Identifying SWEDD in PPMI

Christian Bracher

October 7, 2014

Prologue: PPMI and Machine Learning

Causes of Parkinson's Disease (PD) are poorly understood.

PPMI's strategy is a “big data” approach to these challenges:

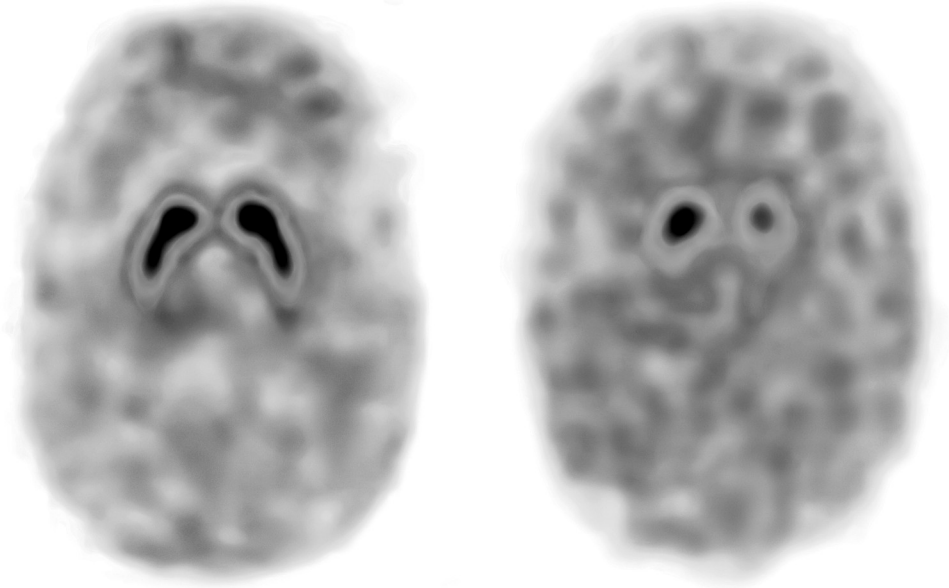
- $80 M five year longitudinal, clinical and observational study
- Extensive horizontal data set:
  Biomarker, psychological, neurological, imaging, genetics data
- Several cohorts with altogether ~ 1,000 subjects

Use machine learning, advanced statistics to examine data:

- Establish classifying algorithms, search for patterns in data
- Be wary: Noisy data, small cohort sizes render validation difficult

SPECT Brain Imaging (DaTSCAN)

- Increasingly common "standard" in PD diagnosis
- Uses short-lived isotope ^{123}I in a marker that binds to dopamine transporters
- FDA approved 2011, used much longer in Europe



HEALTHY PARKINSON'S

SPECT Brain Imaging (DaTSCAN)

- DaTSCAN statistics:
  - Healthy subjects pass DaTSCAN, but
  - 15 % of PD subjects are “misclassified”
- Define a new subject cohort by exclusion:

  SWEDD — subjects without evidence of dopaminergic dysfunction

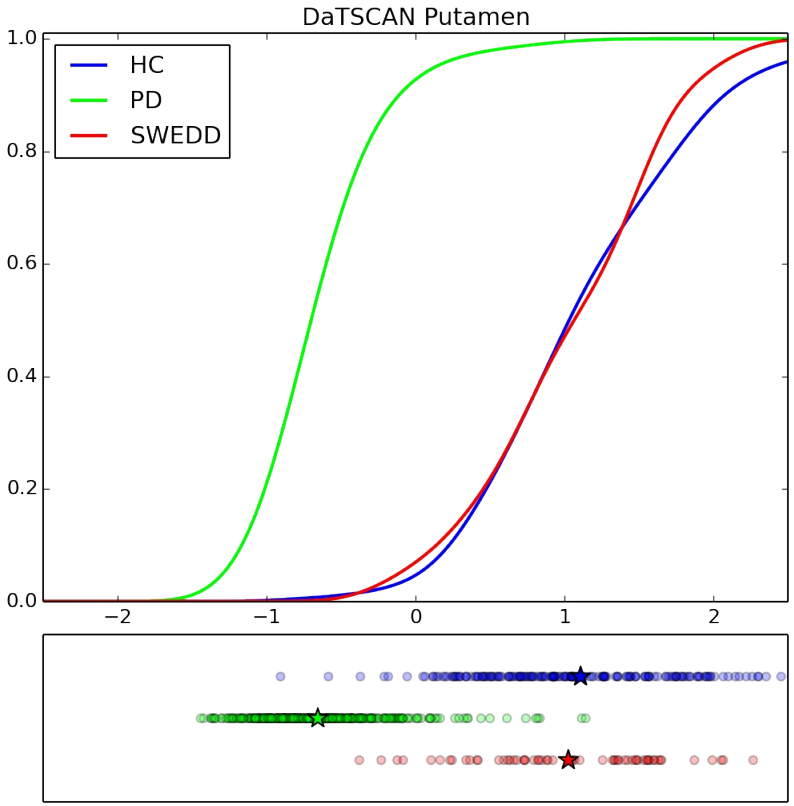
See e.g.: K. Marek *et al.*, Imaging the dopamine system to assess disease-modifying drugs: studies comparing dopamine agonists and levodopa. Neurology **61** (2003), S43 – 48.

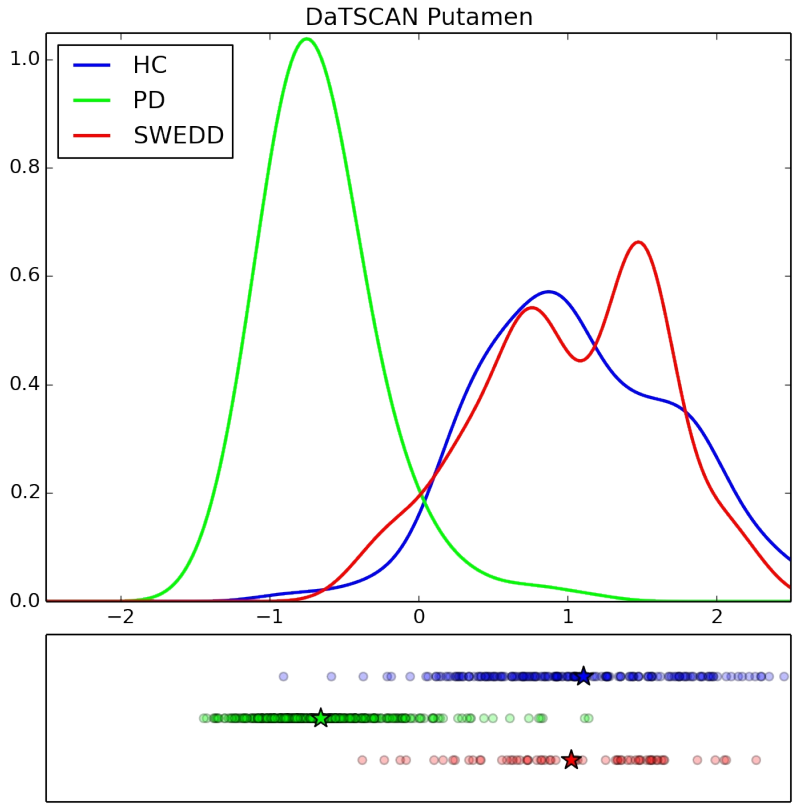
SWEDD

- Question: Does SWEDD represent a distinct syndrome? Are there independent biological markers for SWEDD?

DaTSCAN Putamen Signal

Clear separation between PD and SWEDD distributions (with some outliers):





Cumulative probability (left) and probability density (right) distributions with Gauss filter.
Bottom: Raw data (normalized) and cohort averages

Evaluating Markers of Parkinson's Disease

Many other physiological changes have been linked to PD:

- Lipid metabolism
- Alteration in protein levels in cerebrospinal fluid (CSF)
- Genetic variants, gene expression
- Neurological abnormalities:
  Olfactory, REM sleep, spatiovisual capability, motor skills, …
- Psychological dysfunction (depression, …)

Are they useful in distinguishing SWEDD from “generic” PD?

Part I:

Identifying SWEDD Using Clinical Assessments

Identifying SWEDD Using Clinical Assessments

- Studies indicate that patients with Parkinsonian syndromes deviate from healthy controls in many clinical parameters
- Using test of significance, we examined motor and non-motor neurological assessments for differences between PD and SWEDD:
  - Olfactory (smelling) ability (UPSIT)
  - REM sleep behavior (acting out dreams)
  - Autonomic dysfunction (SCOPA-AUT)
  - Self-evaluation of motor difficulties (UPDRS Part II)

Note: RBD (REM sleep), autonomic, and olfactory dysfunction are early risk factors for PD:

- M. Visser et al., Assessment of Autonomic Dysfunction in Parkinson's Disease: The SCOPA-AUT, Movement Disorders **19** (2004), 1306–1312.
- K. Stiasny-Kolster *et al*., Combination of 'idiopathic' REM sleep behaviour disorder and olfactory dysfunction (...), Brain **128** (2005), 126–137.
- A. Siderowf, A. E. Lang, Pre-Motor Parkinson's Disease: Concepts and Definitions, Mov Disord. **27** (2012), 608–616.

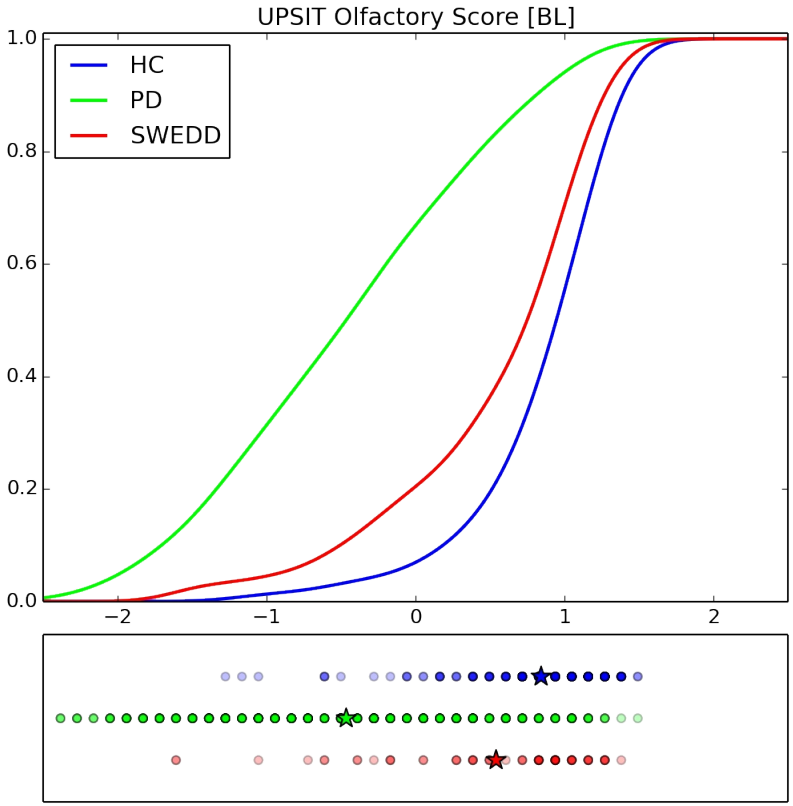
Olfactory – Fine Motor – Autonomic Function

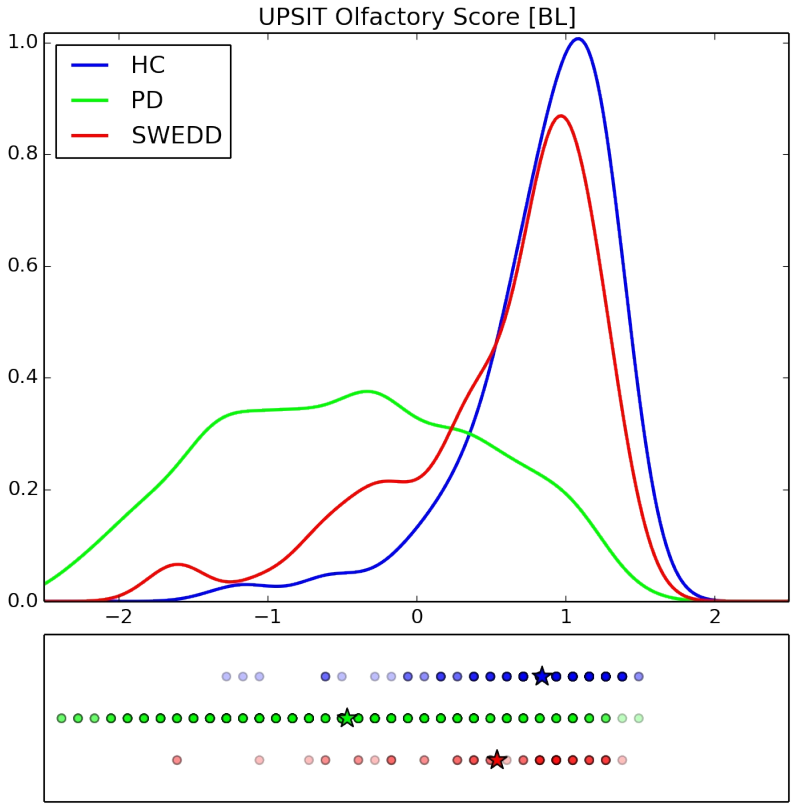
Elimination/Aggregation of features yields simple classifier:

- Feature #1: Olfactory Score (UPSIT)
  - Number of correctly identified scents
  - Feature Importance *(Random Forest Classifier):* 42.4%
- Feature #2: Adaptation Subscore (SCOPA-AUT)
  - *Lightheadedness, Dizzy Standing Up, Fainting, Daytime Perspiration, Night Perspiration, Bright Light Sensitivity, Cold Tolerance, Heat Tolerance* scores
  - Feature Importance *(Random Forest Classifier):* 34.2%
- Feature #3: Fine Motor Skills Subscore (MDS-UPDRS Part II)
  - Sum of *Dressing, Hygiene, Handwriting, Hobbies, Tremor* scores
  - Feature Importance *(Random Forest Classifier):* 23.5%

Distribution of Olfactory Ability

Strong reduction of smelling sense in PD cohort, SWEDD less impacted on average:

