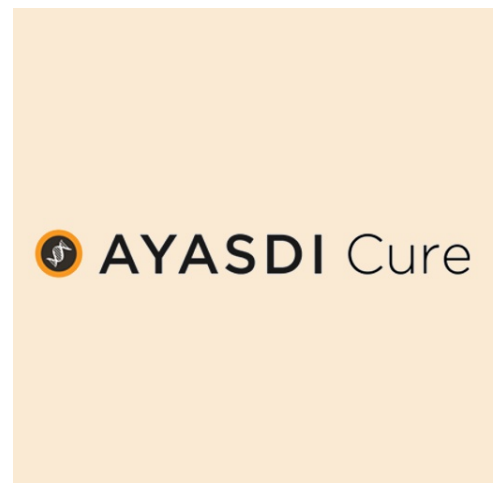# Project PPMI — A Report

Christian Bracher
Mijail Gomez

August 17, 2014

Acknowledgments

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

Overview

- Parkinson's Disease — A Challenge

- Outline of the PPMI Study

- Evaluating Markers of Parkinson's Disease

- The Olfactory – REM Sleep – Motor Symptom Triad

- Leveraging Social Networks for Parkinson's Research

- A Machine Learning Web Application for PPMI

- Suggestions for Future Investigations

Parkinson's Disease — A Challenge

PD presents multiple challenges:

- Progressive, fatal disease; no cure

- Medications only treat symptoms

- No definitive diagnosis without autopsy

- Causes of PD are mostly unknown

- Possibility of distinct PD subtypes

- Early detection of PD
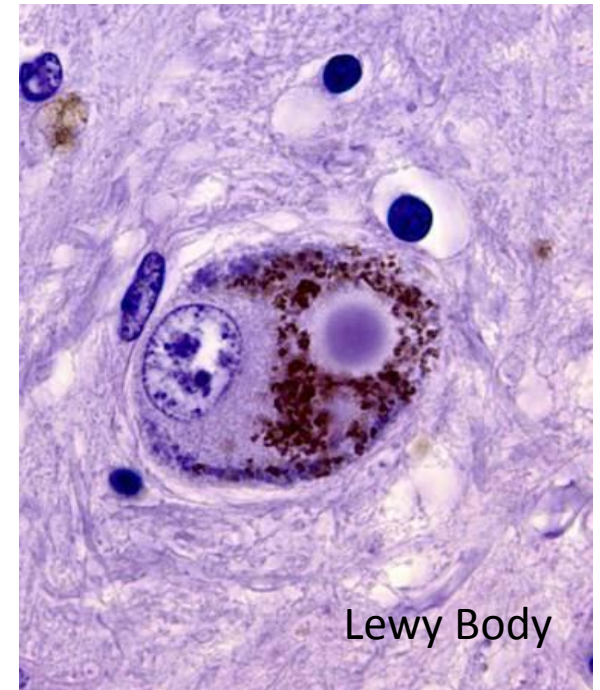
Lewy Body

Parkinson's Progression Marker Initiative

PPMI is designed to yield answers to these challenges:

- $60 M four year longitudinal, clinical and observational study

- Massive horizontal data set:
  Biomarker, psychological, neurological, imaging, genetics data

- Several cohorts with altogether ~ 1,000 subjects

Use machine learning, advanced statistical techniques to evaluate data, find patterns

Evaluating Markers of Parkinson's Disease

Many physiological changes have been linked to PD:

- Brain function (dopamine response)

- Lipid metabolism

- Alteration in protein levels in cerebrospinal fluid (CSF)

- Genetic variants, gene expression

- Neurological abnormalities:
  Olfactory, sleep, spatiovisual capability, ...

- Psychological dysfunction (depression, ...)

We examine some of these claims using the PPMI data set.

SPECT Brain Imaging (DaTSCAN)

- "Gold Standard" in PD diagnosis

- Expensive ($5,000)

- Requires specialized equipment:

  - Radioactive marker

  - SPECT imaging

- Not widely available

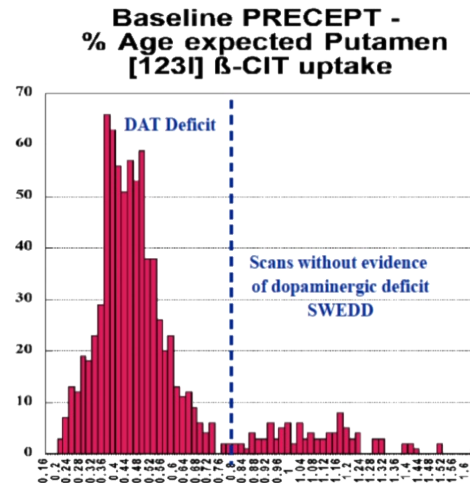HEALTHY   PARKINSON'S
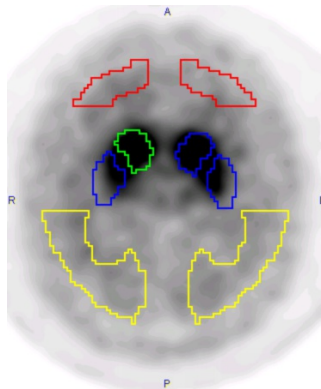
SPECT Brain Imaging (DaTSCAN)

- But DaTSCAN is far from perfect:

  – Healthy subjects pass DaTSCAN, but

  – 15 % of PD subjects are "misclassified"

- DaTSCAN unveils a new subject cohort:

