MRI
Functional MRI
Phenotypic/ Behavioral
Control group
TPM group
ScanDir ID Site Gender Age Handedness DX
1018959 3 0 12.36 1
1019436 3 1 12.98 1
1043241 3 1 9.12 1
1266183 3 0 9.67 1
1535233 3 1 9.64 0
1541812 3 0 8.45 1
1577042 3 1 9.06 1
1594156 3 1 12.87 1
1623716 3 0 12.65 1
1638334 3 0 10.44 1
1652369 3 0 10.6 0
1686265 3 1 8.02 0
1692275 3 0 10.78 1
1735881 3 0 9.39 2
1779922 3 1 10.84 1
1842819 3 1 10.09 1
1846346 3 0 8.91 1
1873761 3 0 10.36 1
1962503 3 0 8.67 0
1988015 3 1 11.17 1
1996183 3 1 11.04 1
2014113 3 1 10.35 1
2018106 3 1 11.66 1
Subject 1
Subject 2
...
Subject K

- LDA proposes topics, or groups of multimodal observations, which we can posit describe the same latent *dimension* in the data, thus answering the question of how features from across modalities relate to each other.
- Using the ADHD-200 competition dataset as our marker, we present our results from *unsupervised* topic-modeling and discuss how they relate to previously-published *supervised* classification models.

LDA vs. ML

- Machine learning models map multimodal features to a common categorical variable. Feature selection algorithms operate within ML algorithms to identify subsets of features which are both *non-redundant* and mutually distinct of each other, yet *predictive* of the outcome variable of interest.
- LDA maps similar multimodal features to a common metric space, or "topics", though a latent generative model.

Applications of LDA

- Feature Selection: Are there features which map to the same topic space as the categorical variable of interest?
- Dimension Reduction: A single subject's multimodal observations can be described as the topics they contain, a sparse representation of the generative process used to construct that patient.
- Feature Interpretation: Which multimodal features are similar?

Topic Models in Computational Linguistics:

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Topic Modeling (Latent Dirichlet Allocation), Blei et al., 2003

http://www.psychology.adelaide.edu.au/personalpages/staff/simondennis/LexicalSemantics/BleiNgJordan03.pdf

Terminology:

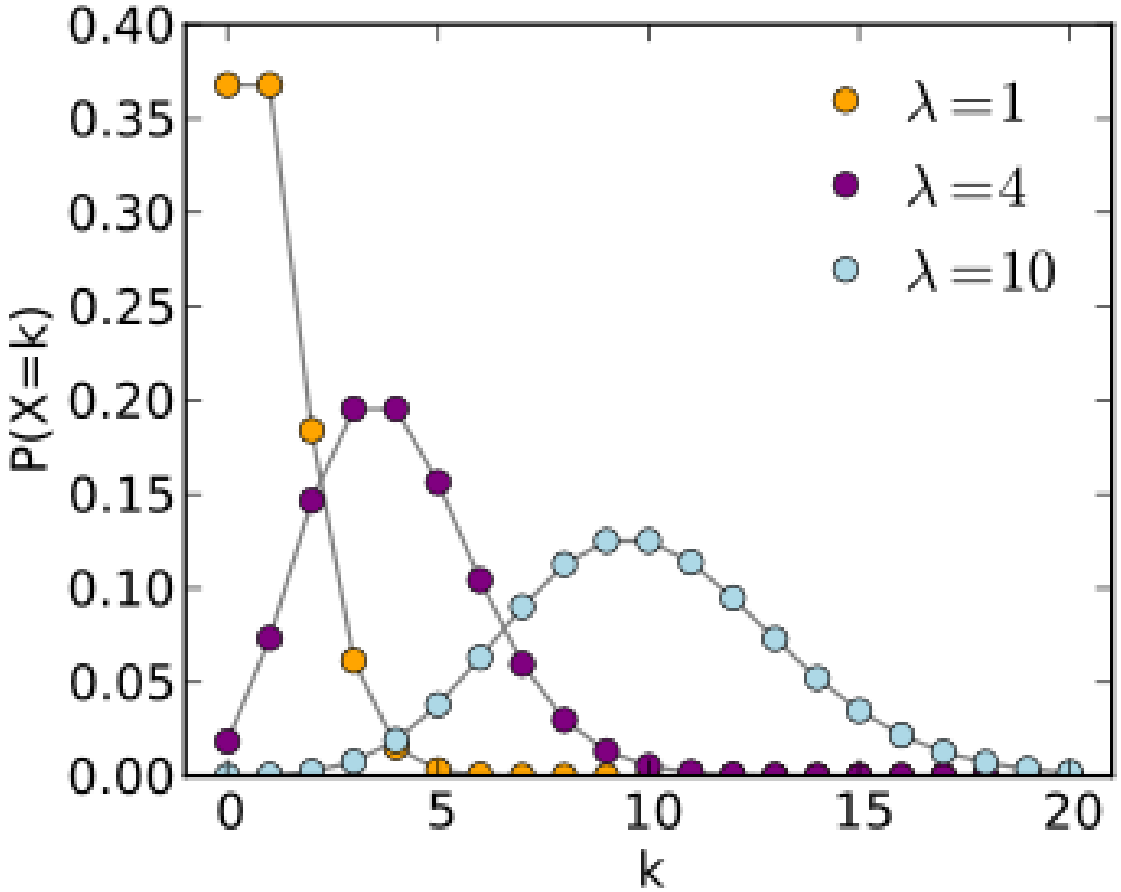
- *Word*: A vocabulary indexed from $1, \ldots, V$. The $v-th$ word is represented by a vector $w^v = 1$ and $w^u = 0$ for $u \neq v$.
- *Document*: A sequence of N words, $\mathbf{w} = (w_1, w_2, \ldots, w_N)$.
- *Corpus*: A collection of M documents denoted by $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$

Latent Dirichlet Allocation: A generative probabilistic model

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$.
3. For each of the N words w_n:
   1. Choose a topic $z_n \sim Multinomial(\theta)$
   2. Choose a word w_n from $p(w_n|z_n, \beta)$ a multinomial probabilistic model conditioned on the topic z_n

$$[\beta]_{kxV}, \beta_{ij} = p(w^j = 1|z^i = 1)$$

Poisson Distribution for Number of Topics



$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$E(X) = Var(X) = \lambda$$

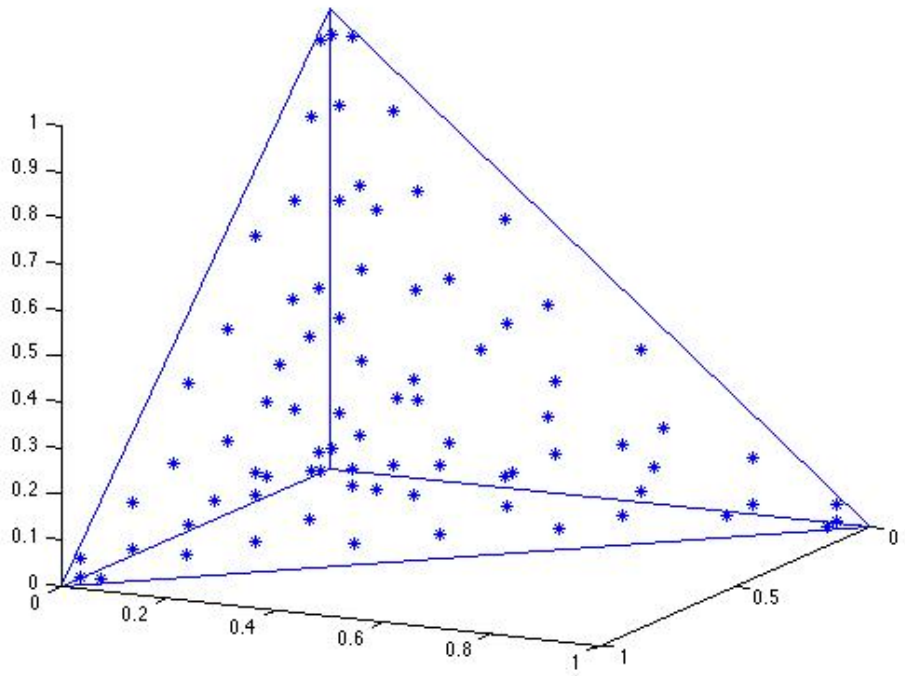
Poisson Distribution in Real Life:

In 1898, Ladislaus Bortkiewicz discovered the number of fatal horse kicks in the Prussian army followed a Poisson distribution.