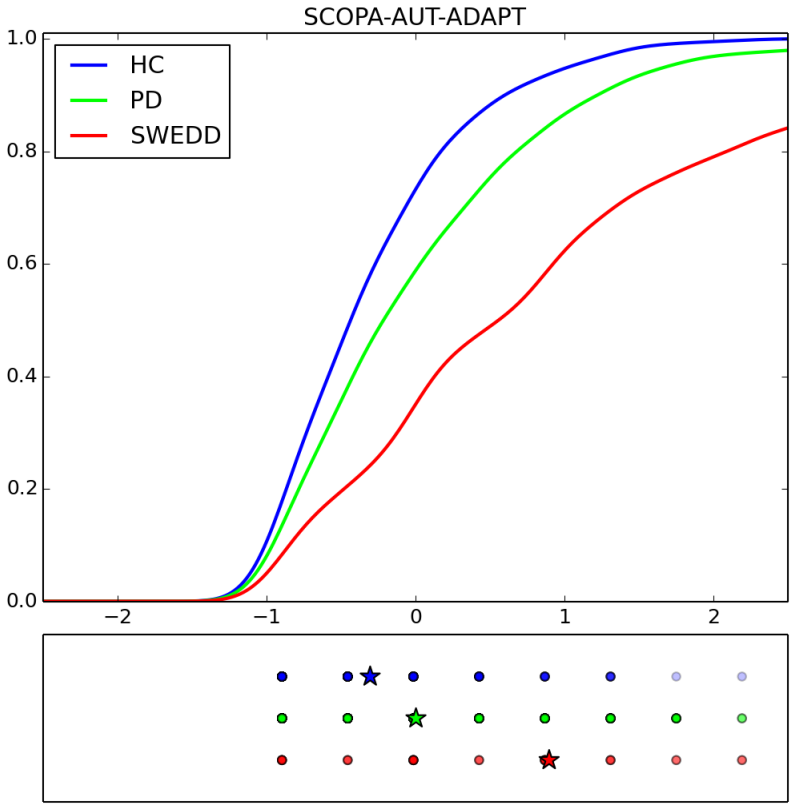


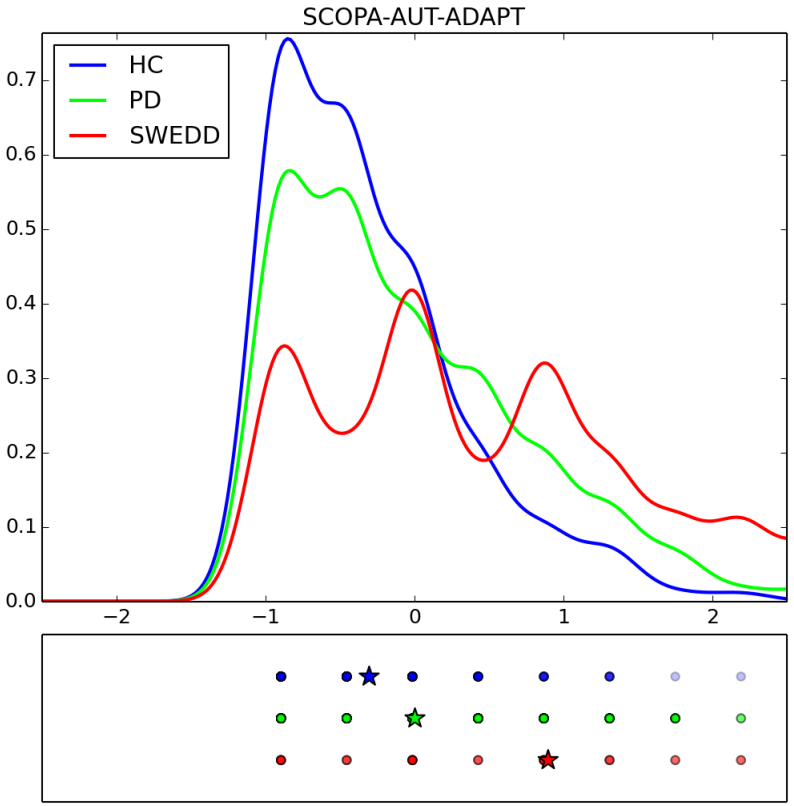


Cumulative probability (left) and probability density (right) distributions with Gauss filter.
Bottom: Raw data and cohort averages

Profiles of SCOPA-AUT Adaptation Score

Adaptation loss is much more pronounced in SWEDD than PD and healthy controls:

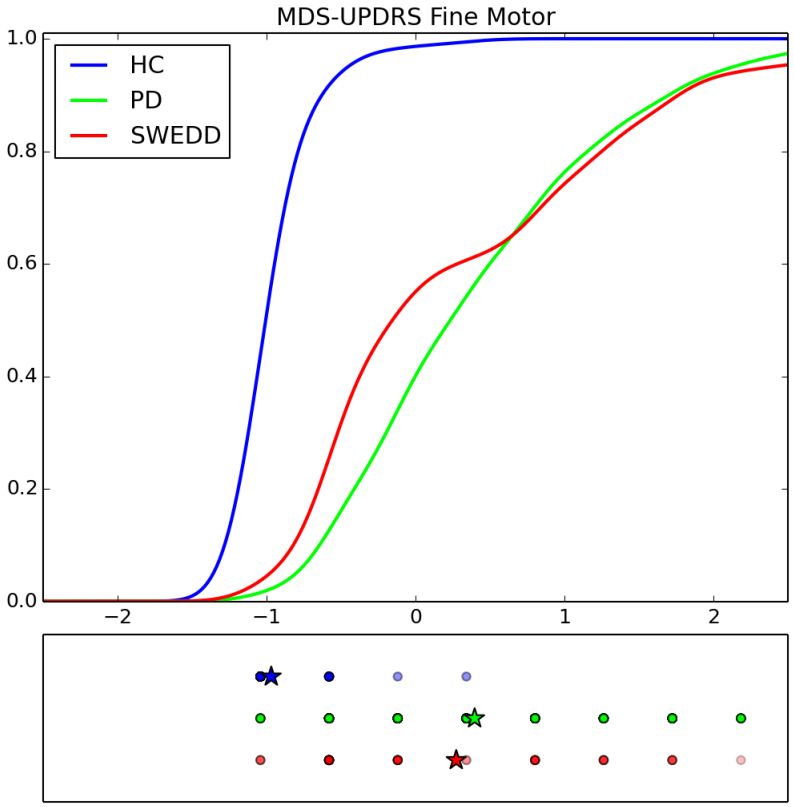


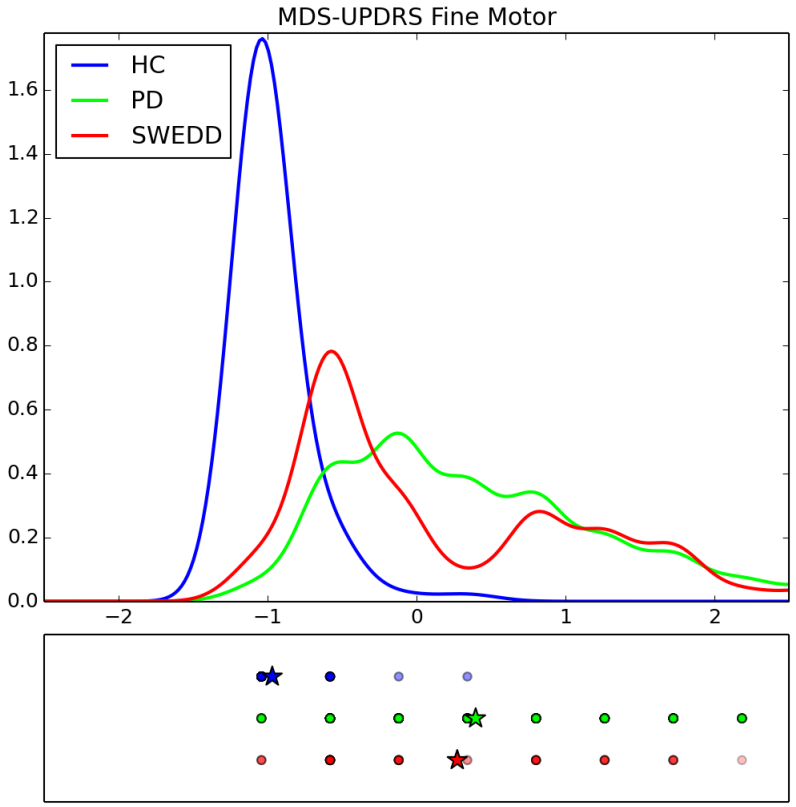


Cumulative probability (left) and probability density (right) distributions with Gauss filter.
Bottom: Raw data and cohort averages

Distribution of Fine Motor Skills

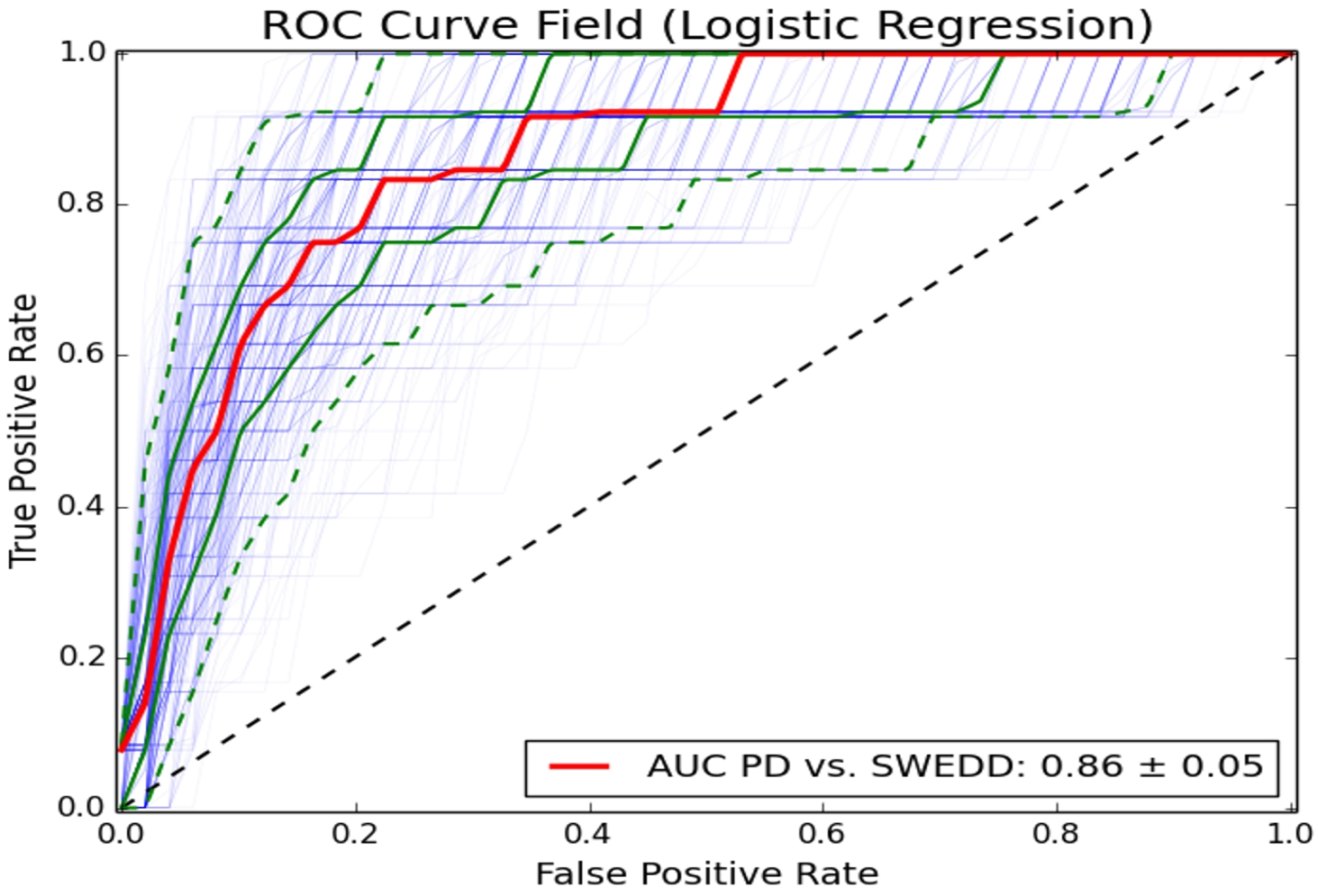
Loss of fine motor skills in PD and SWEDD – bimodal distribution for SWEDD?





Cumulative probability (left) and probability density (right) distributions with Gauss filter.
Bottom: Raw data and cohort averages

Olfactory – Fine Motor – Autonomic Classifier



Cross-validation of regression classifier shows performance range

Topological Data Analysis with AYASDI

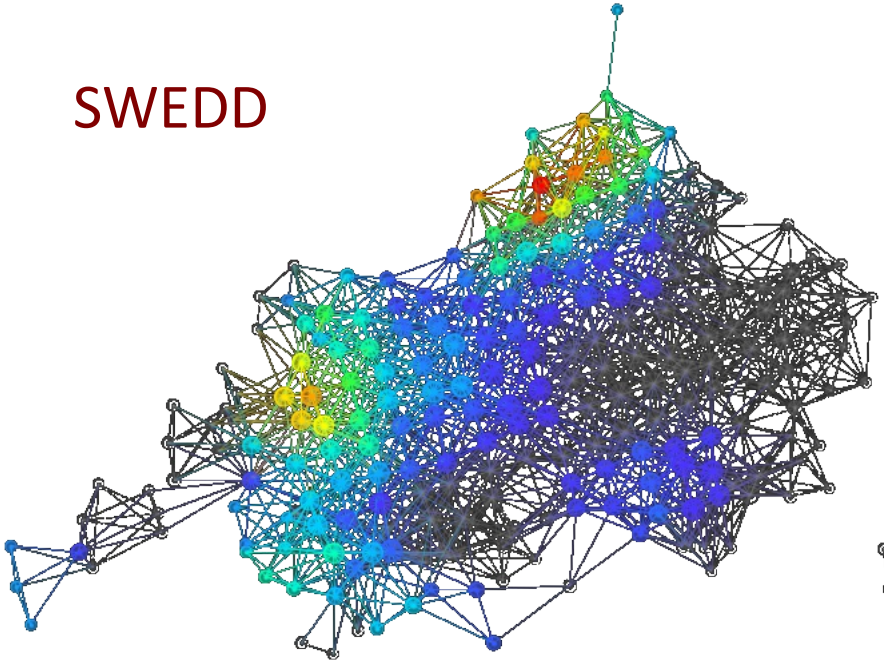
AYASDI is a proprietary software package that performs unsupervised learning and cluster analysis, using Topological Data Analyses (TDA) to discover new insights from data:

- AYASDI is based on the geometry of the data in feature space
  - Metrics (built-in or custom) define a distance between data points
  - Lenses perform a projection of the data
- AYASDI translates results into a network or graph of the data

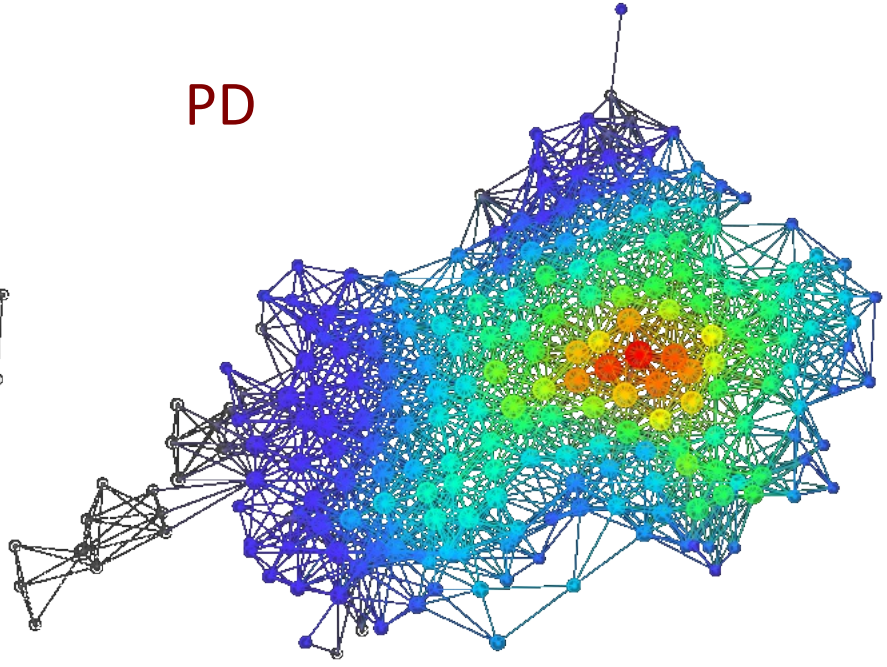
  See e.g.: G. Carlsson, Topology and Data, Bulletin of the American Mathematical Society **46** (2009), 255–308.
- AYASDI displays two-dimensional views of the data network to identify clusters of data points with shared features
- It is easy to perform statistical analysis of parts of the graph to reveal common features of clustered data