  SWEDD — subjects without evidence of dopaminergic dysfunction

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

SWEDD

Classifying SWEDD Using DaTSCAN
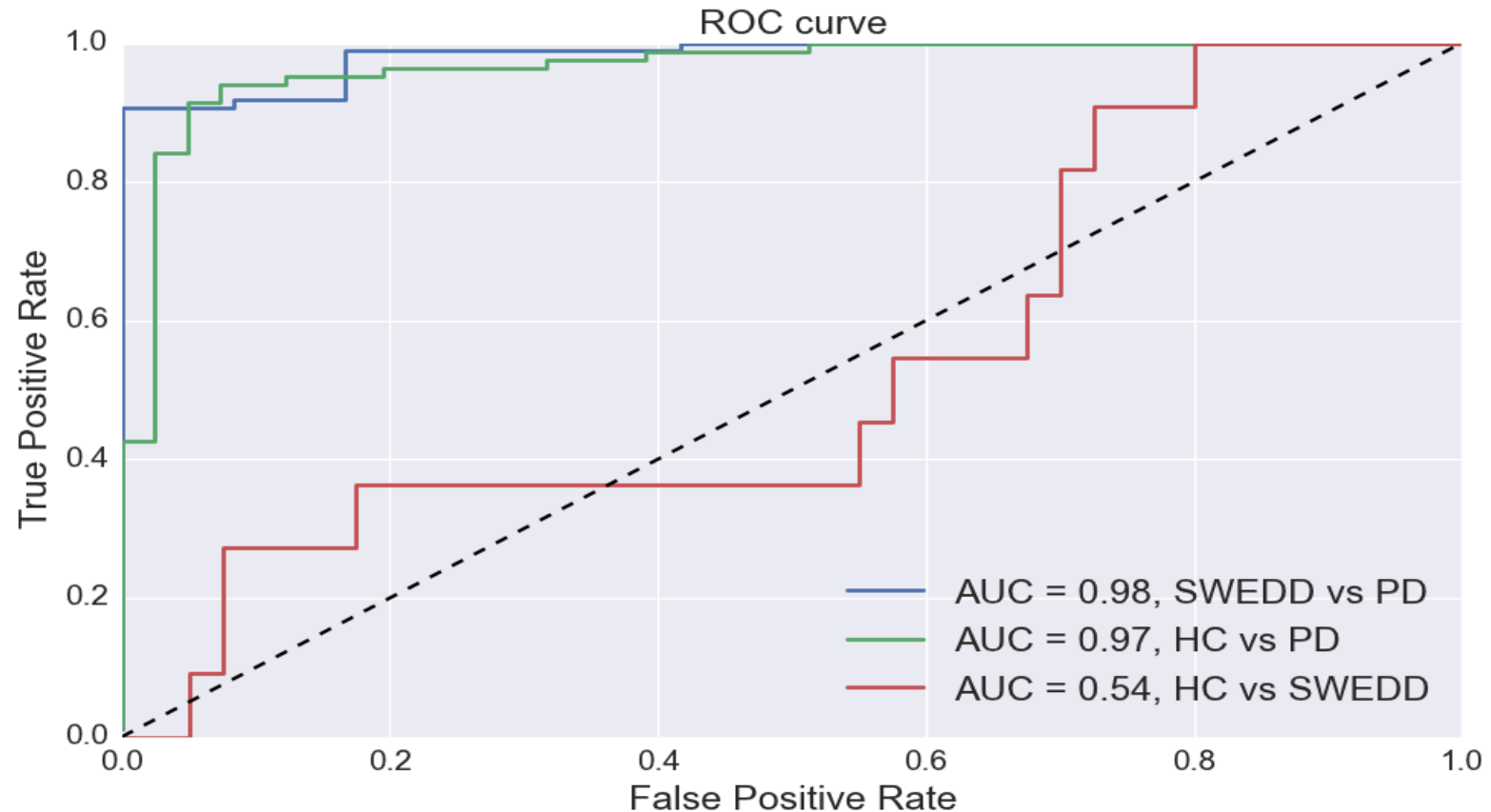

DaTSCAN
(Dopamine
Transporter
SPECT Scan)

→

Find "Striatal Binding
Ratios" (4 parameters)

→

**Baseline PRECEPT -
% Age expected Putamen
[123I] ß-CIT uptake**

DAT Deficit

Scans without evidence
of dopaminergic deficit
SWEDD

> 80% of "normal"?

↗ PD (85 %)

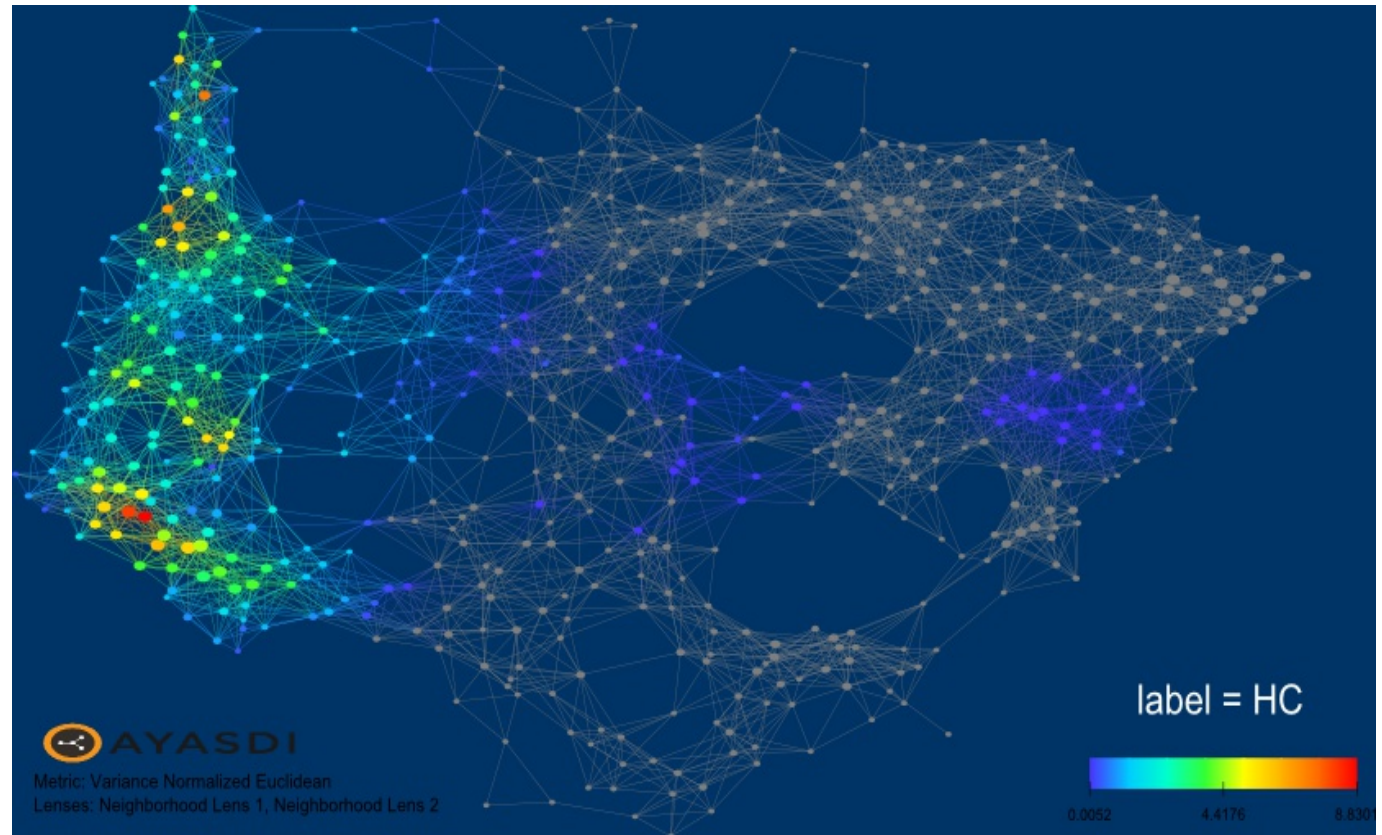↘ SWEDD (15 %)

How Well Does DaTSCAN Perform?