Dirichlet Distribution for θ

θ is k-dimensional Dirichlet variable in the $(k-1)$ simplex when

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_i-1}\cdots\theta_k^{\alpha_k-1}$$



Simplex: $\theta \in Simplex_k \Rightarrow \theta_i \geq 0, \sum_{i=1}^{k}\theta_i = 1$

Joint Distribution

$$p(\theta, \mathbf{z}, \mathbf{w}, \alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

given α and β, the joint distribution of topic mixture θ, N total topics $\mathbf{z}$, and N total words $\mathbf{w}$ where $p(z_n|\theta) = \theta_i$ for i such that $z_n^i = 1$.

Marginal Distributions

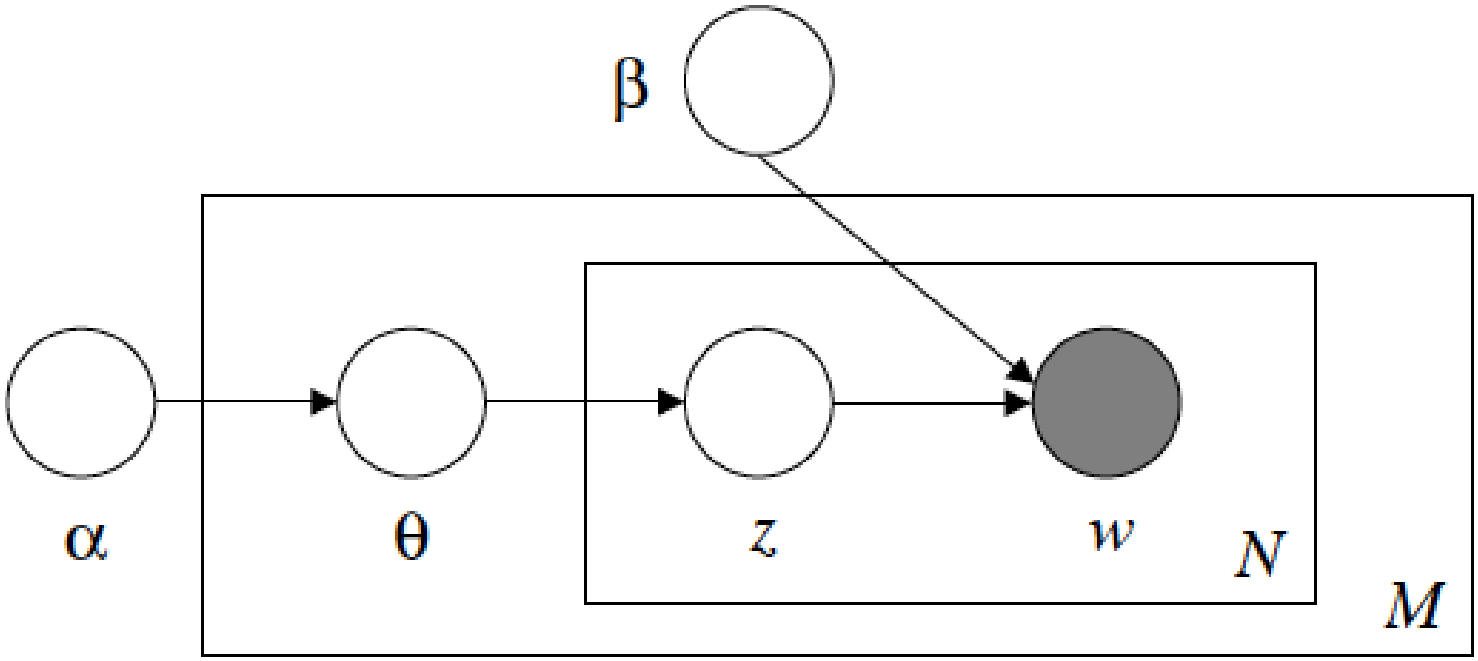
$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\Big(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\Big)\,d\theta$$

$$p(\theta,w,z|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\beta)$$

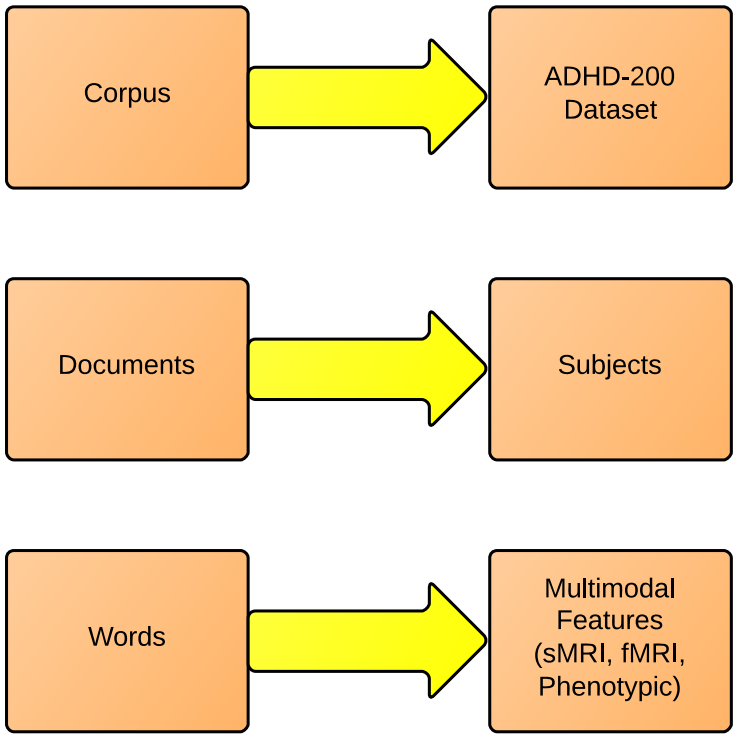
Marginal Corpus Distribution for M topics

$$p(D|\alpha,\beta) = \prod_{i=1}^{M} \int p(\theta_d|\alpha)\Big(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\Big)\, d\theta_d$$

Plate Notation of Latent Dirichlet Allocation



Mapping Multimodal Data to the "Document/Word" Framework



ADHD-200: Motivations

`http://fcon_1000.projects.nitrc.org/indi/adhd200/`

- Attention Deficit Hyperactivity Disorder (ADHD) affects at least 5-10% of school-age children and is associated with substantial lifelong impairment, with annual direct costs exceeding $36 billion/year in the US.
- Despite a voluminous empirical literature, the scientific community remains without a comprehensive model of the pathophysiology of ADHD.
- Further, the clinical community remains without objective biological tools capable of informing the diagnosis of ADHD for an individual or guiding clinicians in their decision-making regarding treatment.

ADHD-200: Data

- 776 resting-state fMRI and anatomical datasets aggregated across 8 independent imaging sites, 491 of which were obtained from typically developing individuals and 285 in children and adolescents with ADHD (ages: 7-21 years old).
- Accompanying phenotypic information includes: diagnostic status, dimensional ADHD symptom measures, age, sex, intelligence quotient (IQ) and lifetime medication status. Preliminary quality control assessments (usable vs. questionable) based upon visual timeseries inspection are included for all resting state fMRI scans.

ADHD-200: Competition

- The competition invited participants to develop diagnostic classification tools for ADHD diagnosis based on functional and structural magnetic resonance (MRI) of the brain. Applying their tools, participants provided diagnostic labels for previously unlabeled datasets.
- The competition assessed diagnostic accuracy of each submission and invited research papers describing novel, neuroscientific ideas related to ADHD diagnosis.
- Twenty-one international teams, from a mix of disciplines, including statistics, mathematics, and computer science, submitted diagnostic labels, with some trying their hand at imaging analysis and psychiatric diagnosis for the first time.

ADHD-200: Results

- The average prediction accuracy across teams was 49.79%.
- Percentage prediction accuracy ranged between 43.08% and 61.54% (mean = 56.02%) when using a two-class classifier to classify TDC vs. ADHD, disregarding the ADHD subtypes.
- Teams were better at predicting TDC (mean accuracy = 71.77%), compared to predicting ADHD regardless of subtype (mean accuracy = 37.44%).
- When teams correctly predicted an ADHD diagnosis, they correctly predicted the subtype in 62.79% of the cases. Of those, 79% were ADHD-1, and 21% were ADHD-3.
- In 37.21% of cases teams correctly predicted ADHD, but assigned an incorrect subtype.
- Considering that 55% of the TestSet were TDC, 30% ADHD-1, and 15% ADHD-3, teams overly favored TDC (67.45%) over ADHD-1 (22.09%) and ADHD-3 (10.46%) in their predictions.