AYASDI: Olfactory – Fine Motor – Adaptation

SWEDD



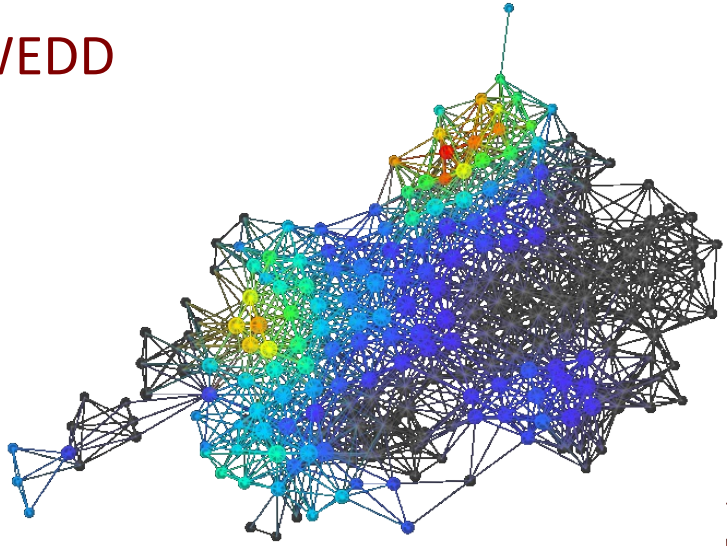
PD



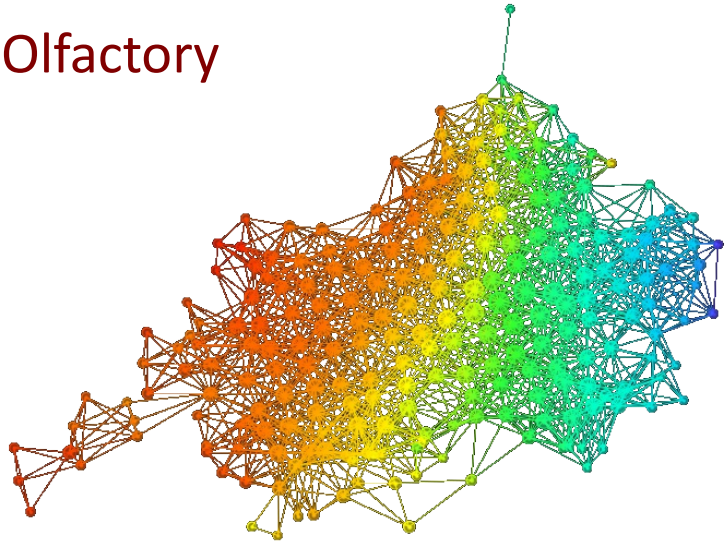
Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

AYASDI Analysis – Parameter Space

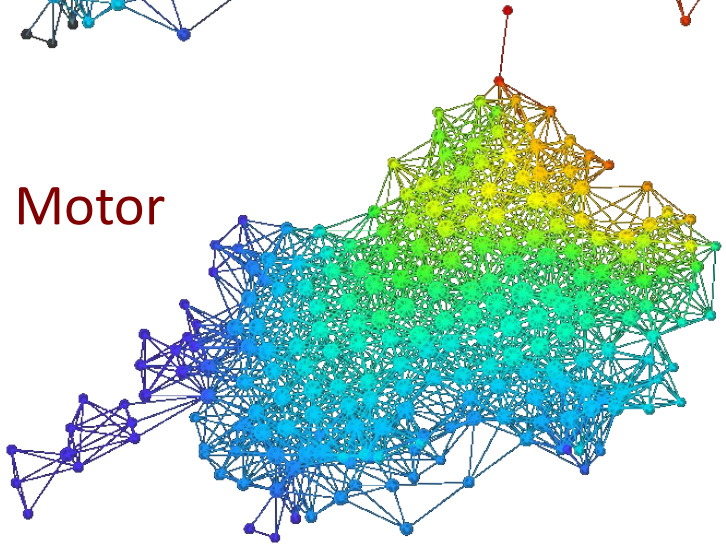
SWEDD



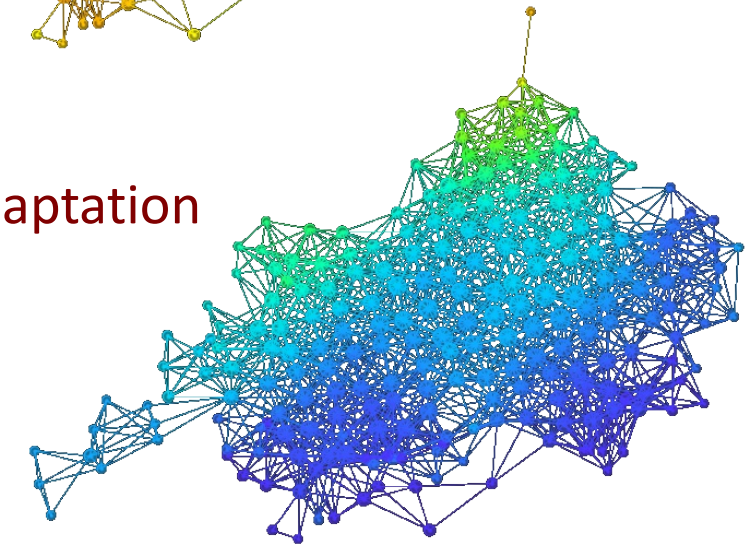
Olfactory



Fine Motor



Adaptation



Implications for Diagnosis & Healthcare

DaTSCAN is gaining recognition as a "standard" in PD diagnosis. But it has problematic aspects:

- Can be performed only at specialized centers (SPECT imaging)
- Requires access to a short-lived radioactive marker (^{123}I)
- Procedure is expensive ($3,000 – $5,000)

While not a full alternative to DaTSCAN, the olfactory–fine motor–adaptation classifier addresses these issues:

- Can be performed anywhere, even in the field or at home
- Requires only a "sniff test" and two screening questionnaires
- Materials are very affordable ($30 – $50)

Implications for Diagnosis & Healthcare

Idea: General practitioner performs "triage" on patients with *de novo* motor symptoms:

- Obtain UPSIT olfactory score (0 – 40)
- Patient fills fine motor screening questionnaire (score 0 – 20)
- Same for autonomous/adaptation problems (score 0 – 32)

Use PPMI results to estimate likelihood of PD:

- Enter scores into logistic regression equation (applet?)
- Compare result to probability threshold (e.g., 98%)
- Reserve DaTSCAN for patients with PD likelihood below threshold (normosmic patients, patients with strong autonomous dysfunction)

Part II:

Identifying SWEDD Using Biomarkers in PPMI

Performance of Biomarkers in PPMI

PPMI captures many biomarkers that have been deemed important in PD. How do they perform?

- Markers that appear unrelated to disease status:
  - Example: Lipid metabolism

    See e.g.: L. de Lau *et al.*, Serum Cholesterol Levels and the Risk of Parkinson's Disease, Am. J. Epidemiol. **164** (2006), 998–1002.
- Markers showing statistically significant differences between cohorts on average, but variation within groups is sufficiently large to lack discriminatory power:
  - Example: PD-related proteins in cerebrospinal fluid (CSF)

    See e.g.: M. Shi *et al.*, Cerebrospinal Fluid Biomarkers for Parkinson Disease Diagnosis and Progression, Ann Neurol. **69** (2011), 570–580.
  - Example: RNA transcription rates of PD-related proteins

    See e.g.: L. Molochnikov *et al.*, A molecular signature in blood identifies early Parkinson's disease, Mol. Neurodegener. **7** (2012), 26.

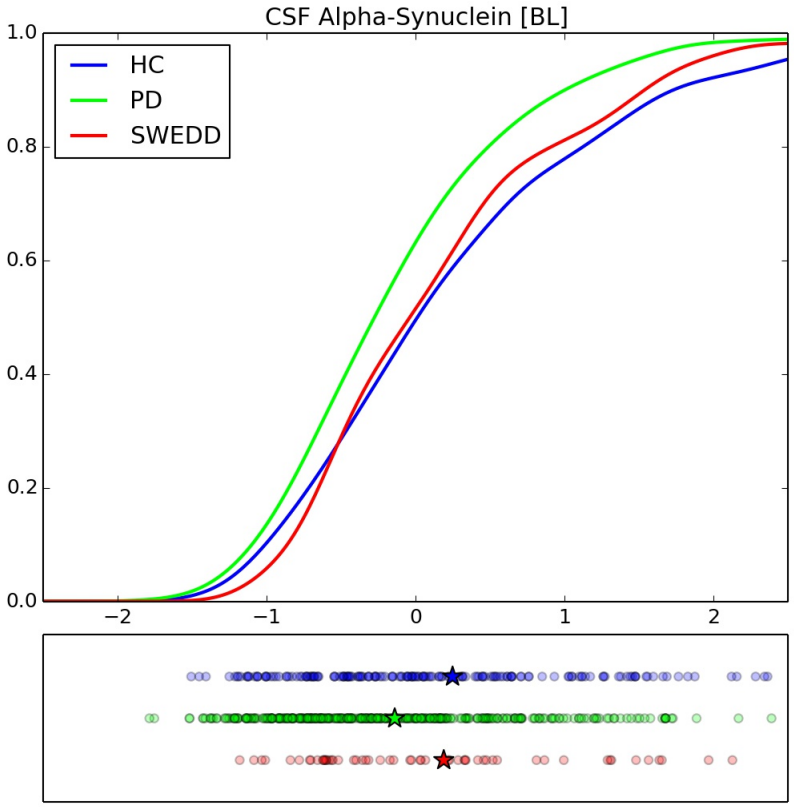
Identifying SWEDD via CSF Protein Levels?

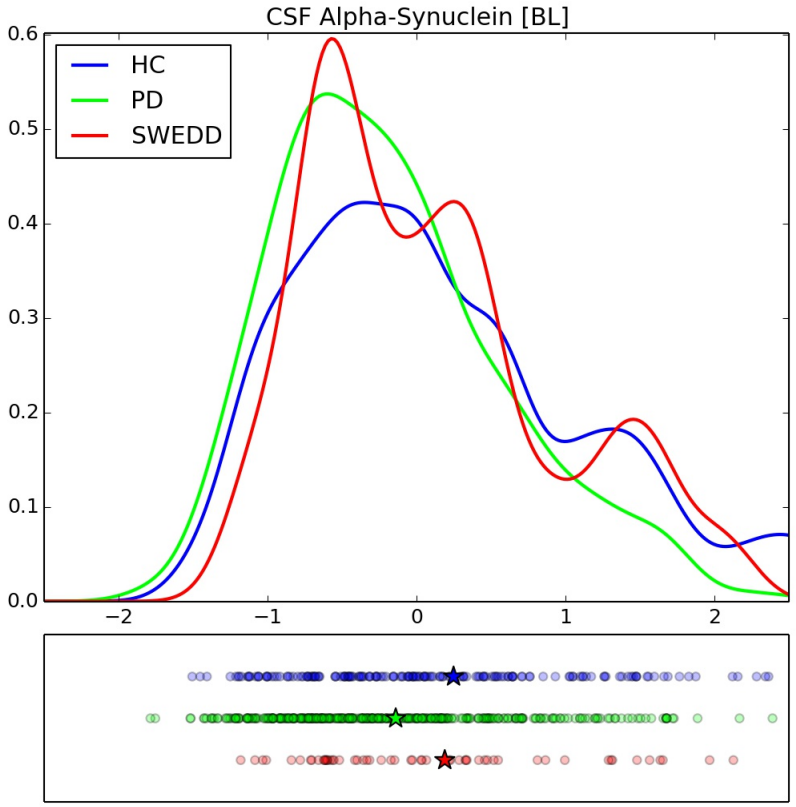
Here, restriction to proteins in the cerebrospinal fluid (CSF):

- Protein levels recorded (at baseline):
  - α-synuclein (the main component of the Lewy bodies seen in PD)
  - Amyloid-β_{42} (a protein fragment implicated in Alzheimer's Disease)
  - Phosphorylated *tau*-protein (another neuropathic protein in AD)
  - Total *tau*-protein
- Statistically significant differences between cohorts on average:
  - CSF α-synuclein is suppressed in PD, but unaltered in SWEDD
  - Amyloid-β_{42} is increased in SWEDD, but unaltered in PD
- But: Large variation within cohorts destroys discriminatory power

CSF Profiles: α-Synuclein

Significant decrease of α-syn in PD cohort (compared to HC, SWEDD) on average:

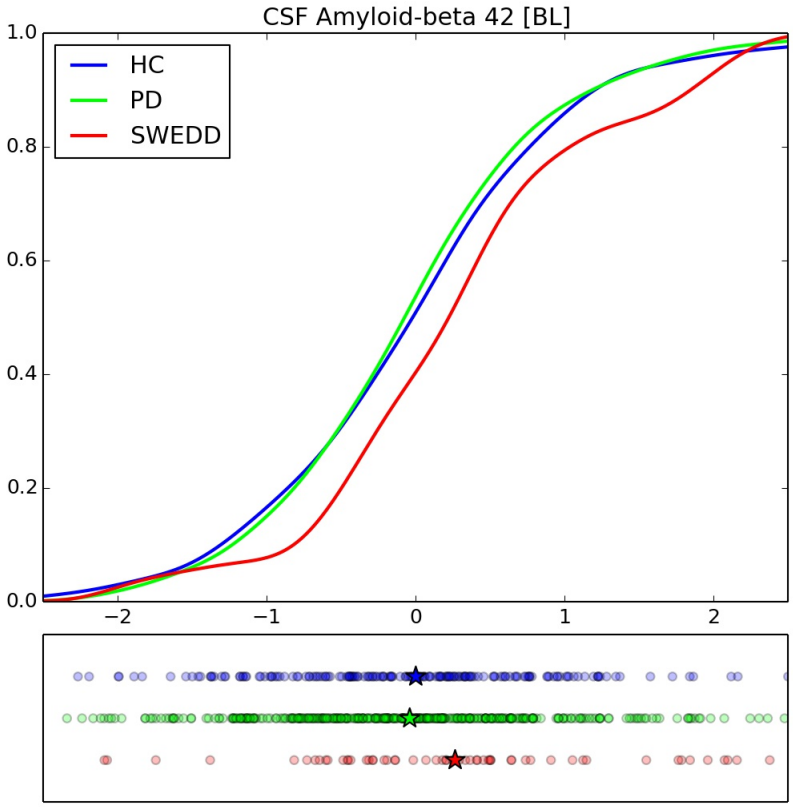


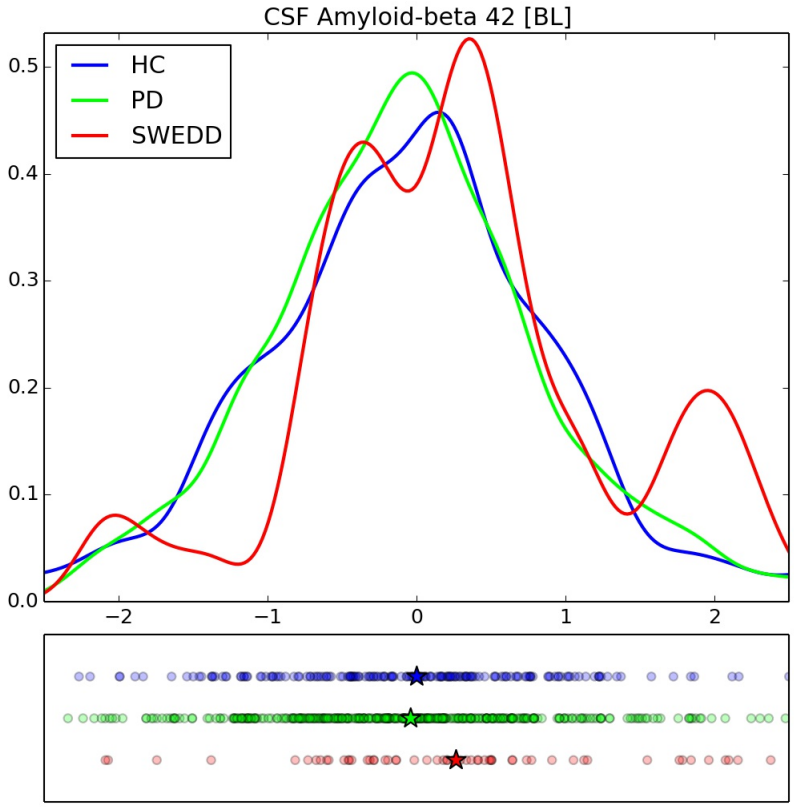


Cumulative probability (left) and probability density (right) distributions with Gauss filter.
Bottom: Raw data and cohort averages