Supervised learning using DaTSCAN fails to separate Healthy Control (HC) and SWEDD subjects

Topological Data Analysis – Healthy Cohort

Analysis:

- DaTSCAN data (4 feature set)

- Healthy subjects

label = HC

Metric: Variance Normalized Euclidean
Lenses: Neighborhood Lens 1, Neighborhood Lens 2
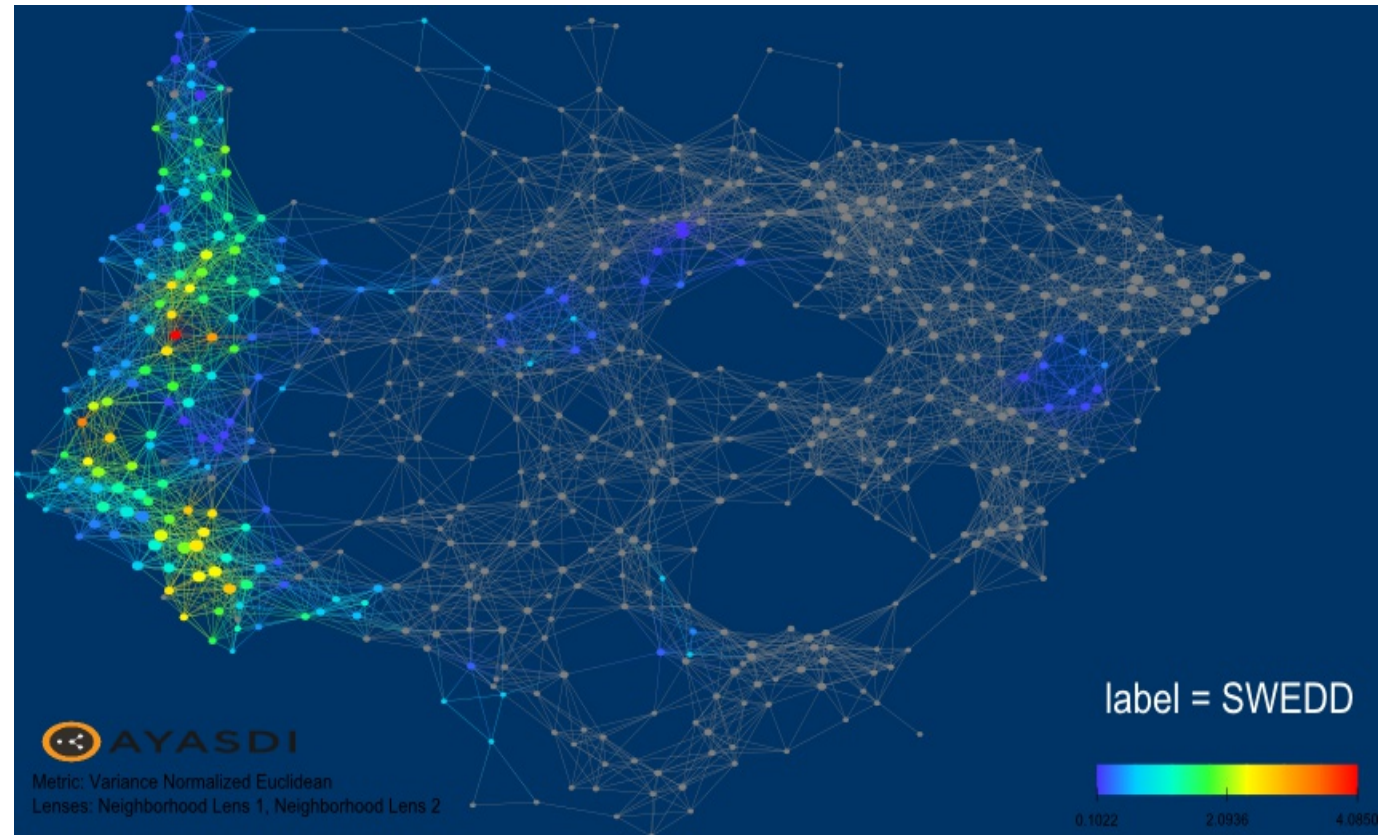
0.0052   4.4176   8.8301

Ayasdi — Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

Topological Data Analysis – SWEDD

Analysis:

- DaTSCAN data (4 feature set)