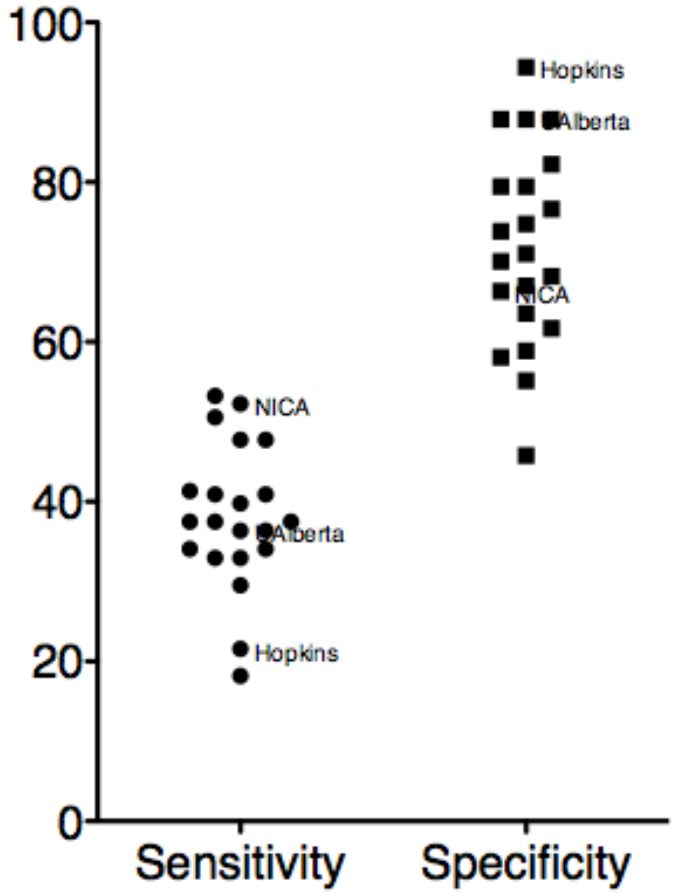
Disqualification

"The team from the University of Alberta did not use imaging data for their prediction model. This was not consistent with intent of the competition. Instead they used only age, sex, handedness, and IQ. **However, in doing so they obtained the most points, outscoring the team from Johns Hopkins University by 5 points, as well as obtaining the highest prediction accuracy (62.52%).**"

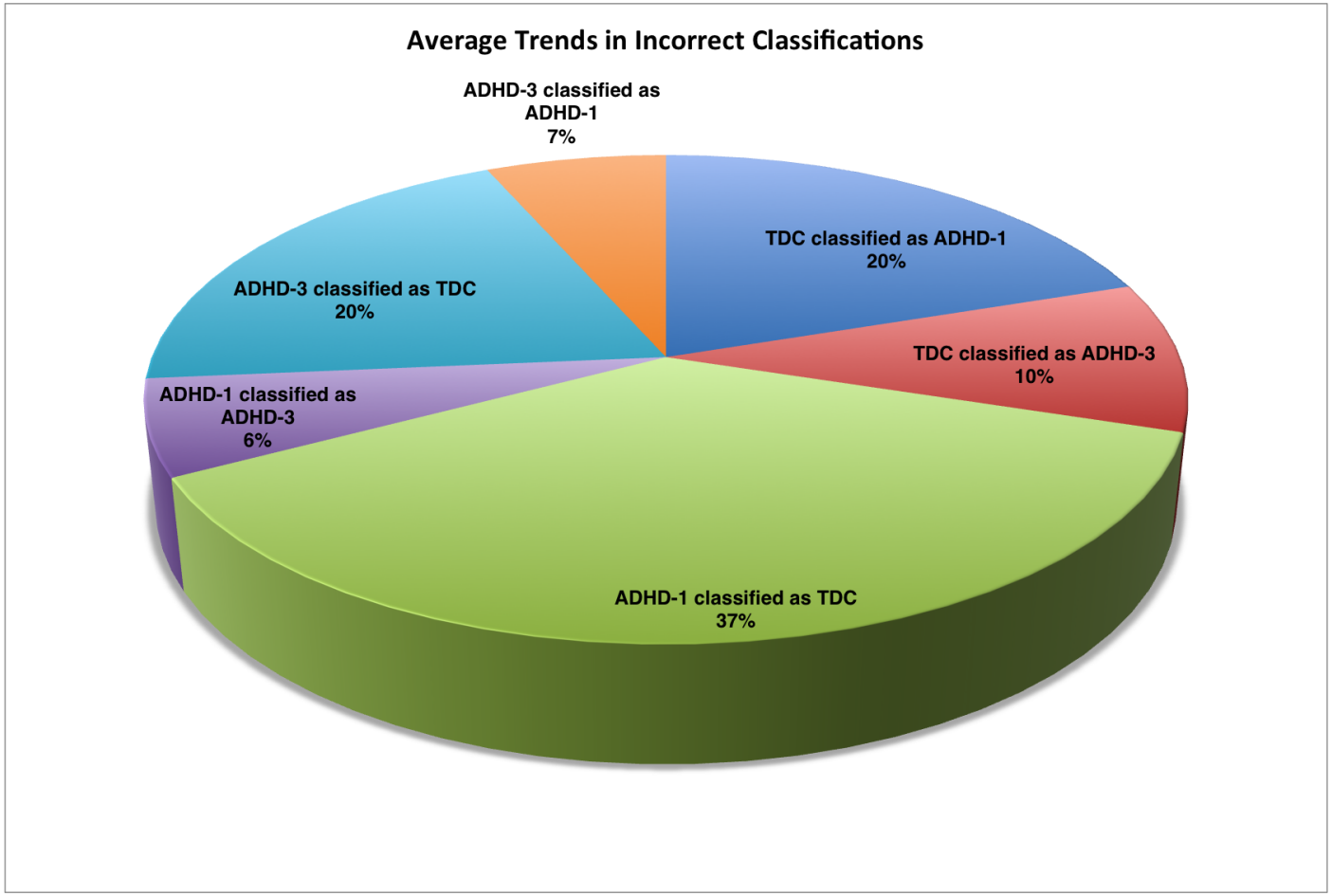
`http://www.talyarkoni.org/blog/2011/10/12/brain-based-prediction-of-adhd-now-with-100-fewer-brains/`

Alberta Response

"For the record, we tried a pile of imaging-based approaches. As a control, we also did classification with age, gender, etc. but no imaging data. It was actually very frustrating for us that none of our imaging-based methods did better than the no imaging results. It does raise some very interesting issues."
Matthew Brown, Project Manager, University of Alberta


100
80
60
40
20
0
Hopkins
Alberta
NICA
NICA
Alberta
Hopkins
Sensitivity
Specificity

ADHD-200 Competition



Phenotypic Features used for Topic Modeling

- Diagnosis: Typically-developing, ADHD-Combined, ADHD-Hyperactive/Impulsive, ADHD-Impulsive
- Handedness (Left/Right/Ambidextrous)
- Gender
- IQ Scores/Instrument
- ADHD Behavioral Measures/Instrument
- Medication Status

Imaging Features

- Structural MRI (681): Cortical and Subcortical measurements obtained from Freesurfer
- Functional MRI: Motion Parameters (12), Number of ICs (1), Connectivity Measures (546)- functional graph theory measures based on the Greicius functional atlas. The diversity coefficient is a measure of how many different modules a given node has connections to.