CSF Profiles: Amyloid-β_{42}

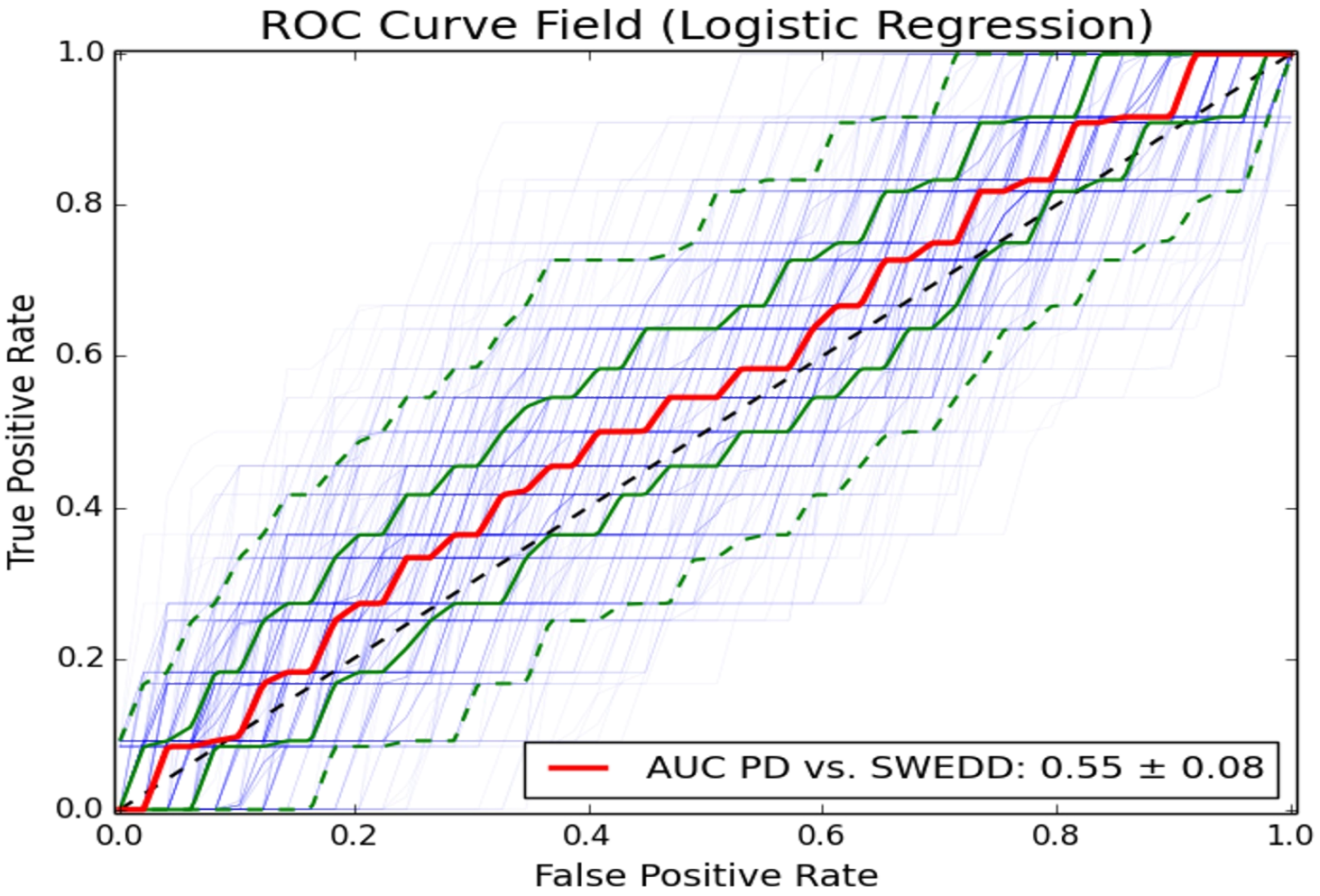
Significant increase of Amyloid-β_{42} in SWEDD cohort (compared to HC, PD) on average:





Cumulative probability (left) and probability density (right) distributions with Gauss filter.
Bottom: Raw data and cohort averages

CSF Protein Level Classifier for SWEDD



Cross-validation of regression classifier shows weak discriminatory power

Genetics Markers in PPMI

PPMI provides different types of genetics markers:

- Apolipoprotein E (*ApoE*) genetic allele data
  - Data for three common forms of cholesterol transporter: ε2, ε3, ε4
- Single-nucleotid polymorphisms (SNPs)
  - Genetic data for 33 point mutations in genes coding for PD-related proteins
- SNCA multiplication data
  - All but one subject had normal number of *SNCA* copies; therefore ignored
- Illumina NeuroX Array and ImmunoChip Array data
  - Data for ca. 500,000 SNPs (!), compressed in two PLINK files
  - Just started to decompress/analyze data (*Leandro Loss, Bayes Impact*)

Apolipoprotein E and Parkinson's Disease

- ε4/ε4 carriers are strongly at risk to develop Alzheimer's Disease:

  See e.g.: E. Corder *et al.*, Gene dose of ApoE type 4 allele and the risk of Alzheimer's disease in late onset families, Science **261** (1993), 921–923.

- Conflicting studies about *ApoE* allele ε2 as risk factor for PD:

  See e.g.: X. Huang *et al.*, APOE-ε2 allele associated with higher prevalence of sporadic PD, Neurology **62** (2004), 2198–2202.

  See e.g.: M. Federoff *et al.*,. A large study reveals no association between APOE and PD, Neurobiol. Dis. **46** (2012), 389–392.

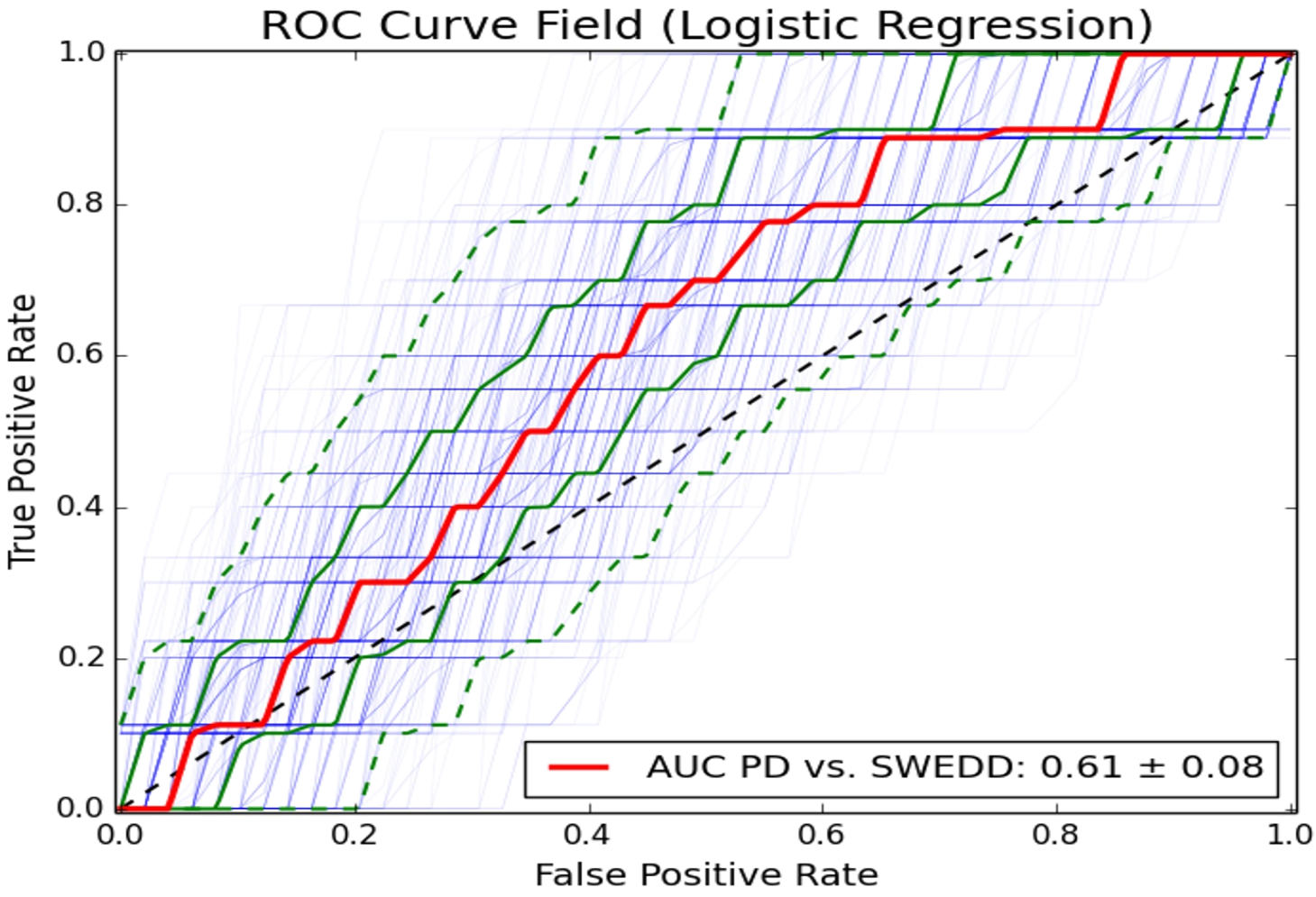
- PPMI: *ApoE* status matters for susceptibility to SWEDD, not to PD:

	ApoE ε2	*ApoE* ε3	*ApoE* ε4
Global average	**0.17**	**1.54**	**0.29**
Standard deviation	0.41	0.62	0.51
HC average	0.17	1.54	0.29
PD average	0.16	1.53	0.28
SWEDD average	0.23	1.35	0.42

Single-Nucleotid Polymorphisms in PPMI

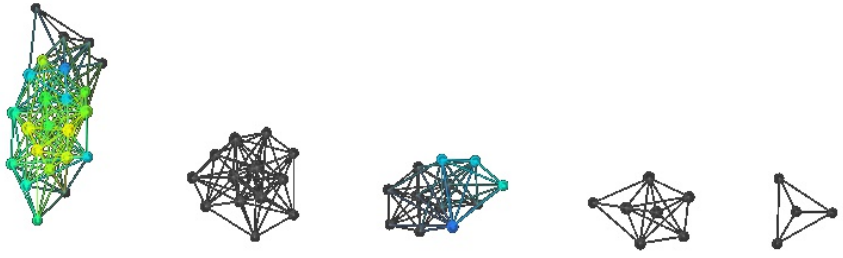
- PPMI biomarkers database has genotypes for 33 PD-related loci
- Examined relationship between PD, SWEDD, and genetic variants
- Result: Notable deviations from healthy cohort for several SNPs:
  - Some SNPs are risk factors for SWEDD, but not PD:
    Examples: *TMEM175, GCH1, NOTCH4* genes
  - Some SNPs are risk factors for PD, but not SWEDD:
    SNCA, LRRK2 among affected genes
  - A few SNPs are risk factors for both SWEDD and PD:
    In particular, strong risk associated with gene complex *SIPA1L2*
- Caveat: Small size of SWEDD cohort impairs statistical significance

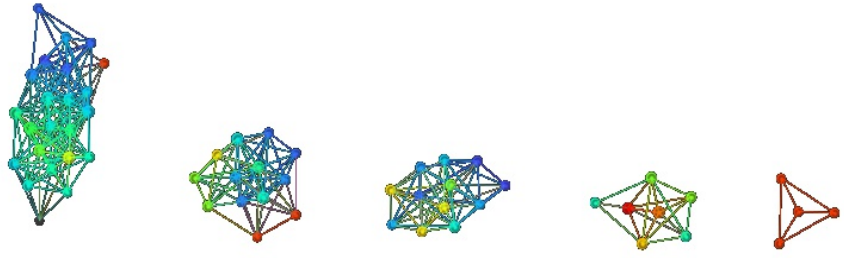
Evaluating a Genetic Classifier for SWEDD



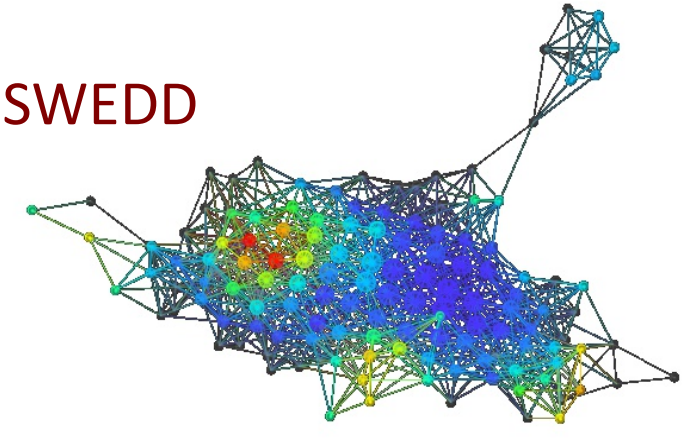
Genotype information provides only a weak classification mechanism