- SWEDD subjects

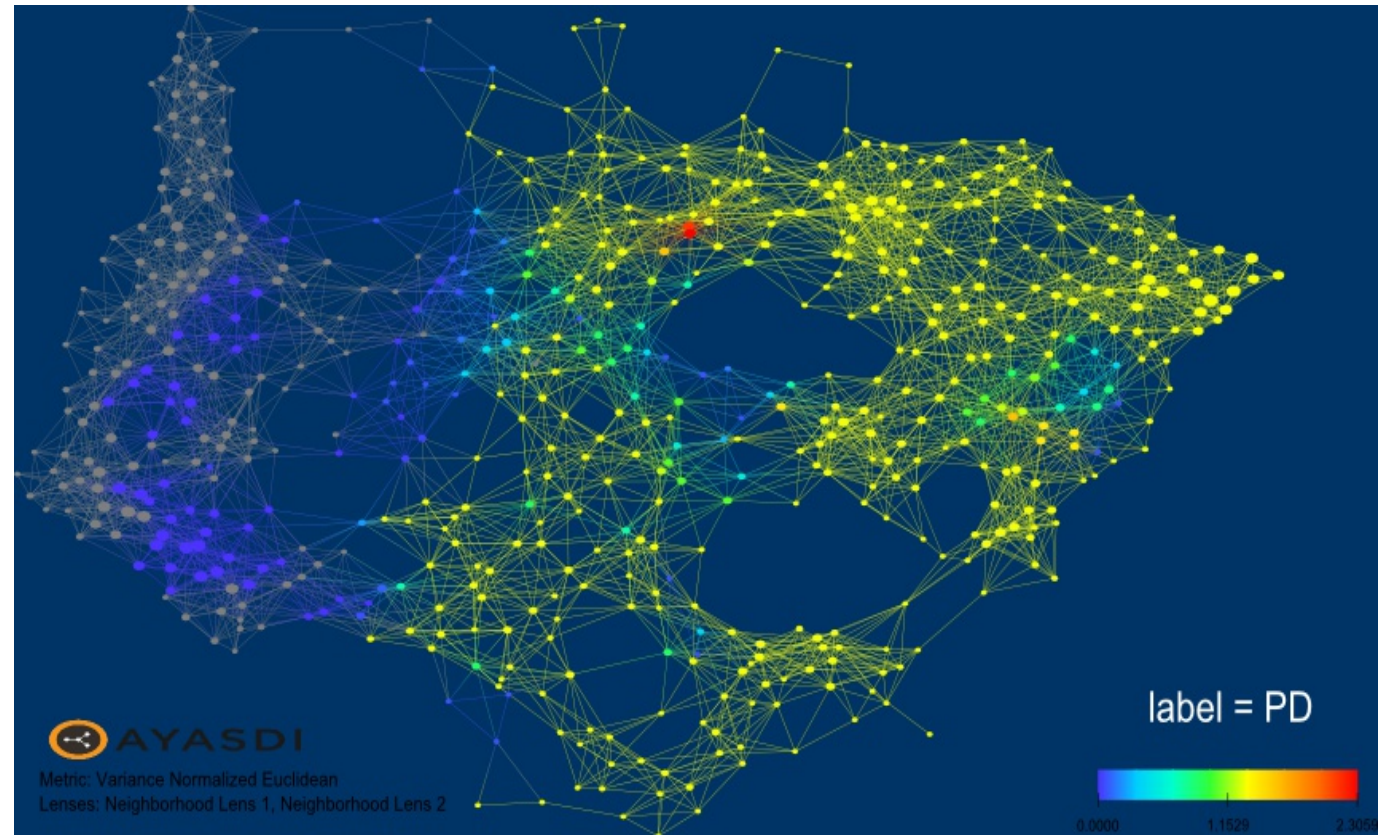- Not separated from healthy population!

Ayasdi — Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

Topological Data Analysis – PD Patients

Analysis:

- DaTSCAN data (4 feature set)

- PD subjects

- Clearly form separate cluster

Ayasdi — Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

Identifying PD Using Clinical Assessments

Are there other markers of Parkinson's Disease?

- Earlier studies indicate that PD patients deviate from healthy controls in many clinical parameters

- Using test of significance, we found three areas that were good predictors for PD and SWEDD:

  - Olfactory (smelling) ability (UPSIT)

  - REM sleep behavior (acting out dreams)

  - Self-evaluation of motor difficulties (UPDRS Part II)

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

- 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.

The Olfactory – REM Sleep – Motor Triad

Features used for 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

REM Sleep – Trends Among Cohorts

Analysis of REM sleep behavior:

	REM Aggressive Dreams [BL]	REM Complex Movements [BL]	REM Disturbed Sleep [BL]	REM Hurt Bed Partner [BL]	REM Move Arms/Legs [BL]	REM Movements Awake Me [BL]	REM Nocturnal Behavior [BL]	REM Remember Dreams [BL]	REM Speaking in Sleep [BL]	REM Sudden Limb Movements [BL]	REM Things Fell Down [BL]	REM Vivid Dreams [BL]
global mean	0.243451	0.077042	0.288136	0.107858	0.348228	0.211094	0.197227	0.368259	0.300462	0.280431	0.050847	0.619414
std dev	0.429496	0.266863	0.453244	0.310441	0.476776	0.408400	0.398212	0.482704	0.458813	0.449557	0.219856	0.485905
HC mean	0.217391	0.038043	0.211957	0.043478	0.244565	0.125000	0.201087	0.418478	0.255435	0.163043	0.038043	0.679348
PD mean	0.262376	0.089109	0.292079	0.133663	0.371287	0.235149	0.185644	0.353960	0.324257	0.314356	0.047030	0.589109
SWEDD mean	0.196721	0.114754	0.491803	0.131148	0.508197	0.311475	0.262295	0.311475	0.278689	0.409836	0.114754	0.639344

- Involuntary motion-related features are enhanced: SWEDD > PD > HC

- No clear picture regarding dream-related features

Motor Problems – Trends Among Cohorts
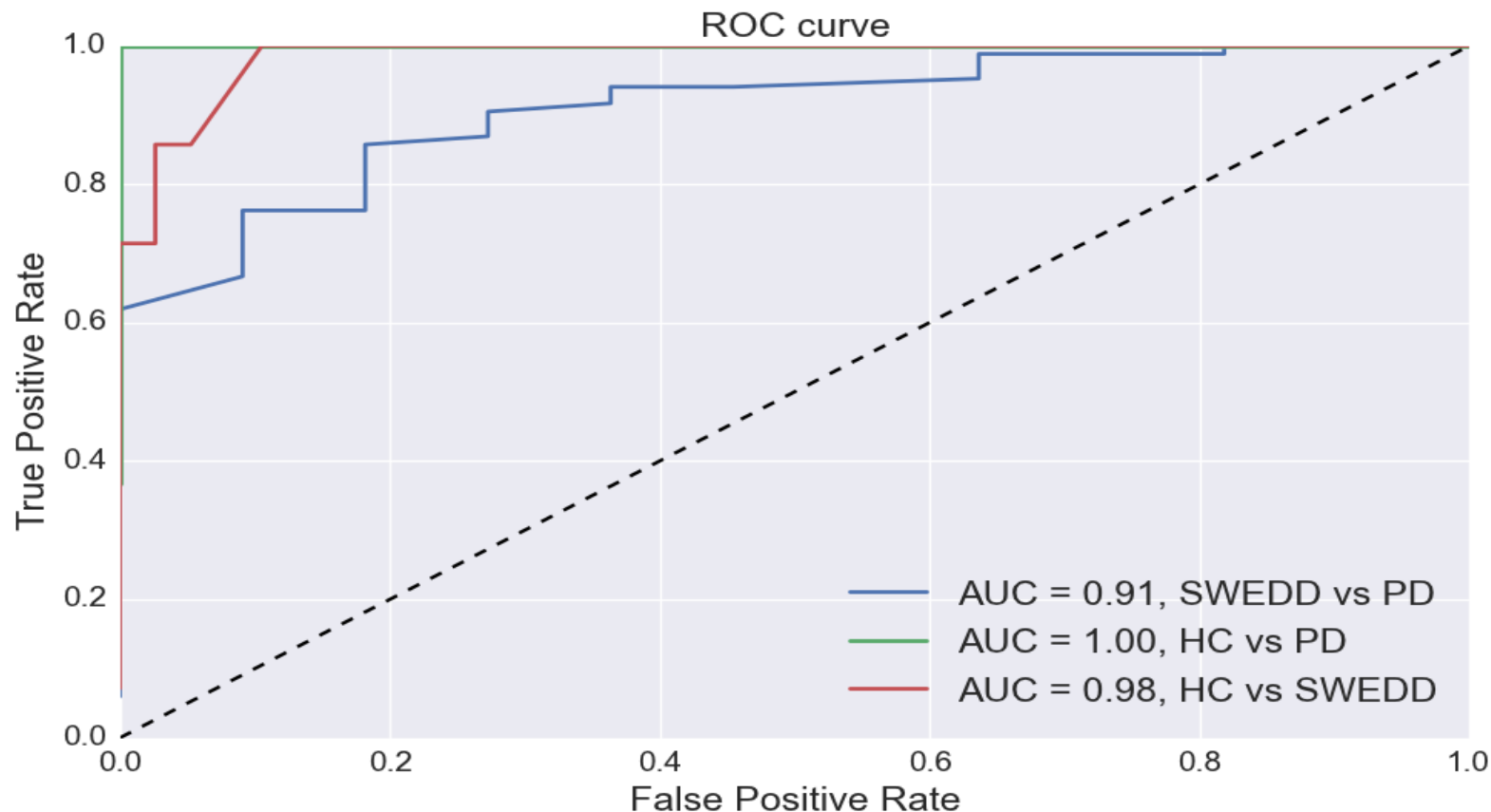
Analysis of self-reported motor problems:

	UPDRS-2 Dressing [BL]	UPDRS-2 Eating [BL]	UPDRS-2 Freezing [BL]	UPDRS-2 Getting up [BL]	UPDRS-2 Handwriting [BL]	UPDRS-2 Hobbies [BL]	UPDRS-2 Hygiene [BL]	UPDRS-2 Saliva [BL]	UPDRS-2 Speech [BL]	UPDRS-2 Swallowing [BL]	UPDRS-2 Tremor [BL]	UPDRS-2 Turning in Bed [BL]	UPDRS-2 Walking & Balance [BL]
global mean	0.315871	0.218798	0.044684	0.329738	0.637904	0.345146	0.163328	0.483821	0.320493	0.120185	0.787365	0.194145	0.291217
std dev	0.538984	0.459688	0.234732	0.568510	0.871273	0.641525	0.378201	0.917996	0.635058	0.397961	0.727180	0.411144	0.526988
HC mean	0.032609	0.005435	0.000000	0.065217	0.054348	0.038043	0.000000	0.097826	0.016304	0.021739	0.032609	0.027174	0.054348
PD mean	0.452970	0.297030	0.054455	0.438119	0.886139	0.448020	0.247525	0.655941	0.420792	0.146040	1.076733	0.252475	0.386139
SWEDD mean	0.262295	0.344262	0.114754	0.409836	0.754098	0.590164	0.098361	0.508197	0.573770	0.245902	1.147541	0.311475	0.377049

- Speech, swallowing, eating problems more pronounced in SWEDD
- Problems with dressing, handwriting, hygiene, saliva more prominent in PD

Analysis of olfactory performance:

- More impaired in PD (22.3/40) than SWEDD (31.4/40). Best in HC (34.2/40).

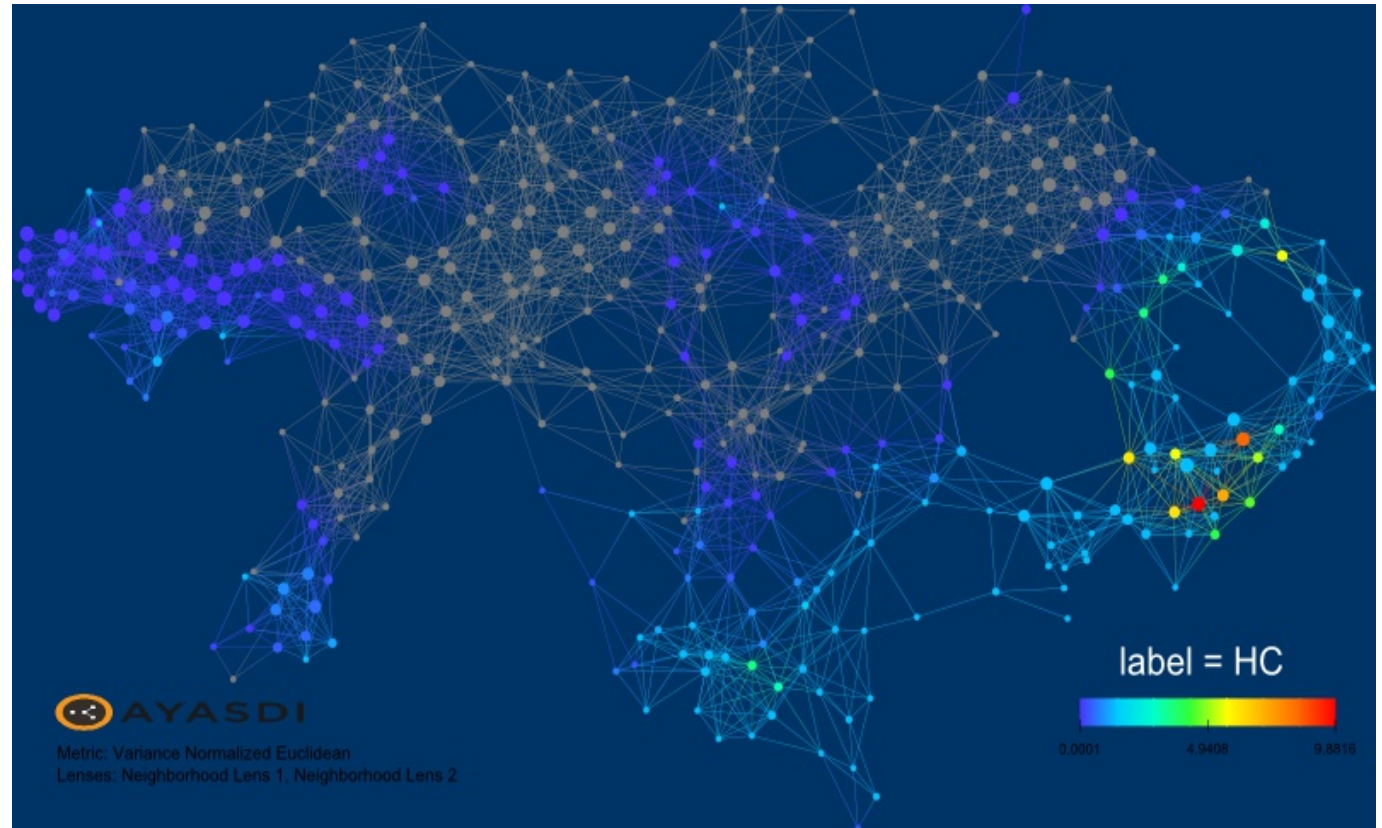# Classification Using Smell, REM, and Motor

ROC curve, True Positive Rate vs False Positive Rate:
- AUC = 0.91, SWEDD vs PD
- AUC = 1.00, HC vs PD
- AUC = 0.98, HC vs SWEDD

Smell, REM, and motor assessments together outperform DaTSCAN!