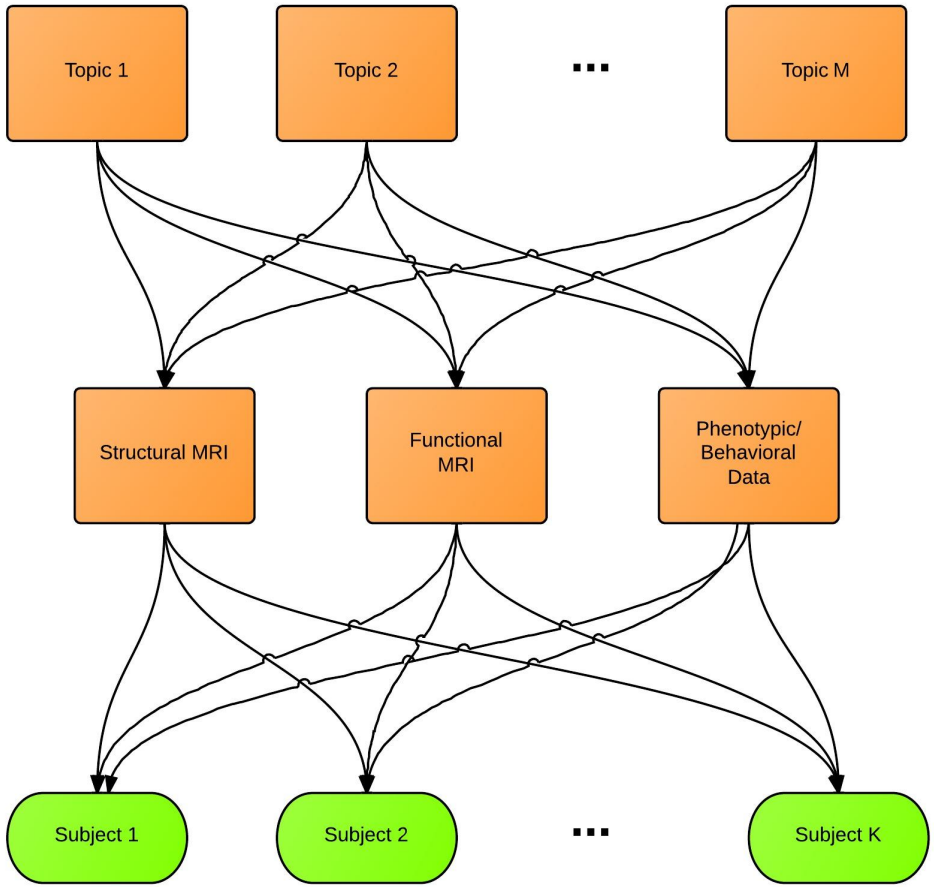
Patient Demographics

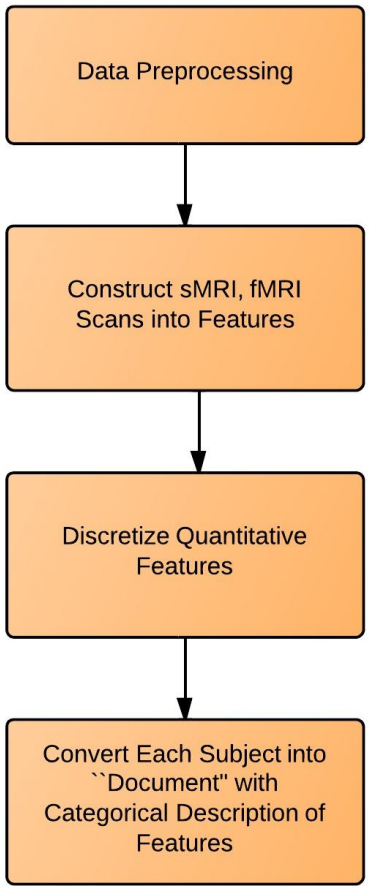
Site	N	ADHD (%)	RightHanded (%)	Male (%)	Age	SD Age
Kennedy Krieger Institute	83	0.27	0.9	0.55	10.24	1.35
NeuroIMAGE Sample	48	0.52	0.88	0.65	16.99	2.74
New York University Child Study Center	216	0.55	0.99	0.65	11.67	2.92
Oregon Health & Science University	79	0.47	1	0.54	8.84	1.12
Peking University	194	0.4	0.98	0.74	11.98	1.86
University of Pittsburgh	89	0	0.96	0.52	15.11	2.9
Washington University in St. Louis	50	0	1	0.54	11.33	3.57

Table: Summary Statistics by Site

	Typically Developing	ADHD Combined	ADHD Hyperactive	ADHD Inattentive	% Medicated Patients
Kennedy Krieger Institute	0.73	0.19	0.01	0.06	0.27
NeuroIMAGE Sample	0.48	0.38	0.12	0.02	
New York University Child Study Center	0.45	0.34	0.01	0.20	0.47
Oregon Health & Science University	0.53	0.29	0.03	0.15	0.29
Peking University	0.60	0.15		0.25	0.33
University of Pittsburgh	1.00				
Washington University in St. Louis	1.00				

Table: ADHD Statistics by Site


Topic 1
Topic 2
...
Topic M
Structural MRI
Functional MRI
Phenotypic/ Behavioral Data
Subject 1
Subject 2
...
Subject K


Data Preprocessing
Construct sMRI, fMRI Scans into Features
Discretize Quantitative Features
Convert Each Subject into ``Document'' with Categorical Description of Features

Subject-9956994
Handedness-1
Secondary.Dx-
ADHD.Measure-
Med.Status-
Site-4
DX-0
Gender-0
Age-HI
NumICS-LO
M1-LO
M2-LO
M3-HI
M4-HI
M5-HI
M6-LO
M7-LO
M8-LO
M9-HI
M10-HI
M11-HI
M12-LO
bankssts_SurfArea-HI
caudalanteriorcingulate_SurfArea-HI
caudalmiddlefrontal_SurfArea-HI
cuneus_SurfArea-LO
entorhinal_SurfArea-HI
fusiform_SurfArea-LO
inferiorparietal_SurfArea-LO
inferiortemporal_SurfArea-HI
isthmuscingulate_SurfArea-HI
lateraloccipital_SurfArea-LO
lateralorbitofrontal_SurfArea-HI

Model Parameters

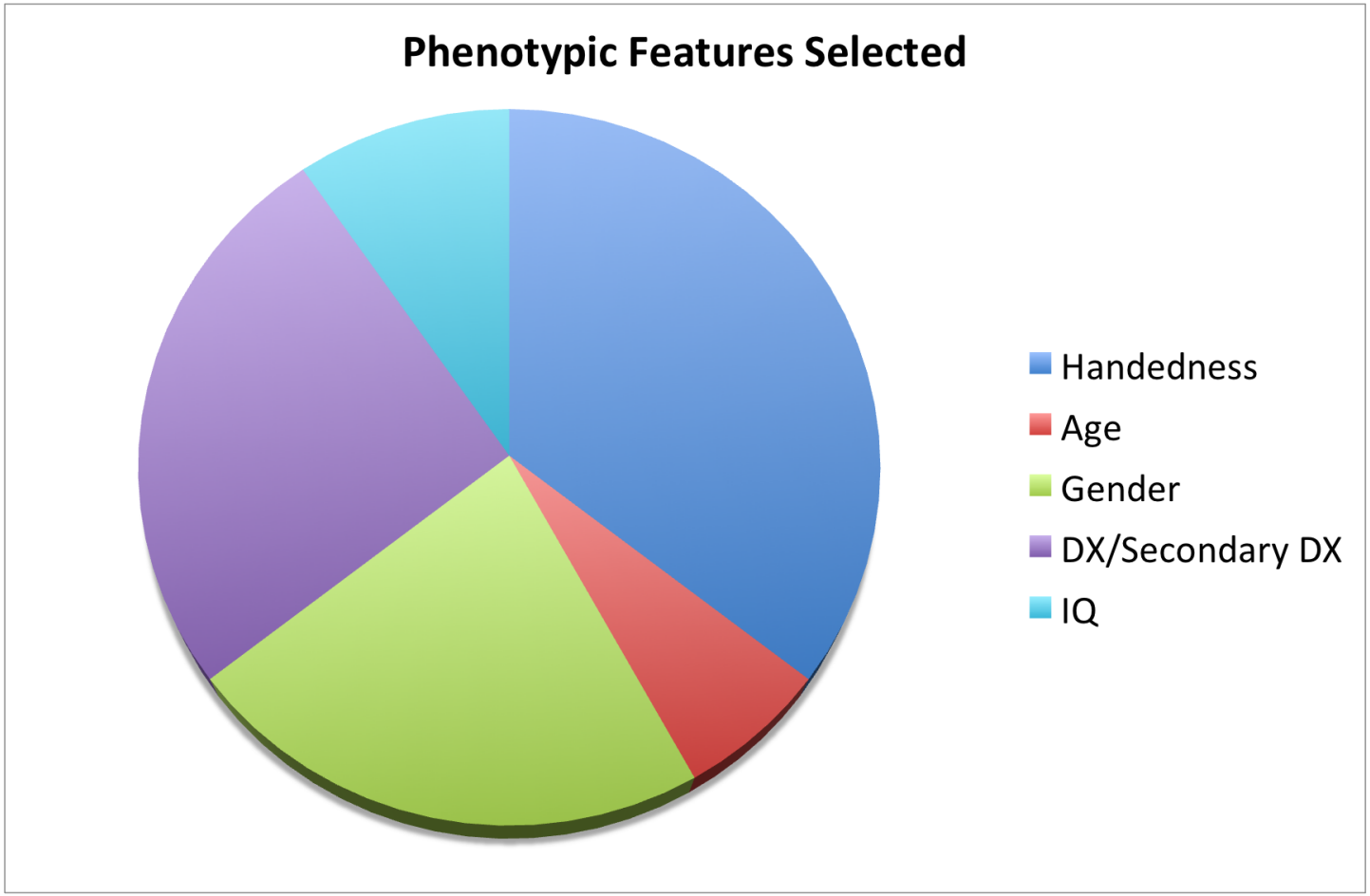
- Convert all observations into categorical by describing whether the quantitative observation is above or below the median value for that variable, i.e. isthmuscingulate_SurfArea-HI, isthmuscingulate_SurfArea-LO
- Run LDA using Variational EM (VEM) with R packages tm and ltm
- 20 Topics
- 200 "words" per topic
- Full results for topics at `http://ariana82.bol.ucla.edu/downloads-2/files/Terms.csv`