AYASDI Analysis: Genotype Data

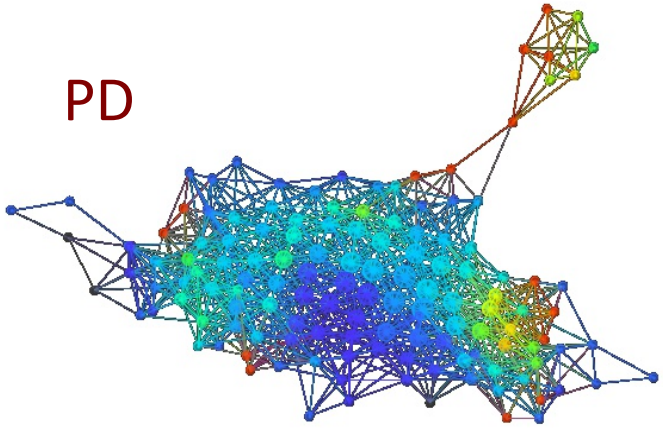




SWEDD



PD



Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

Part III:

Visualizing the SWEDD Population

Characterizing the SWEDD Population

- "We're able to identify SWEDD. But what is it?"
  - What are the defining features of SWEDD?
  - Is SWEDD a single, clearly demarcated syndrome? A set of distinct, independent conditions? A "continuum of cases"?
  - What are risk factors for SWEDD? How do they differ from PD?
- Employ "unsupervised learning" algorithms:
  - Feature reduction: Remove irrelevant "dimensions" from data
  - Identifying "clusters" of related data points
  - Visualization of higher dimensional data

Visualization by Feature Reduction

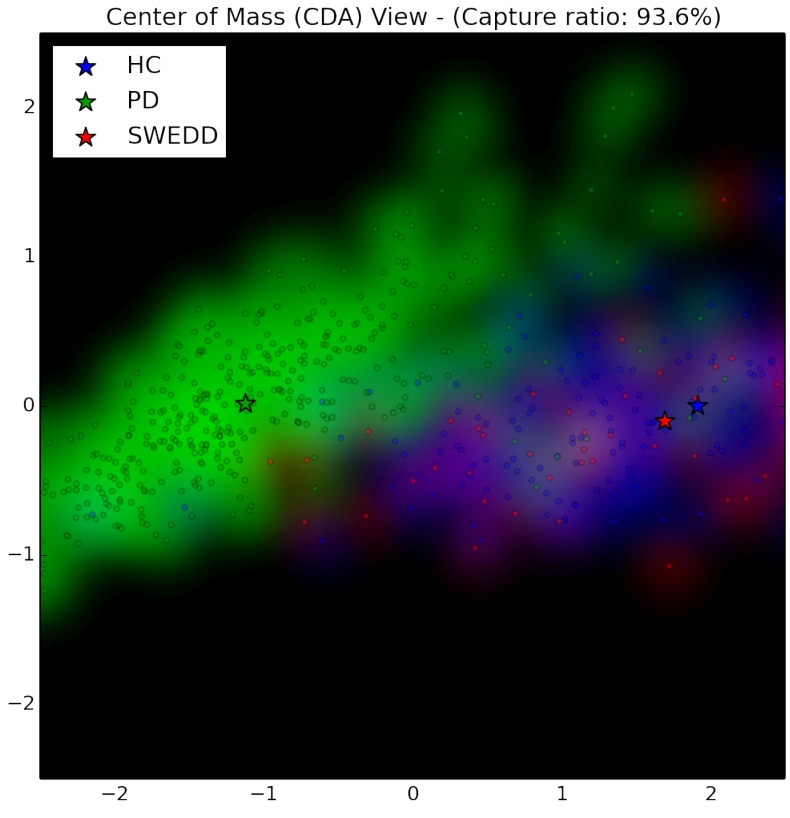
"Mechanism" behind classification schemes?

- Data "lives" in high-dimensional feature space; difficult to visualize

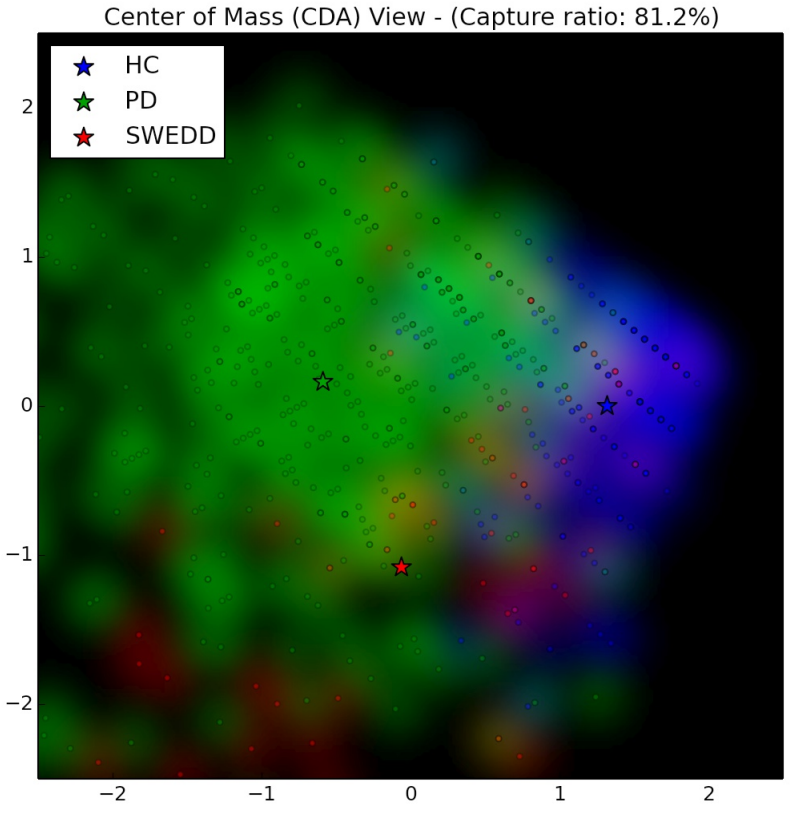
Idea: "Project" data for planar representation (plot)

- Linear method #1: Principal Component Analysis (PCA)
  Projection on plane that most closely aligns to data points in feature space
- Linear method #2: Canonical Discriminant Analysis (CDA)
  Projection on plane spanned by cohort averages ("center of mass"), faithfully reproduces distances between cohort averages in feature space
- Non-linear methods: Clustering, graphs, ...
  Examples: *k*-means, stochastic embedding, topological data analysis

Projection: DaTSCAN, Olfactory-Motor-Adaptation



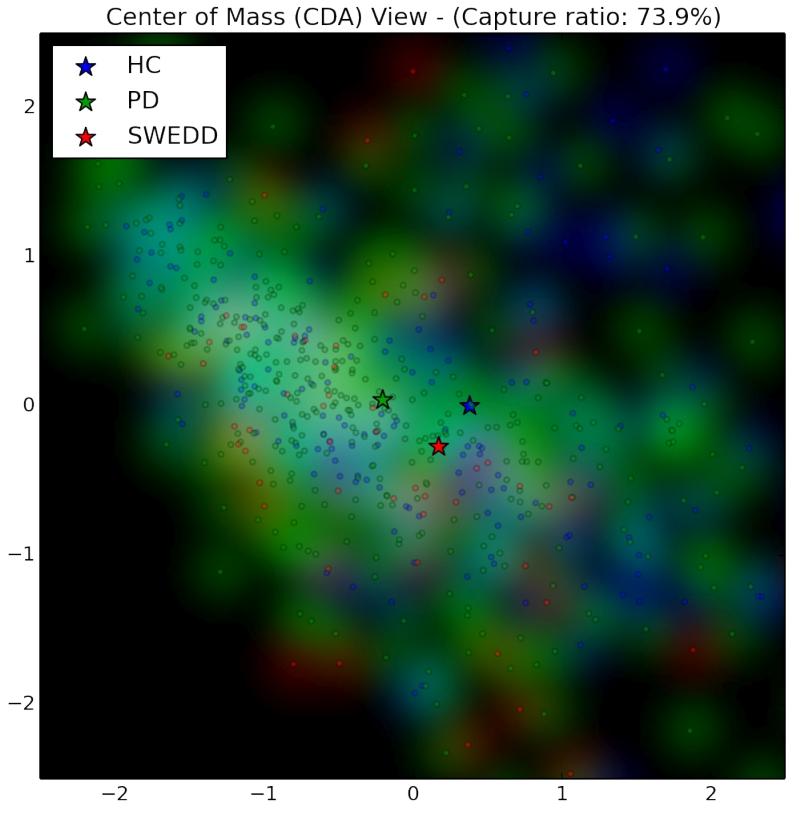
DaTSCAN classification



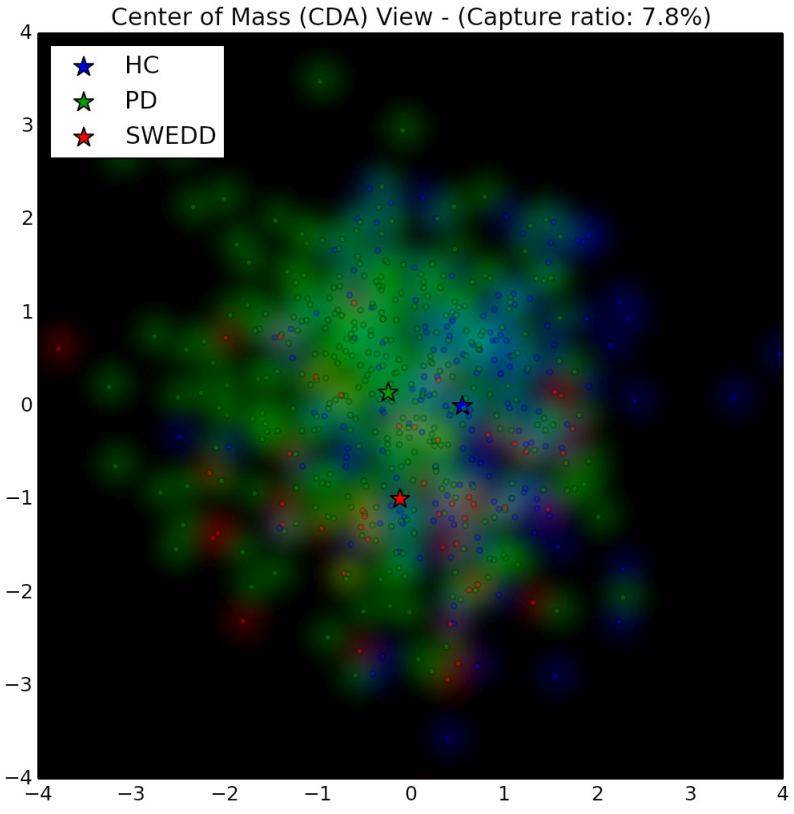
Olfactory – Fine Motor – Adaptation scheme

Sharp separation in DaTSCAN graph reflects PPMI classification scheme

Projection: CSF Proteins, Genetics



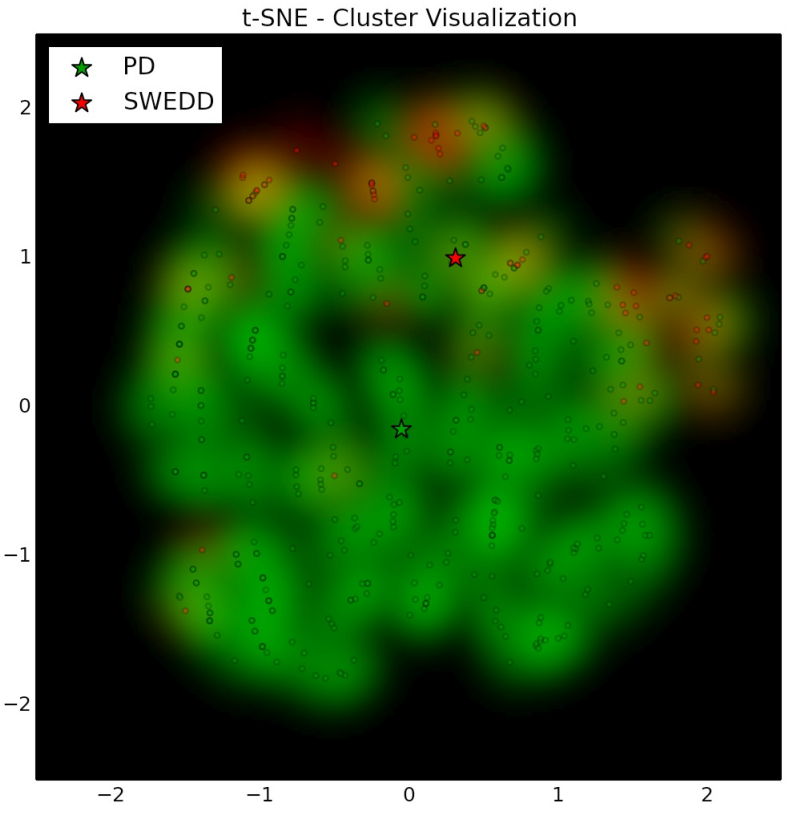
CSF Protein Level Classification



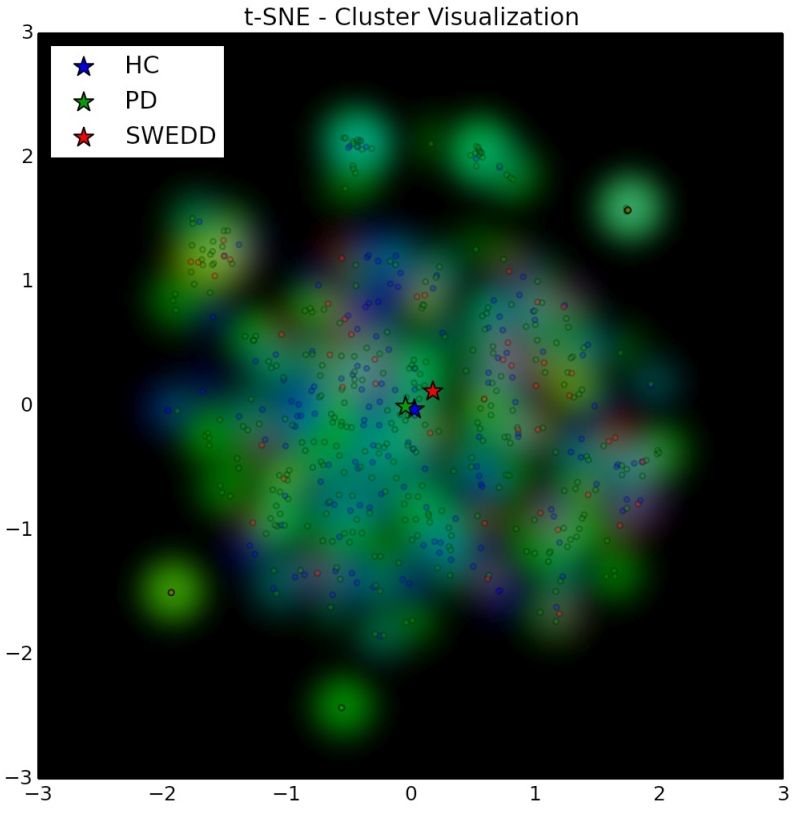
Genetics Classification

Lack of separation of cohorts “explains” poor performance of classifiers

Clustering Analysis: Subgroups of SWEDD?