Topological Data Analysis – Healthy Cohort

Analysis:

- Data: Olfactory, REM Sleep, Motor (26 feature set)
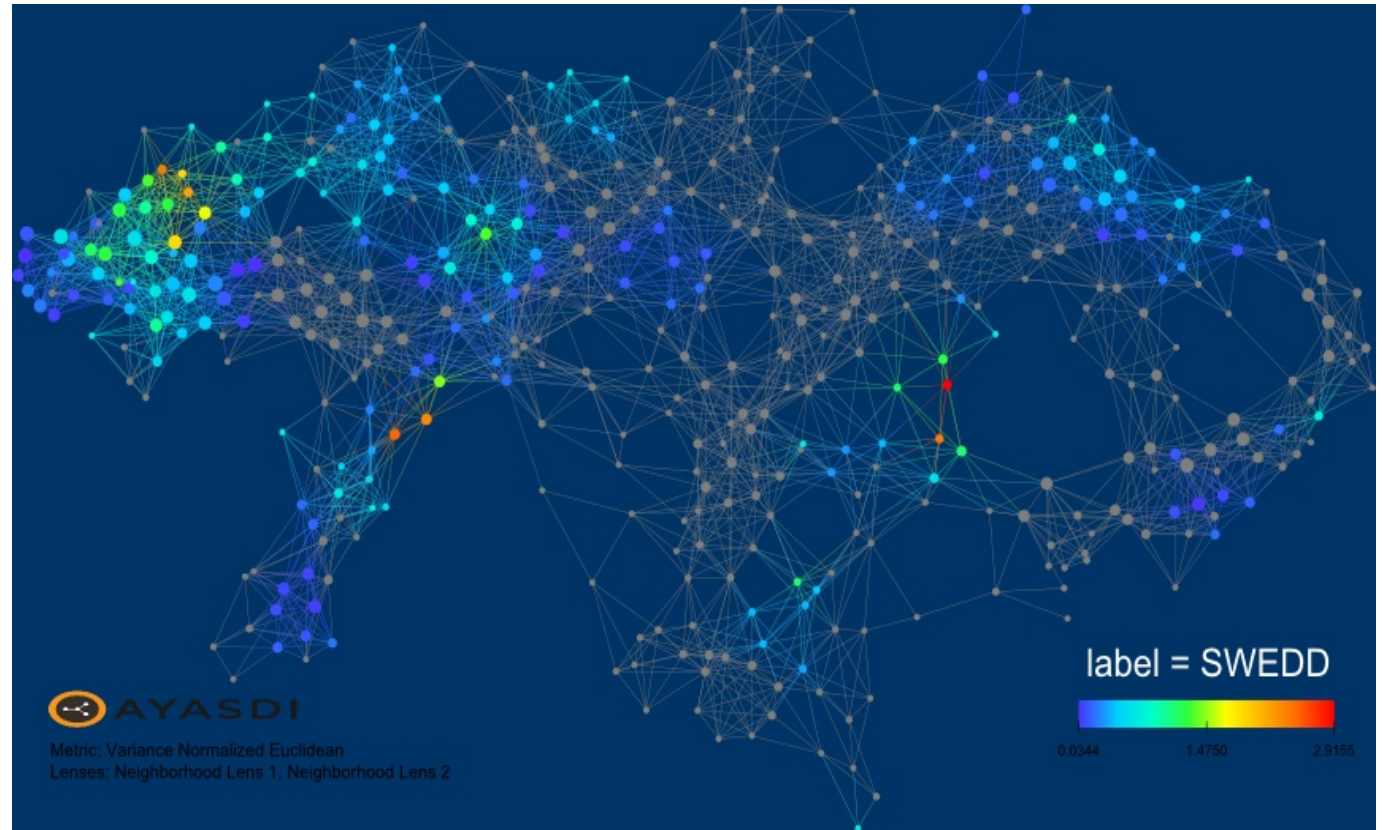
- Healthy subjects

label = HC

Metric: Variance Normalized Euclidean
Lenses: Neighborhood Lens 1, Neighborhood Lens 2

0.0001   4.9408   9.8816

Ayasdi — Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

Topological Data Analysis – SWEDD

Analysis:

- Data: Olfactory, REM Sleep, Motor (26 feature set)

- SWEDD subjects

- Virtually no overlap with healthy cohort!

- Several clusters?

Ayasdi — Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

Topological Data Analysis – PD Patients

Analysis:

- Data: Olfactory, REM Sleep, Motor (26 feature set)