Results

Topic 1	Topic 2	Topic 3	Topic 4
divers_pos_language_07-lo	partic_neg_auditory_03-lo	posteriorcingulate_thickstd-lo	handedness-1
paracentral_thickavg-hi	superiortemporal_gauscurv-lo	precuneus_curvind-lo	partic_neg_ventral_dmn_03-hi
medialorbitofrontal_thickavg-hi	inferiortemporal_gauscurv.1-lo	frontalpole_gauscurv-lo	caudalmiddlefrontal_meancurv.1-hi
divers_pos_anterior_salience_06-hi	divers_pos_dorsal_dmn_09-lo	lingual_curvind-lo	parsopercularis_meancurv.1-lo
precentral_curvind-hi	middletemporal_surfarea-lo	precuneus_meancurv-lo	supramarginal_meancurv-hi
pericalcarine_meancurv-lo	superiorfrontal_thickavg-lo	postcentral_curvind-hi	supramarginal_gauscurv-hi
partic_neg_post_salience_03-hi	divers_pos_ventral_dmn_09-lo	lingual_foldind.1-hi	superiortemporal_meancurv-hi
strength_pos_language_04-hi	insula_surfarea.1-hi	handedness-1	strength_neg_visuospatial_05-lo
strength_neg_recn_06-lo	partic_neg_ventral_dmn_01-lo	strength_pos_precuneus_01-hi	inferiortemporal_thickavg.1-lo
supramarginal_foldind.1-lo	partic_neg_visuospatial_01-lo	dx-0	parstriangularis_meancurv.1-hi
pericalcarine_gauscurv-lo	superiortemporal_curvind.1-hi	superiortemporal_thickstd.1-lo	superiorfrontal_surfarea-hi
fusiform_gauscurv-lo	partic_pos_recn_06-lo	supramarginal_thickstd.1-hi	lateraloccipital_thickstd.1-lo
divers_neg_precuneus_02-hi	insula_foldind-hi	strength_pos_visuospatial_10-hi	strength_neg_sensorimotor_06-lo
postcentral_thickavg-lo	divers_neg_visuospatial_05-hi	precuneus_surfarea.1-hi	caudalmiddlefrontal_curvind-hi
divers_neg_recn_01-hi	parsopercularis_foldind.1-lo	caudalanteriorcingulate_foldind.1-lo	postcentral_foldind-lo
divers_pos_precuneus_02-hi	postcentral_meancurv-lo	superiortemporal_meancurv-lo	strength_neg_precuneus_01-lo
inferiorparietal_gauscurv-lo	pericalcarine_curvind-lo	parsopercularis_meancurv-hi	divers_neg_visuospatial_05-hi
partic_neg_sensorimotor_05-lo	partic_neg_basal_ganglia_01-lo	divers_pos_post_salience_05-lo	partic_pos_post_salience_03-hi
handedness-1	strength_pos_ventral_dmn_01	middletemporal_surfarea-lo	middletemporal_curvind-hi
superiorfrontal_meancurv.1-lo	parstriangularis_gauscurv-hi	strength_neg_post_salience_02-lo	partic_pos_basal_ganglia_05-lo
divers_neg_sensorimotor_03-lo	partic_pos_dorsal_dmn_05-hi	cuneus_foldind.1-lo	frontalpole_meancurv.1-lo
frontalpole_meancurv-lo	lateraloccipital_gauscurv-lo	frontalpole_grayvol.1-lo	superiortemporal_curvind.1-hi
lingual_curvind.1-lo	superiorfrontal_gauscurv.1-lo	divers_neg_dorsal_dmn_01-lo	partic_neg_basal_ganglia_03-hi
strength_pos_ventral_dmn_05-hi	caudalmiddlefrontal_thickstd.1	rostralanteriorcingulate_grayvol.1-lo	divers_pos_precuneus_02-hi
isthmuscingulate_thickstd.1-hi	parsopercularis_curvind.1-lo	frontalpole_grayvol-lo	bankssts_curvind-hi
inferiortemporal_gauscurv-lo	inferiortemporal_foldind.1-lo	partic_pos_recn_03-hi	inferiortemporal_gauscurv.1-lo
partic_pos_post_salience_12-lo	divers_neg_precuneus_02-lo	caudalmiddlefrontal_thickstd.1-lo	lateralorbitofrontal_surfarea.1-lo
divers_pos_visuospatial_09-lo	bankssts_curvind.1-hi	paracentral_thickstd.1-lo	strength_pos_basal_ganglia_02-hi
partic_pos_basal_ganglia_05-lo	partic_neg_anterior_salience_	cuneus_foldind-hi	supramarginal_curvind.1-hi
transversetemporal_curvind-lo	parstriangularis_surfarea.1-lo	divers_neg_post_salience_09-lo	partic_pos_visuospatial_05-hi
strength_neg_visuospatial_11-hi	caudalmiddlefrontal_meancurv	caudalmiddlefrontal_gauscurv.1-hi	strength_pos_anterior_salience_05-lo
caudalmiddlefrontal_meancurv.1-lo	strength_pos_ventral_dmn_05	partic_neg_post_salience_05-hi	caudalanteriorcingulate_gauscurv.1-hi
strength_pos_ventral_dmn_02-hi	divers_neg_lecn_05-hi	gender-1	partic_neg_anterior_salience_05-lo
divers_pos_language_07-hi	partic_neg_ventral_dmn_09-lo	parsorbitalis_grayvol-lo	precuneus_foldind.1-hi
partic_pos_anterior_salience_01-hi	partic_pos_dorsal_dmn_03-lo	partic_pos_lecn_04-hi	transversetemporal_grayvol.1-hi
medialorbitofrontal_gauscurv.1-lo	m9-lo	lingual_curvind-hi	partic_neg_recn_04-lo
strength_pos_visuospatial_01-hi	divers_pos_dorsal_dmn_06-hi	strength_pos_post_salience_11-hi	supramarginal_thickstd.1-hi
parahippocampal_foldind-lo	parsopercularis_grayvol-lo	divers_pos_basal_ganglia_04-hi	inferiorparietal_meancurv-hi
divers_pos_sensorimotor_04-lo	transversetemporal_foldind-lo	partic_neg_post_salience_12-hi	parsorbitalis_foldind-hi
bankssts_thickavg.1-hi	precentral_curvind.1-lo	partic_neg_dorsal_dmn_03-lo	paracentral_meancurv-hi
medialorbitofrontal_curvind-hi	superiorparietal_curvind.1-lo	partic_pos_visuospatial_03-hi	temporalpole_grayvol-hi
divers_neg_visuospatial_02-lo	m1-lo	partic_neg_ventral_dmn_03-hi	m1-lo
divers_pos_auditory_01-lo	partic_neg_dorsal_dmn_02-lo	partic_pos_recn_01-lo	partic_pos_high_visual_02-hi
strength_pos_recn_02-lo	pericalcarine_curvind.1-lo	partic_neg_recn_04-lo	temporalpole_gauscurv.1-hi
partic_pos_lecn_04-lo	strength_neg_visuospatial_10-	temporalpole_foldind-lo	middletemporal_meancurv-hi
inferiortemporal_surfarea-hi	superiorparietal_surfarea.1-lo	partic_pos_post_salience_10-hi	strength_neg_ventral_dmn_05-hi
dx-0	lateraloccipital_surfarea-lo	divers_neg_sensorimotor_01-hi	divers_neg_precuneus_03-lo