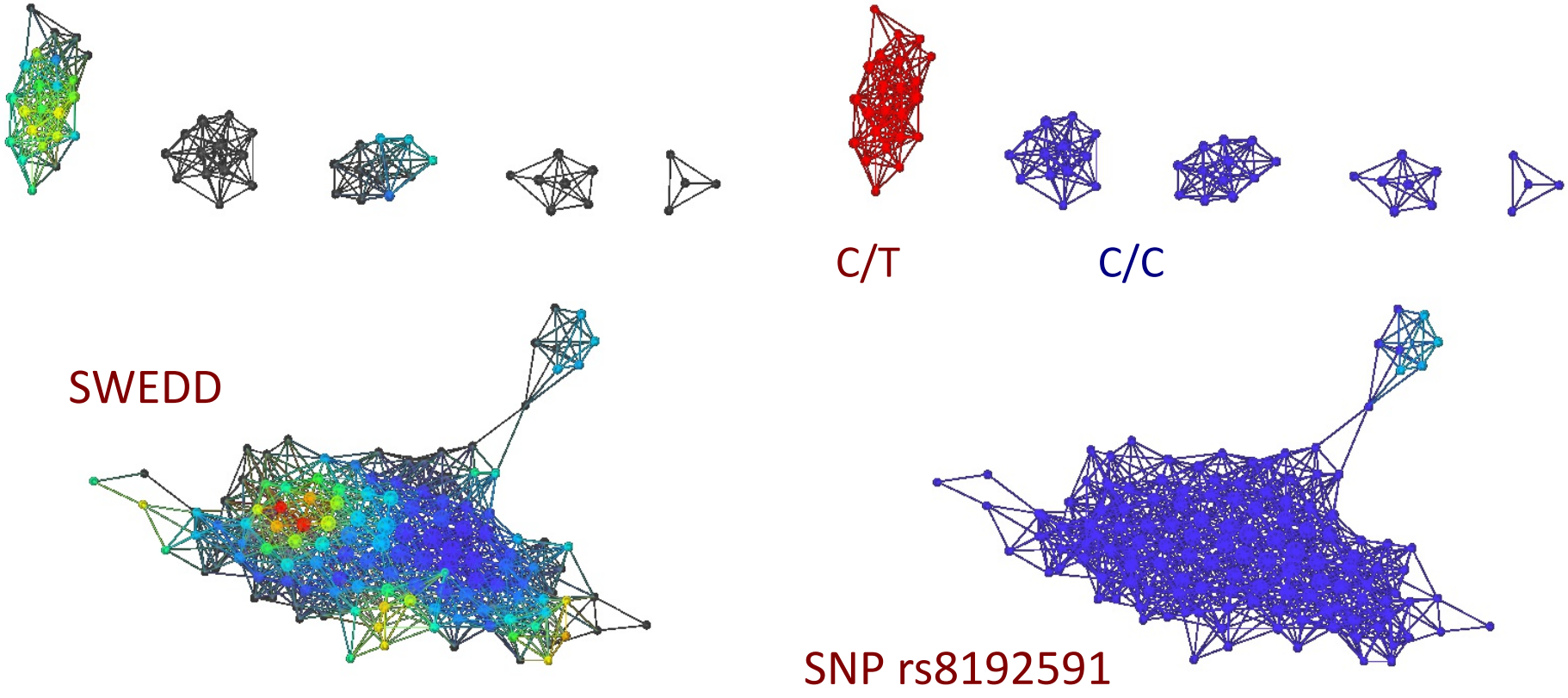
Olfactory – Fine Motor – Adaptation scheme



Genetic classification

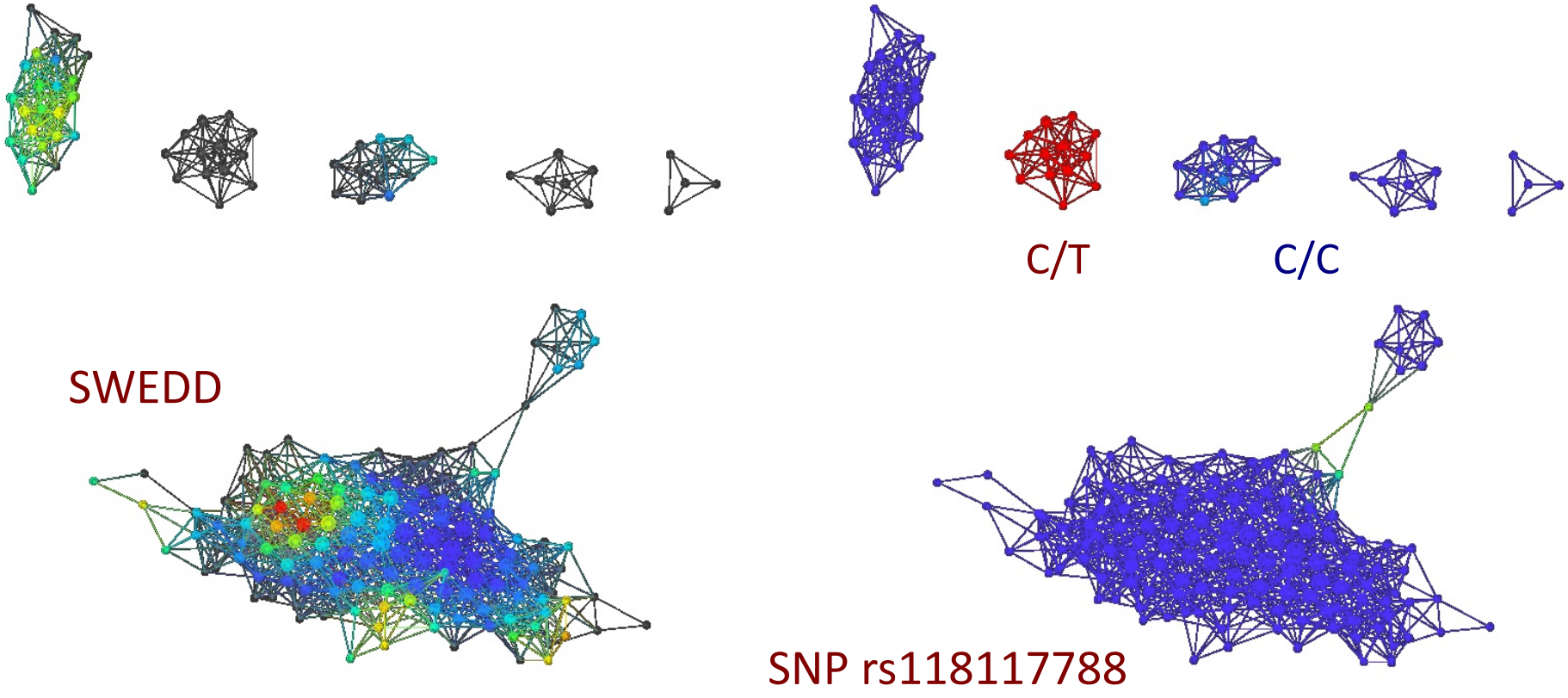
Tentatively: Are there two separate clusters of SWEDD patients? Effects of genotype?

AYASDI Analysis: Genotype Data Examples



The C/T mixed genotype for SNP *rs*8192591 (*Notch4*) is overrepresented for SWEDD

AYASDI Analysis: Genotype Data Examples



The C/T mixed genotype for SNP *rs*118117788 (*MicroRNA 4682*) has no SWEDD subjects

Summary

Summary: Assessing Markers for SWEDD

- PPMI identifies SWEDD patients by exclusion through DaTSCAN.
- I applied machine learning methods to PPMI data to evaluate markers suggested in the PD literature.
- A combined screen for olfactory ability, fine motor problems, and adaptation difficulties yields an independent predictor for SWEDD.
- CSF levels of PD-related proteins, while significantly altered on average, are not good predictors for disease status.
- Tentatively, SWEDD subjects may form two separate clusters.
- PPMI data suggests that SWEDD has distinct genetic risk factors from PD, including involvement of Apolipoprotein E.

Shortcomings and Future Work

Problem: The size of the SWEDD sample is small.

- There are only 50 – 60 SWEDD patients in PPMI
- Low count increases uncertainty, affects statistical significance
- Suggestion: Add data from designated study, other studies.

Problem: The genetic classifier is based on few markers.

- Only 33 genetic loci have been studied.
- Suggestion: Unpack array data, extend analysis to ~500,000 loci.

Problem: Obtain additional clinical markers for SWEDD.

- There are some obvious directions, e.g., record response of subjects to a dose of PD medication (levodopa).

Acknowledgments

For their assistance with this project, I would like to thank:

- Eric Liu, Paul Duan, and Andrew Jiang (Bayes Impact)
- Ken Kubota and the Michael J. Fox Foundation
- The Parkinson's Progression Markers Initiative Team
- Devi Ramanan and Pek Lum (Ayasdi)
- Mijail Gomez and Leandro Loss (Bayes Impact Fellows)
- Vesela Gateva and Uri Laserson (mentors)

Appendix A:

Machine Learning in a Nutshell

Appendix: Machine Learning in a Nutshell

- Supervised learning:
  - Training and classification
  - Cross-validation
  - Evaluation: ROC curves
- Unsupervised learning:
  - Feature reduction
  - Clustering
  - Visualization

"Supervised Learning": A Primer

- Principle of supervised learning:
  - "Training": Feed computer algorithm with data (e.g., from DaTSCAN), and classification of outcome (here: healthy, PD, SWEDD) for a sample of subjects
  - "Classification": Trained algorithm uses the equivalent data set for a distinct sample of subjects to predict their disease status
- Quality of a classification scheme:
  - Strong classifiers predicts the correct disease status of most subjects in the sample: Few "false positives" and "negatives"
  - Weak classifiers perform little better than random guessing

"Supervised Learning": A Primer

- Algorithms really yield probabilities:
  - Output of classification algorithm is a set of probabilities for class membership, calculated from input data:

	prob(PD)	prob(SWEDD)
Subject #1	0.75	0.25
Subject #2	0.10	0.90
Subject #3	0.45	0.55

  - Use threshold to translate probabilities into cohorts:

    e.g., "subject is classified as SWEDD if $p_{SWEDD} > 0.75$"

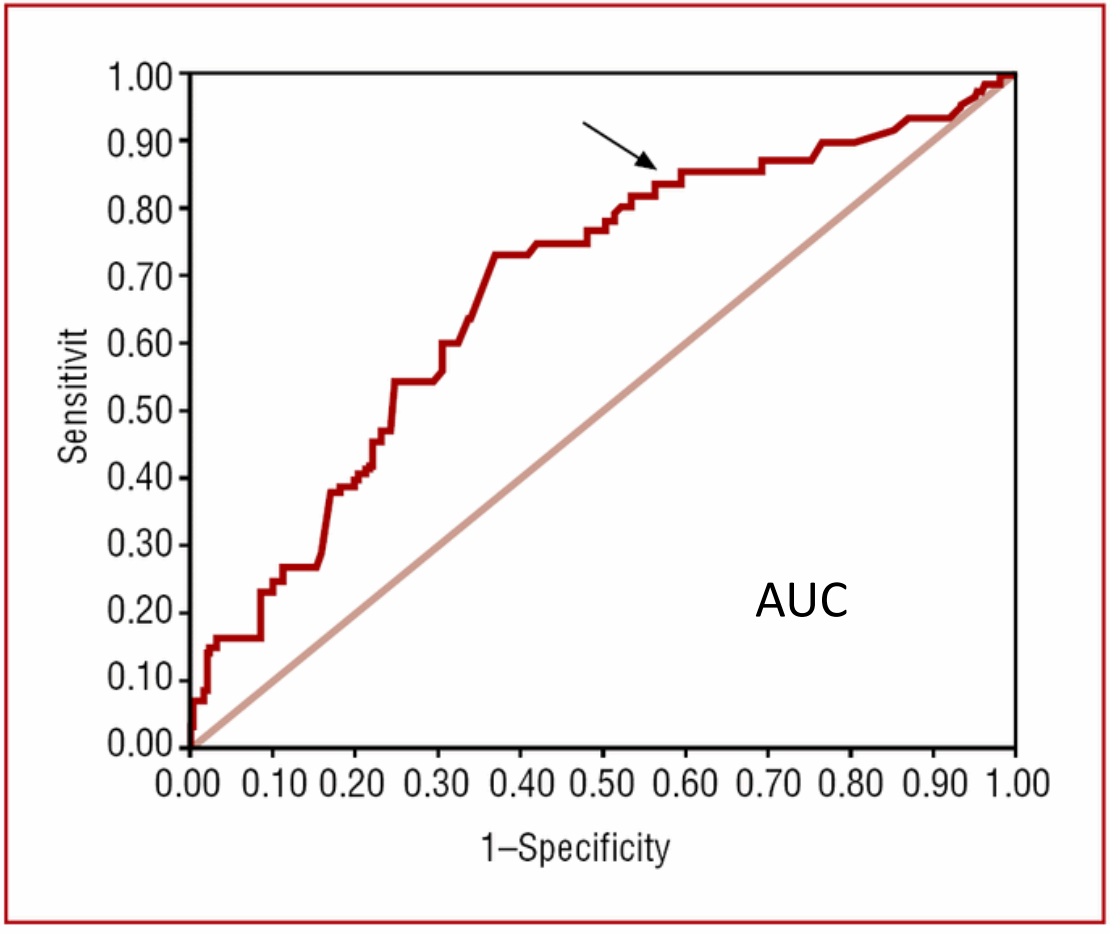
  - Classification depends on threshold value
  - Generally, trade-off between false positives and negatives

Assessing Classifiers: The ROC Curve

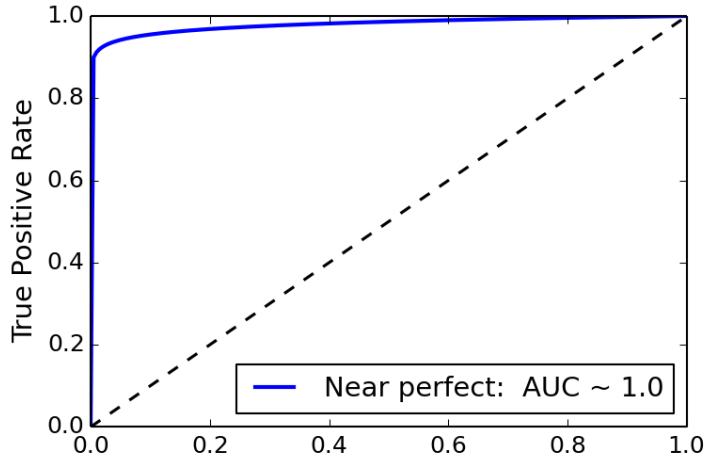
"Receiver Operating Characteristic" (ROC):

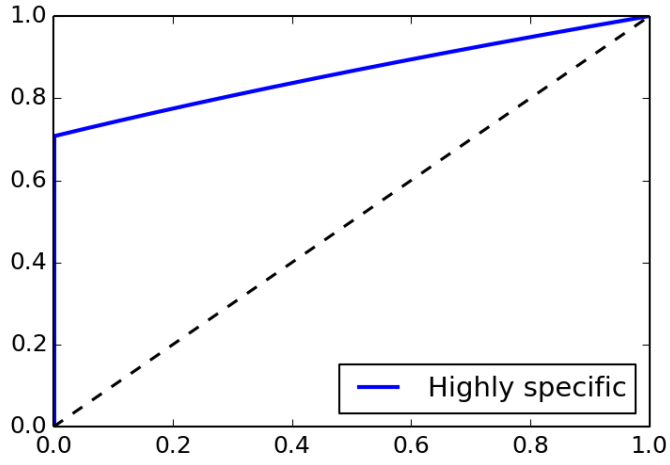
Compares rates of assigning subjects to classes correctly as the probability threshold value changes from zero to one