- PD patients

- Network "split" by SWEDD cluster?
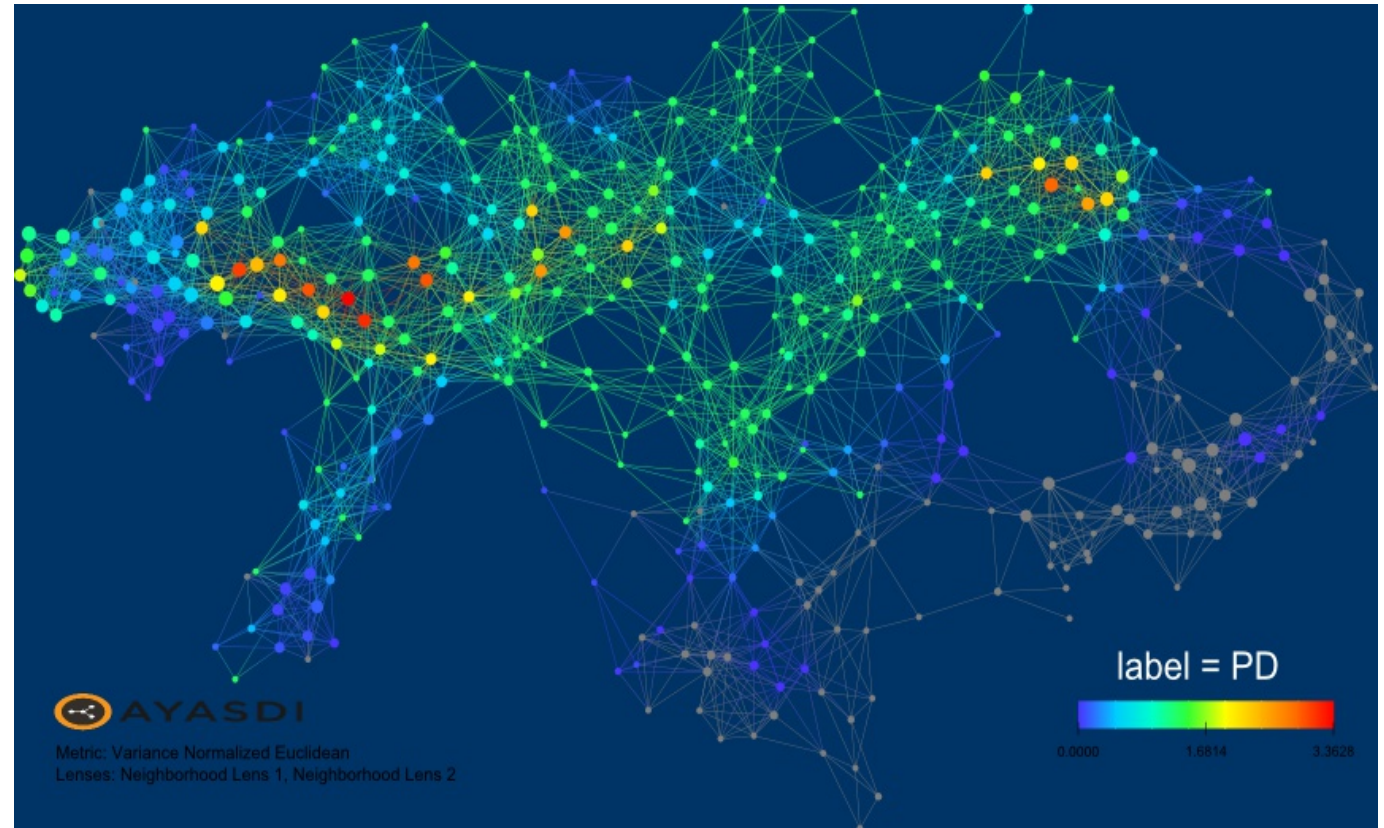
Ayasdi — Algorithm: Variance Normalized Euclidean Metric using Neighborhood Lenses

Visualization by Feature Reduction

"Mechanism" behind Smell – REM – Motor identification?

- Subject data "lives" in 26-dimensional feature space; difficult to visualize or gain insight

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

Problem: Projection creates loss of information

Smell – REM – Motor: Feature Reduction
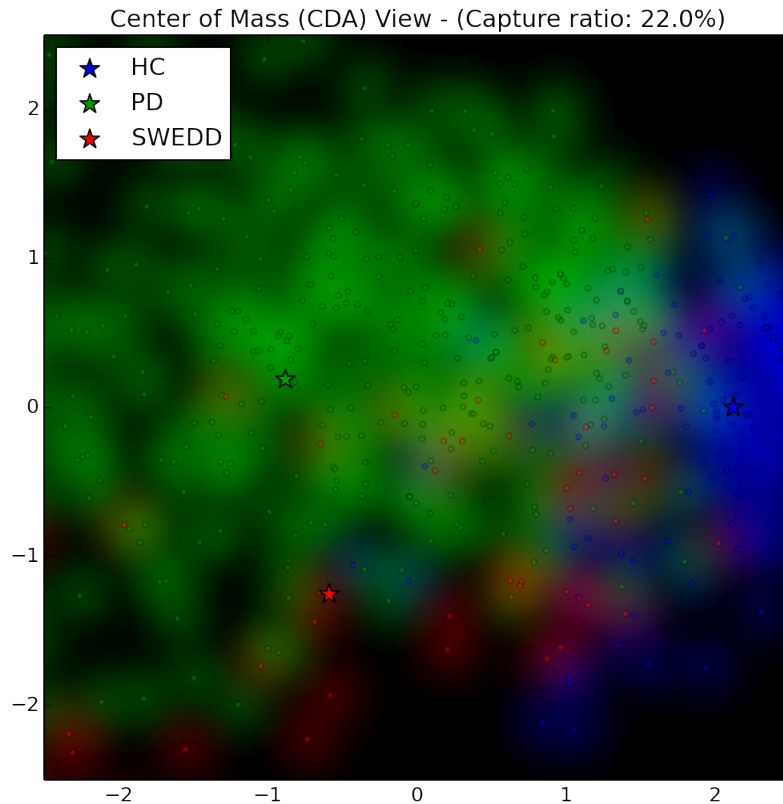
Canonical Discriminant Analysis
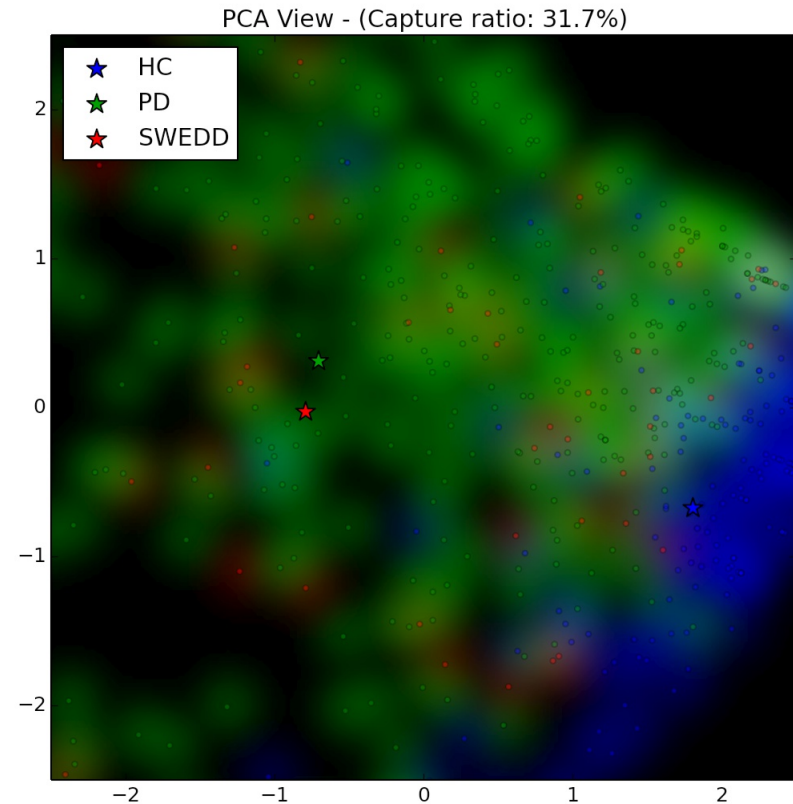
Principal Component Analysis

Low capture ratio points to importance of multidimensional patterns in feature space!