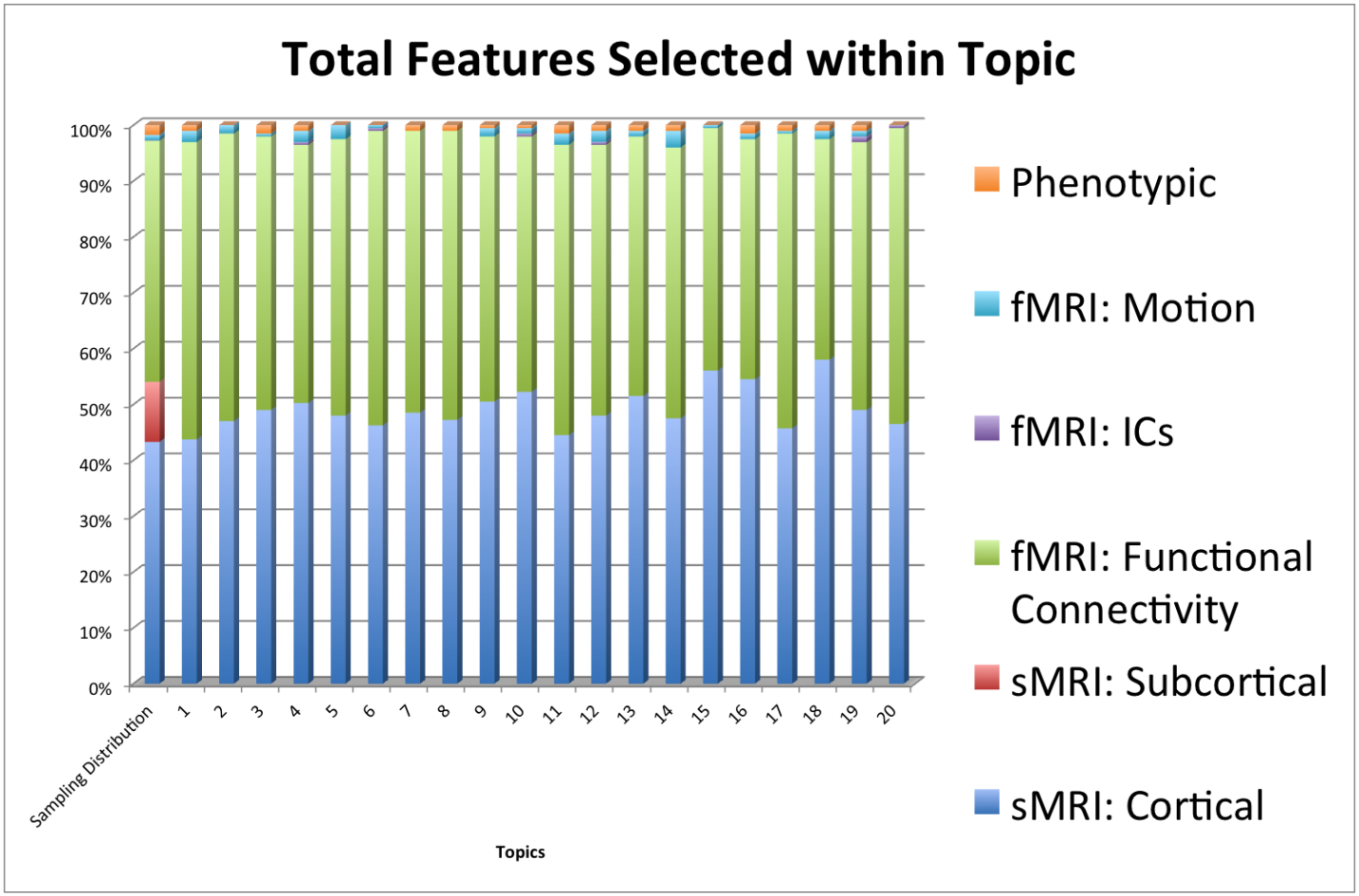
Topic 3 Interpretation:

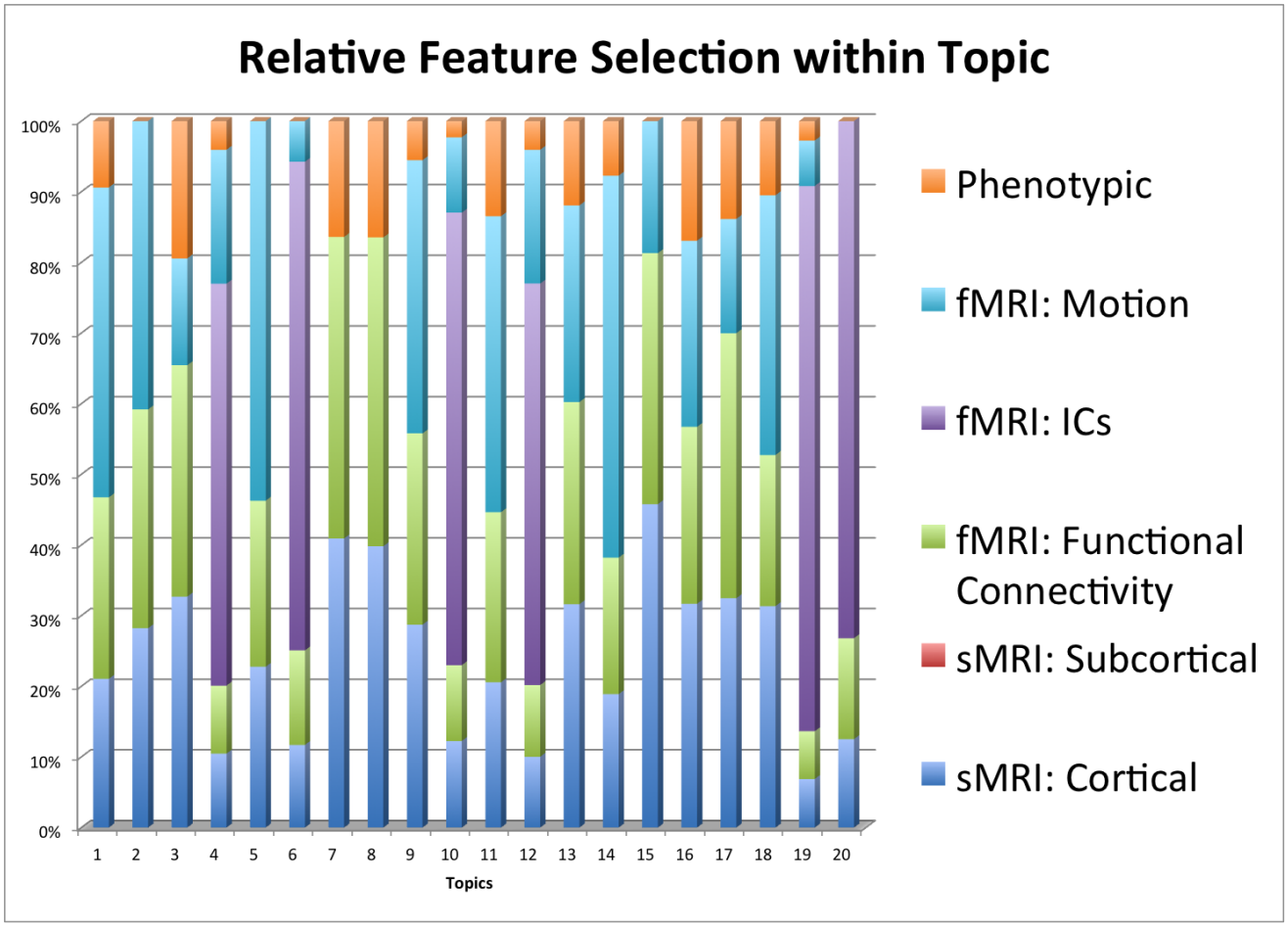
- (TYPICALLY DEVELOPING) looks like a cluster of areas that may be either in the default mode network (e.g. the precuneus), or areas that are involved in switching between salience networks and default mode network.
- Damien Fair (2011) published a paper hypothesizing that inattentive ADHD individuals may have problems suppressing Default mode so that they can focus and attend continuously to the task at hand.
- Given that these cluster with the Dx=0 (meaning that they are typically developing children) and male gender, the DMN areas must be associated with healthy levels. (which as a general rule look low for the DMN areas).
- Full Topics at `http://ariana82.bol.ucla.edu/downloads-2/files/Terms.csv`