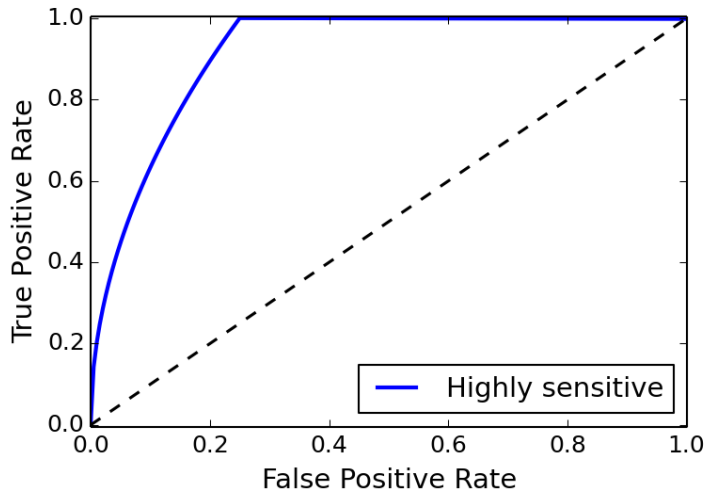
AUC (area under curve) is a rough measure of strength of classification scheme

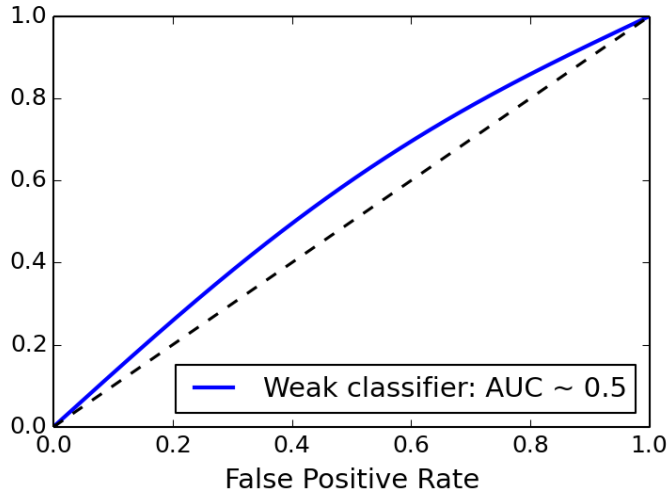


Discussing ROC Curves









"Unsupervised Learning": A Primer

- Principle of unsupervised learning:
  - Feed computer algorithm with high-dimensional data set (e.g., genotype) to analyze the "geometry" of the data
  - There is no "classification" or assignment of an outcome
- Goals of unsupervised learning:
  - Feature reduction: Remove irrelevant "dimensions" from data
  - Identifying "clusters" of related data points
  - Visualization of high-dimensional data

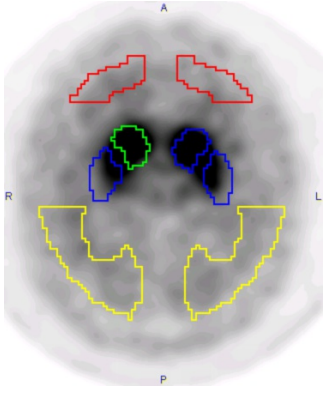
Appendix B:

Subject Classification with the DaTSCAN Protocol

Classification of SWEDD Using DaTSCAN

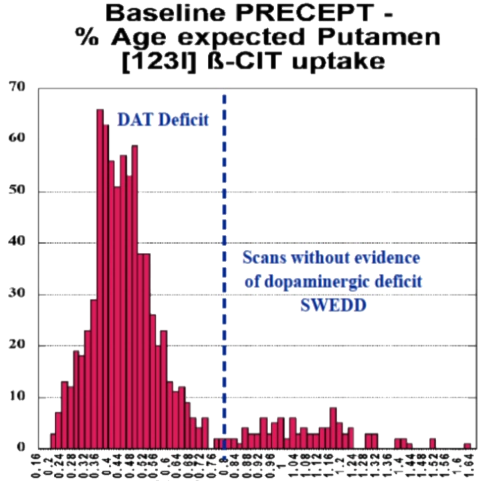
DaTSCAN (Dopamine Transporter SPECT Scan)

→



Find "Striatal Binding Ratios" (4 parameters)

→



> 80% of "normal"?

↗

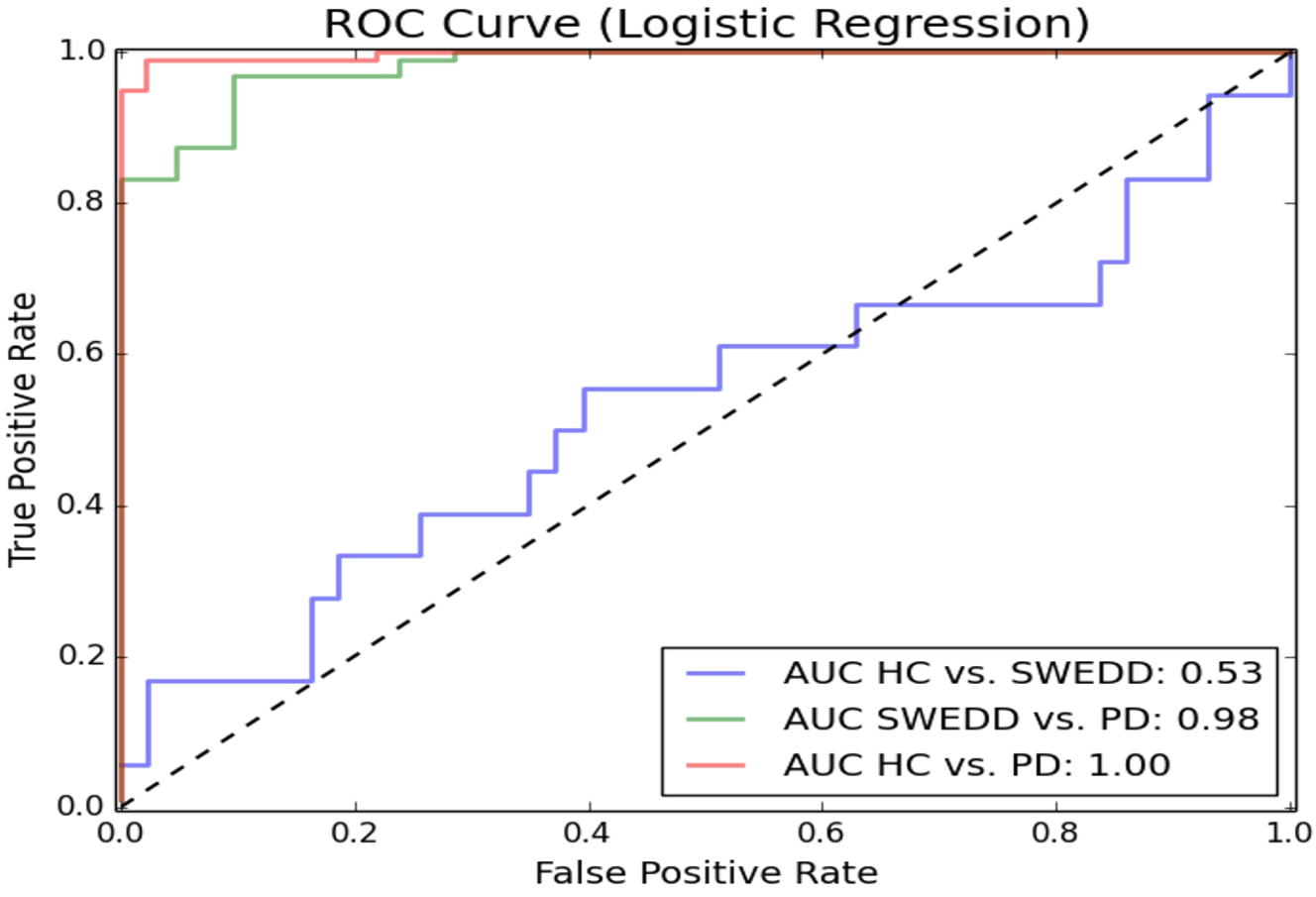
PD (85 %)

↘

SWEDD (15 %)

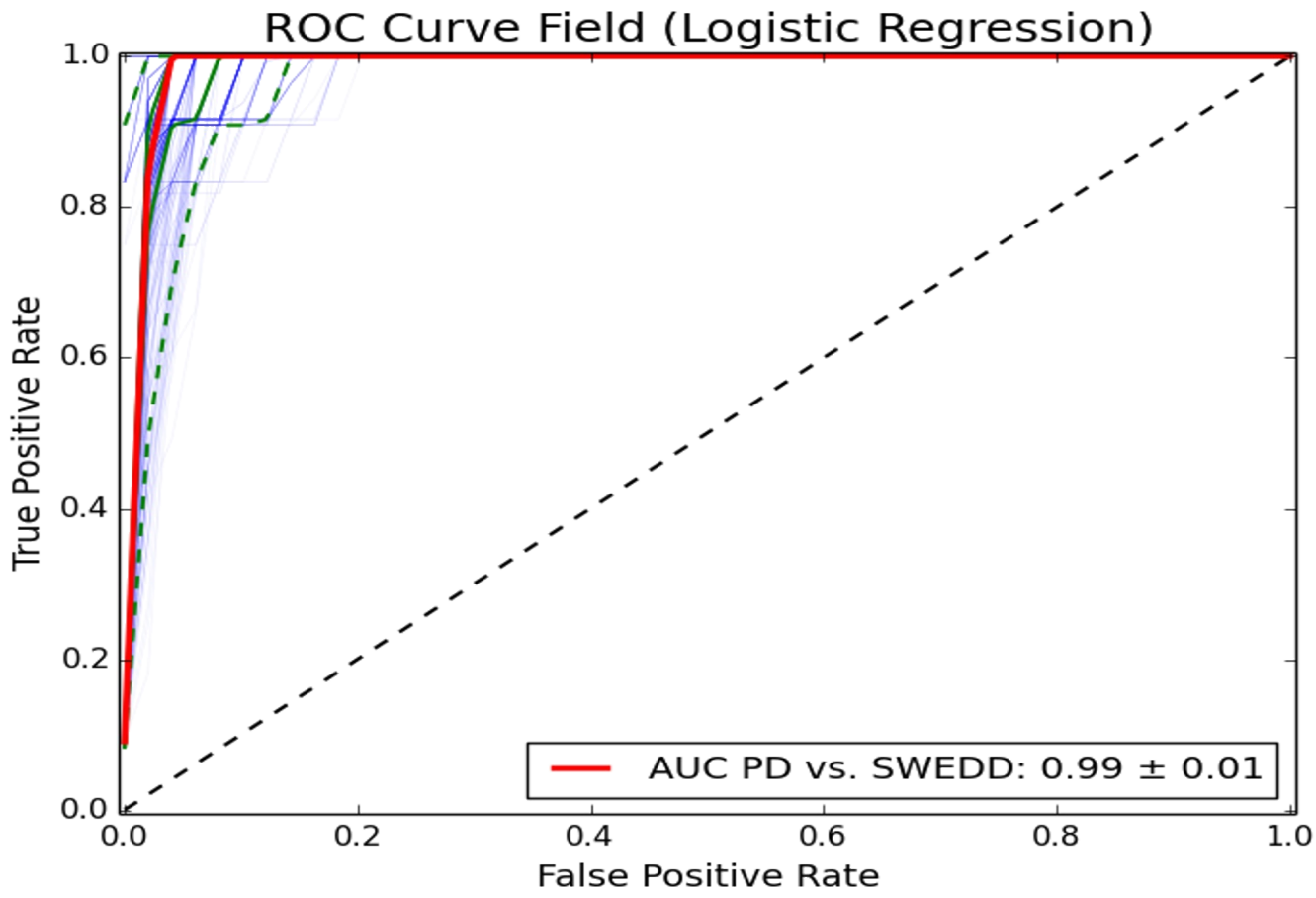
Note: Putamen only – sensitive to "tail" of structure

Classification of SWEDD Using DaTSCAN



DaTSCAN classification fails to separate Healthy Control (HC) and SWEDD subjects

Classification of SWEDD Using DaTSCAN



Random assignment of subjects (shuffle-split cross validation) shows performance range

Appendix C:

Olfactory – Fine Motor – Adaptation Classifier Development & Detail

Olfactory – REM Sleep – Motor Screens

Features contained in this test set (evaluated at baseline):

- UPSIT Olfactory Test (1 feature)
  - Identify 40 different smells (score from 0 to 40)
- REM Sleep Behavior Screening Questionnaire (12 features)
  - Aggressive Dreams, Complex Movements, Disturbed Sleep, Hurt Bed Partner, Move Arms/Legs, Movements Awake Me, Nocturnal Behavior, Remember Dreams, Speaking in Sleep, Sudden Limb Movements, Things Fell Down, Vivid Dreams
- Motor Difficulties Questionnaire (UPDRS Part II, 13 features)
  - Dressing, Eating, Freezing, Getting up, Handwriting, Hobbies, Hygiene, Saliva, Speech, Swallowing, Tremor, Turning in Bed, Walking & Balance

The SCOPA-AUT Screening Questionnaire

Scale for Outcomes in Parkinson's Disease – Autonomic (21 features):

- Eating/Ingestion:
  - Swallowing, Salivation, Food Gets Stuck, Feeling Full Quickly
- Digestion:
  - Constipation, Hard Stools
- Hygiene:
  - Loss of Stool, Loss of Urine, Difficulty Retaining Urine
- Urination:
  - Bladder Not Empty, Weak Urine Stream, Pass Urine Often, Urinate at Night
- Adaptation:
  - Lightheadedness, Dizzy Standing Up, Fainting, Perspiration at Daytime, Perspiration at Night, Bright Light Sensitivity, Tolerating Cold, Tolerating Heat

Preparing a Clinical Classifier for SWEDD

Problem: Avoid overfitting?