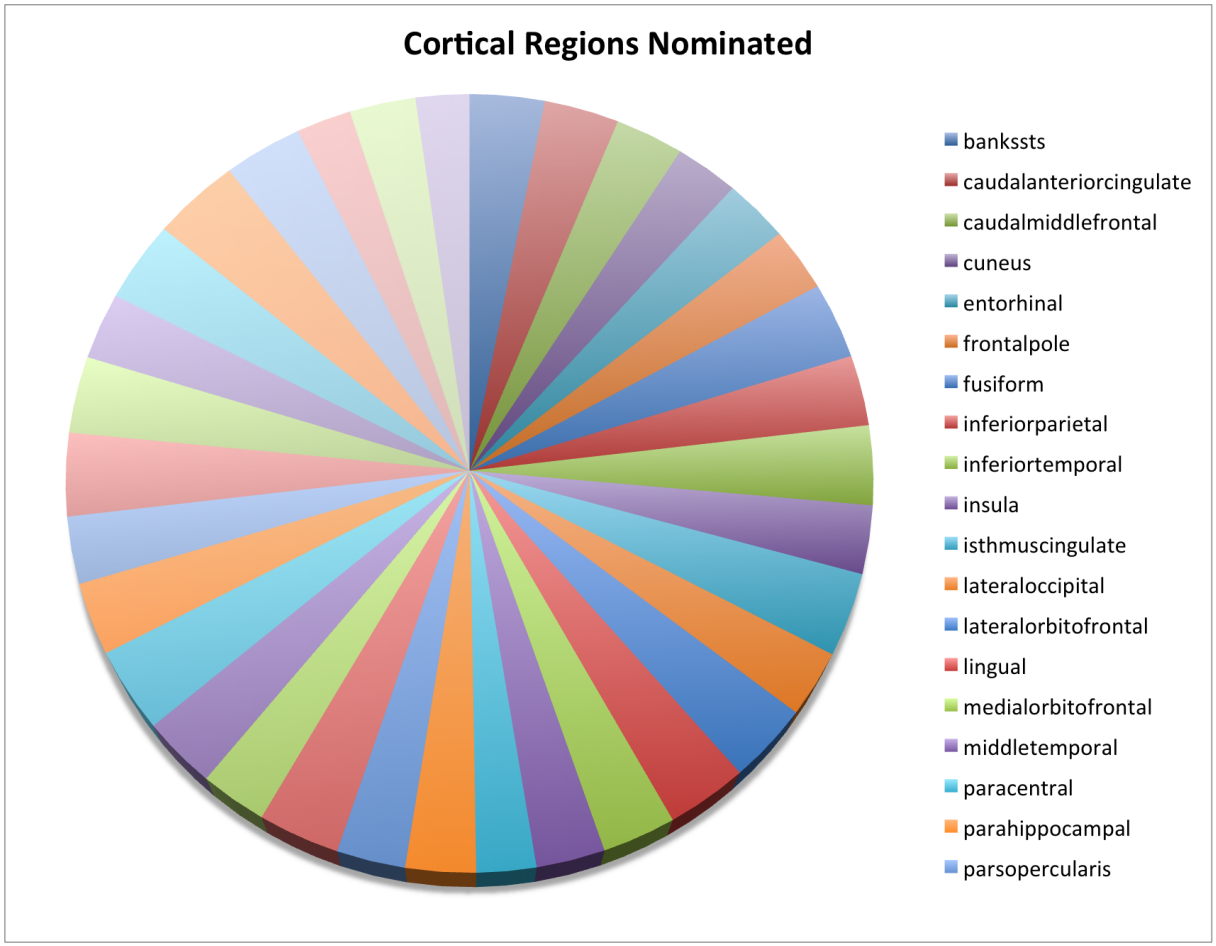

Phenotypic Features Selected
Handedness
Age
Gender
DX/Secondary DX
IQ

Phenotypic Variables Selected

- Age: 66% Age-LOW variables ($<$ 11.56 years old)
- Gender: 100% Male
- Handedness: 100% Right-Handed
- IQ: 100% Low
- DX: 100% Healthy controls

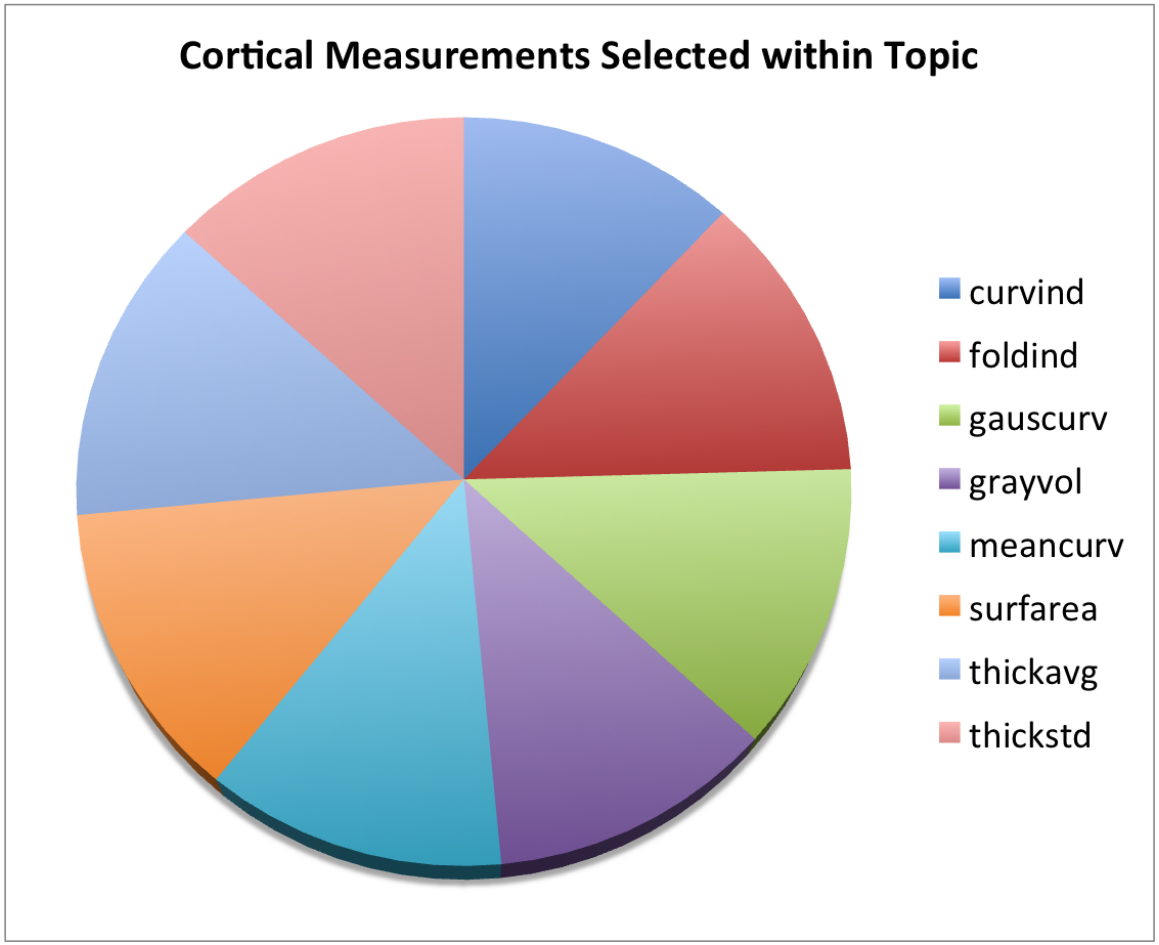

Total Features Selected within Topic
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Sampling Distribution
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Topics
Phenotypic
fMRI: Motion
fMRI: ICs
fMRI: Functional Connectivity
sMRI: Subcortical
sMRI: Cortical