- There are 46 features for ≈ 60 SWEDD subjects in the study

Strategies to improve and simplify the classification system:

- Feature Elimination
  - Remove features that are unspecific, e.g., dream-related questions
- Feature Aggregation
  - Sum up scores for questions that address related problems and show similar response patterns, e.g., lump *Dressing, Hygiene, Handwriting, Hobbies, Tremor* in UPDRS-2 into a single *Fine Motor Skills* score
  - Avoid bias by pre-selection of groups

Parameters for the Clinical Classifier

Input: Scores from screening, questionnaires:

- UPSIT olfactory score (*OLS*, range 0 – 40)
- Fine motor questionnaire score (*FMS*, range 0 – 20)
- Autonomous/adaptation screen score (*AAS*, range 0 – 32)

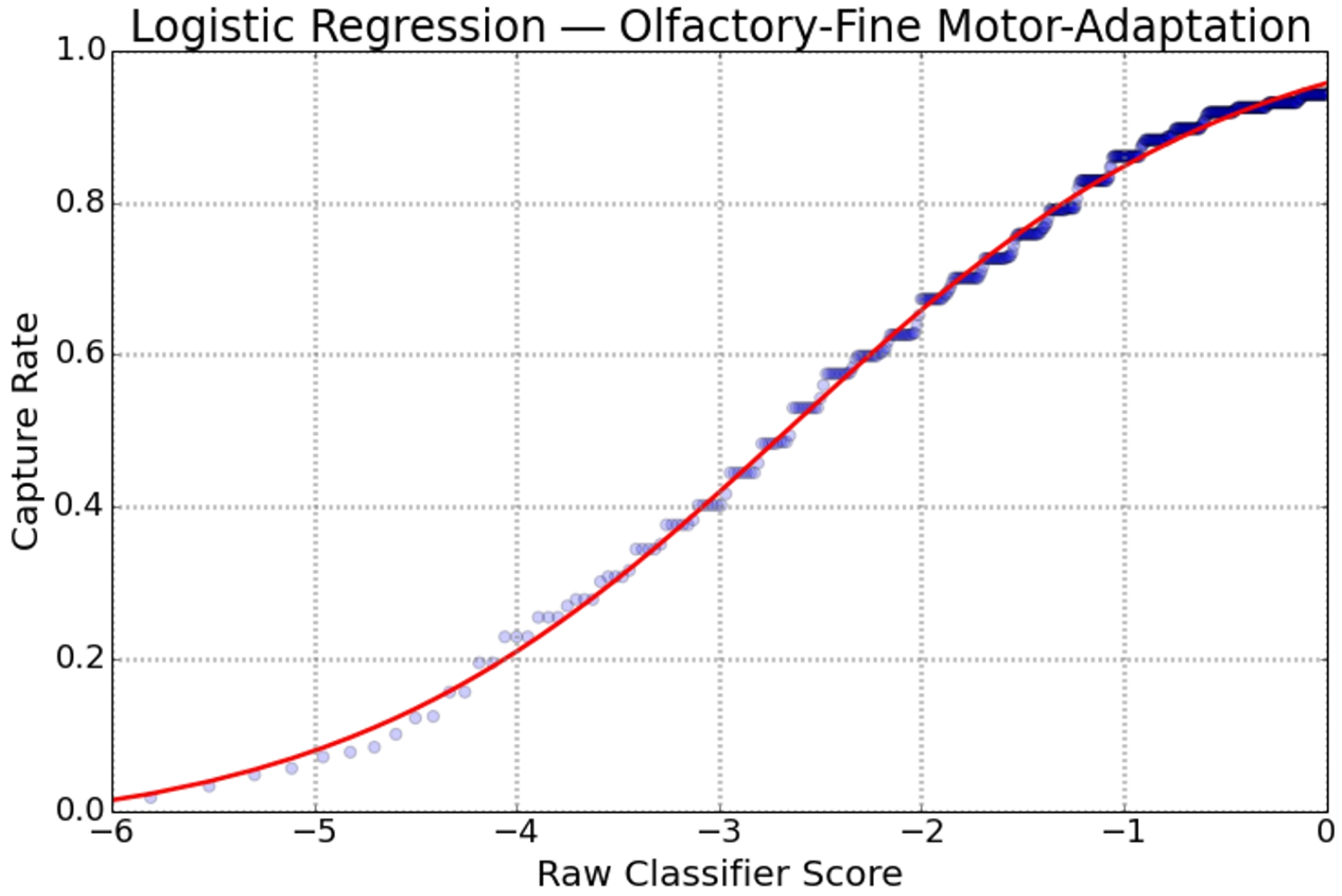
Output: Calculate raw score *z* by regression equation:

$$z = -6.585 + 0.158 \cdot OLS - 0.159 \cdot FMS + 0.311 \cdot AAS$$

Interpretation:

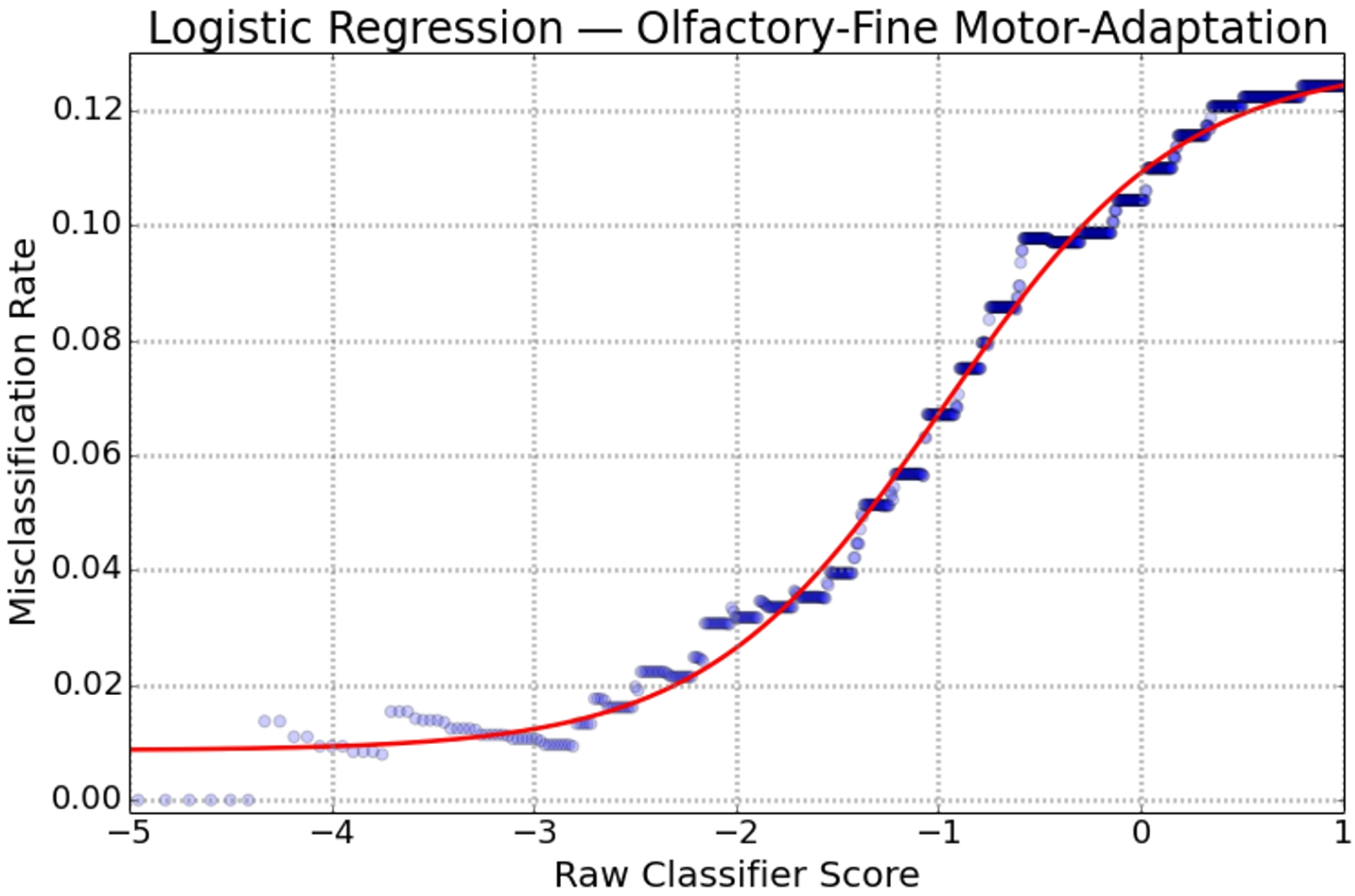
- Smaller values of *z* imply higher likelihood for PD
- Logistic model: $prob(\text{PD}|z) = 1 / (1 + \exp(z))$
- Cutoff values for *z* determine capture ratio, false classification rate

Clinical Classifier: Raw Score vs. Capture Ratio



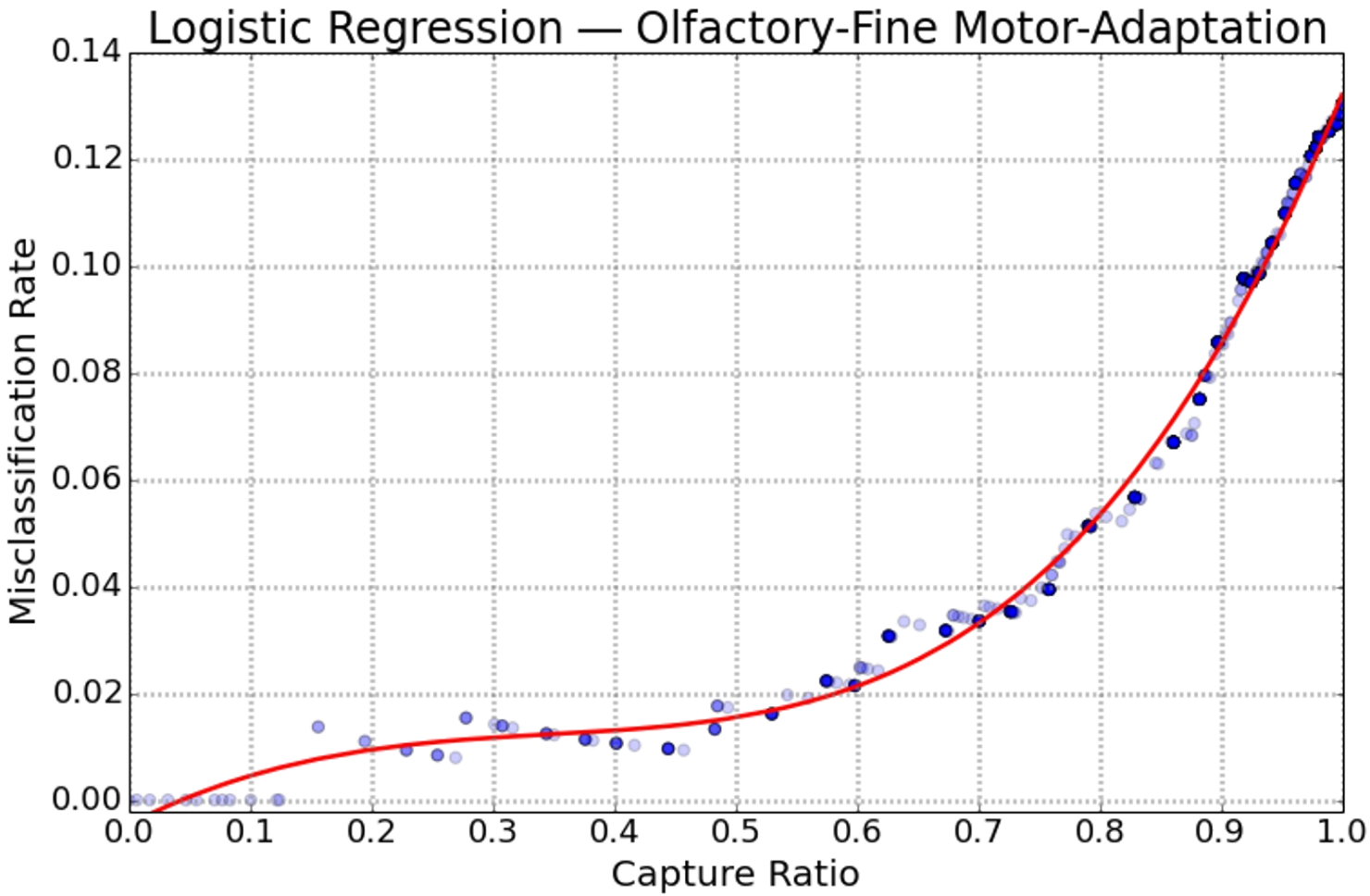
Scores $z < -2.6$ capture around half of the non-healthy cohort populations

Raw Score vs. False Classification Rate



For scores $z < -2.5$, less than 2% of the captured subjects are SWEDD

Capture Ratio vs. False Classification Rate



Possible to capture half of the PD cohort with < 2% SWEDD admixture!

Appendix D:

Statistical Significance of Selected Genetic Loci

Hypergeometric *p*-Values for *ApoE* Genotypes

Type	Significance?	*p*(PD vs HC)	*p*(SWEDD vs HC)
ε2/ε2	PD	0.033 (–)	0.370 (+)
ε2/ε3		0.141 (+)	0.455 (–)
ε2/ε4		0.066 (–)	0.403 (+)
ε3/ε3	SWEDD	0.375 (–)	0.023 (–)
ε3/ε4		0.499 (+)	0.129 (+)
ε4/ε4		0.364 (–)	0.470 (+)

Rare genotypes in grey

Source: AYASDI analysis

Hypergeometric *p*-Values for Selected SNPs

SNP	Type	Gene	Significance?	*p*(PD vs HC)	*p*(SWEDD vs HC)
rs10797576	C/C	SIPA1L2	SWEDD PD	0.008 (–)	0.001 (–)
rs10797576	C/T	SIPA1L2	SWEDD PD	0.013 (+)	0.001 (+)
rs11158026	T/T	GCH1	SWEDD	0.165 (–)	0.002 (–)
rs115462410	C/C	HLA-DQB1	PD	0.008 (+)	0.264 (–)
rs115462410	C/T	HLA-DQB1	PD	0.004 (–)	0.447 (+)
rs11724635	A/A	BST1	PD	0.006 (+)	0.219 (–)
rs11724635	G/G	BST1	PD	0.014 (–)	0.466 (+)
rs2414739	G/G	*Intergenic, c15*	SWEDD PD	0.015 (–)	0.004 (–)

Source: AYASDI analysis

Hypergeometric *p*-Values for Selected SNPs

SNP	Type	Gene	Significance?	*p*(PD vs HC)	*p*(SWEDD vs HC)
rs34311866	A/A	TMEM175	SWEDD	0.087 (–)	0.002 (–)
rs34311866	A/G	TMEM175	SWEDD	0.330 (+)	0.018 (+)
rs356181	C/C	SNCA	PD	0.004 (+)	0.266 (–)
rs3910105	C/C	SNCA	PD	0.010 (–)	0.449 (–)
rs6430538	C/C	*Intergenic, c2*	SWEDD	0.192 (+)	0.007 (+)
rs6430538	T/T	*Intergenic, c2*	SWEDD PD	0.013 (–)	0.020 (–)
rs76904798	C/T	LRRK2	PD	0.006 (+)	0.078 (+)
rs76904798	C/C	LRRK2	PD	0.003 (–)	0.024 (–)
rs8192591	C/C	NOTCH4	SWEDD	0.030 (–)	0.004 (–)

Source: AYASDI analysis

Appendix E:

Topological Data Analysis with AYASDI

Topological Data Analysis with AYASDI – Concepts

Metrics are used with lenses to construct the Ayasdi graph:

- A metric represents a notion of similarity (distance) between rows in the data (i.e., data points).
- A lens is a filter that converts the data set into a vector, with each row in the original data set represented by a coordinate. Basically, a lens operation turns every data point into a single number.
- Ayasdi creates the network or graph using the distances between pairs of data points, and their coordinates from a pair of lenses.

Topological Data Analysis – Metrics and Lenses

Metric used in the Ayasdi analyses in this talk:

- Variance Normalized Euclidean:

  Each feature score is first shifted and scaled so its mean is zero, and its variance is unity. Then, the Pythagorean theorem is used to determine the distance between two "data points" (subjects).

Lenses used in the Ayasdi analyses in this talk:

- Neighborhood Lenses #1 and #2:

  These lenses generate a projection of high-dimensional data into two dimensions by embedding a k-nearest neighbors graph of the data, using Ayasdi's proprietary graph layout algorithm.

Contact Information

- Christian Bracher

cbracher69@gmail.com

http://www.christianbracher.net/

http://www.linkedin.com/in/christianbracher/

https://github.com/cbracher69/PD-Learn/