Relative Feature Selection within Topic
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Topics
Phenotypic
fMRI: Motion
fMRI: ICs
fMRI: Functional Connectivity
sMRI: Subcortical
sMRI: Cortical


Cortical Regions Nominated
bankssts
caudalanteriorcingulate
caudalmiddlefrontal
cuneus
entorhinal
frontalpole
fusiform
inferiorparietal
inferiortemporal
insula
isthmuscingulate
lateraloccipital
lateralorbitofrontal
lingual
medialorbitofrontal
middletemporal
paracentral
parahippocampal
parsopercularis

Cortical Regions Interpretation:

- Regions involved in memory (hippocampus, entorhinal cortex), which are not necessarily noted in ADHD (at least as of yet)
- Some executive function regions from frontal cortex (these are known to be effected in ADHD)
- Full Topics at `http://ariana82.bol.ucla.edu/downloads-2/files/Terms.csv`


Cortical Measurements Selected within Topic
curvind
foldind
gauscurv
grayvol
meancurv
surfarea
thickavg
thickstd

(Unofficial) 1st Place Findings

University of Alberta, including Gagan Sidhu, Matthew Brown, Russell Greiner, Nasimeh Asgarian, and Meysam Bastani.

- Functional Features: Functional (Time-averaged voxels, PCA-projected time courses, low-frequency voxel Fourier components, voxel weightings on ICA connectivity maps)
- Phenotypic Features: Site, Gender, Age, Handedness, IQ
- Noted that functional features performed worse than phenotypic features, so selected a logistic classifier using *only* phenotypic information.
- Spectral Features performed best in the test set overall. Optimal at identifying TD, but suboptimal for ADHD.
- `http://papersdb.cs.ualberta.ca/~papersdb/uploaded_files/1070/paper_30171_Brown_ProvisionalPDF.pdf`

(Official) 1st Place Findings

Johns Hopkins University, including Brian Caffo, Ciprian Crainiceanu, Ani Eloyan, Fang Han, Han Liu, John Muschelli, Mary Beth Nebel, and Tuo Zhao.

- Voting scheme across 4 different algorithms,
- The most promising imaging biomarker was a correlation graph from a motor network parcellation.
- `http://www.frontiersin.org/Systems_Neuroscience/10.3389/fnsys.2012.00061/full`

Subteam	Covariates	Processing	Methods
1	All IQ, age, gender, handedness, site	NITRC	Motor network parcellation, random forest random forest for prediction.
2	All IQ, age, gender, handedness, site	NITRC	Feature extraction, clustering, LDA, multi-class SVM.
3	Composite IQ, age, gender, handedness, site	NITRC	CUR decomposition feature extraction, gradient boosting.
4	Composite IQ, age, gender, handedness, site	NITRC NB Athena	264 seed voxels, motion parameters, PCA, machine learning algorithms.

Composite IQ uses the average of all available IQs. All IQ suggests the use of all available IQ measurements. NITRC for image processing implies the use of the 1000 Functional Connectome processing scripts. NB refers to the NeuroBureau pipelines.

First, image features were extracted using online clustering and latent Dirichlet allocation (Blei et al., 2003) based topic models. Here each sample was considered to be one document (collection of words) and the label of each measurement as a word in the vocabulary.

3rd Place Findings

UCLA/Yale, including John B. Colby, Jeffrey D. Rudie, Jesse A. Brown, Pamela K. Douglas, Mark S. Cohen, and Zarrar Shehzad

- Features: Sub(Cortical) Morphology, rs-fcMRI and REHo, Power Spectra, Graph Theory, Global Connectivity
- Trained Classifiers within each site, within each Feature Set. Optimal features were not consistent across sites.
- Spectral Features performed best in the test set overall. Optimal at identifying TD, but suboptimal at identifying ADHD.
- `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3419970/`

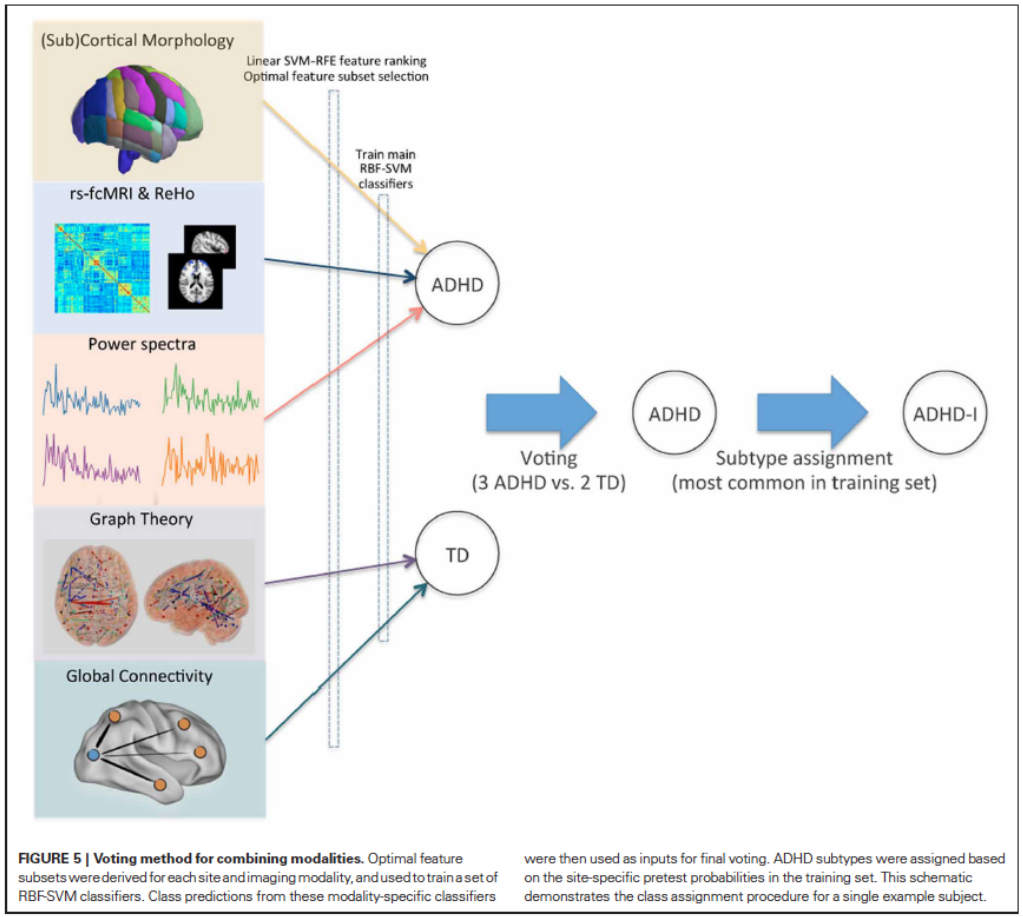


FIGURE 5 | Voting method for combining modalities. Optimal feature subsets were derived for each site and imaging modality, and used to train a set of RBF-SVM classifiers. Class predictions from these modality-specific classifiers were then used as inputs for final voting. ADHD subtypes were assigned based on the site-specific pretest probabilities in the training set. This schematic demonstrates the class assignment procedure for a single example subject.

Topic Modeling Conclusion:

- Phenotypic features were strong performers within the topics, consistent with the Alberta team's findings.
- Diagnostic Possiblities: There was no "ADHD" topic identified.
- There are two possible explanations for this: "ADHD" is not distinct enough neurologically from normal controls to identify such or domain, or the diagnostic practices were wildly different across sites which made pooled information vague.
- Example: Females with low-IQ in Peking were almost certainly identified as ADHD.
- Future work will address this using author-topic models `http://www.datalab.uci.edu/author-topic/398.pdf`, and by labeling features as high/low with respect to Site median rather than global median.

Future Work:

- Model using author/topic models, where author=site.
- Redo modeling just within ADHD Patients.
- Assess topic modeling as a feature selection method. Do items belonging with a clinical feature imply predictive power?