DaTSCAN: Feature Reduction

Center of Mass (CDA) View - (Capture ratio: 93.6%)

Canonical Discriminant Analysis

PCA View - (Capture ratio: 93.9%)
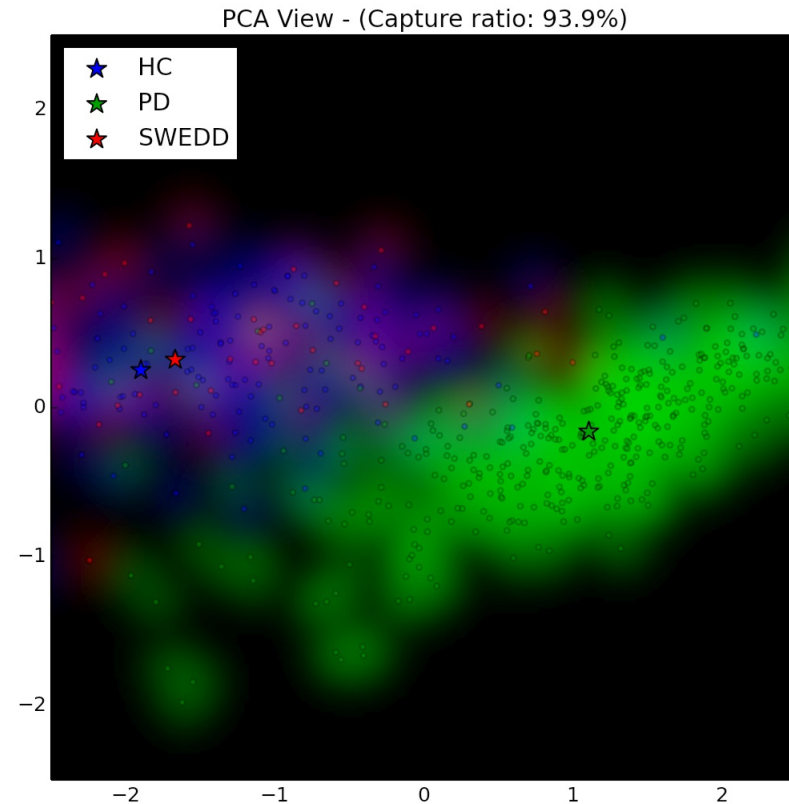
Principal Component Analysis

DaTSCAN classification is predominantly based on a single direction in feature space

Assessments That Lack Resolving Power

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

- Markers that appear altogether independent of 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.

Blood Lipids vs. PD Status

PCA View - (Capture ratio: 78.3%)

Principal Component Analysis

ROC Curve (Random Forest)

AUC HC vs. PD: 0.55
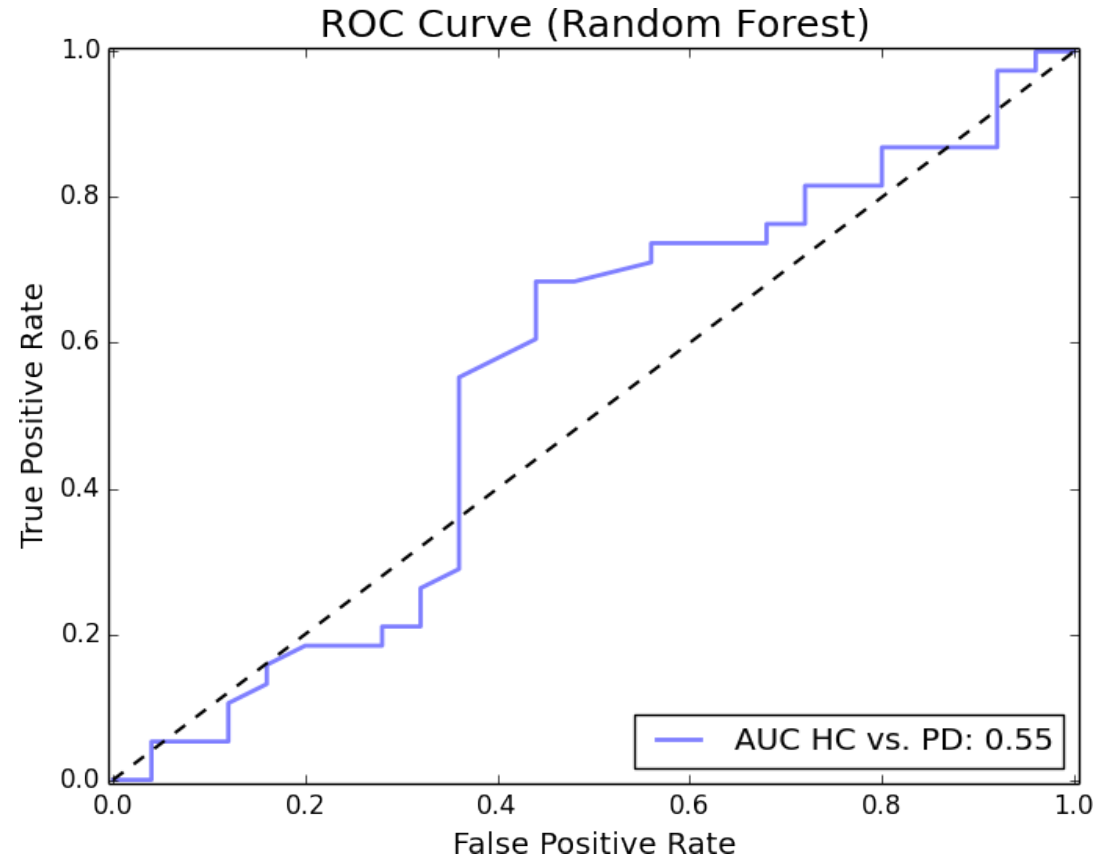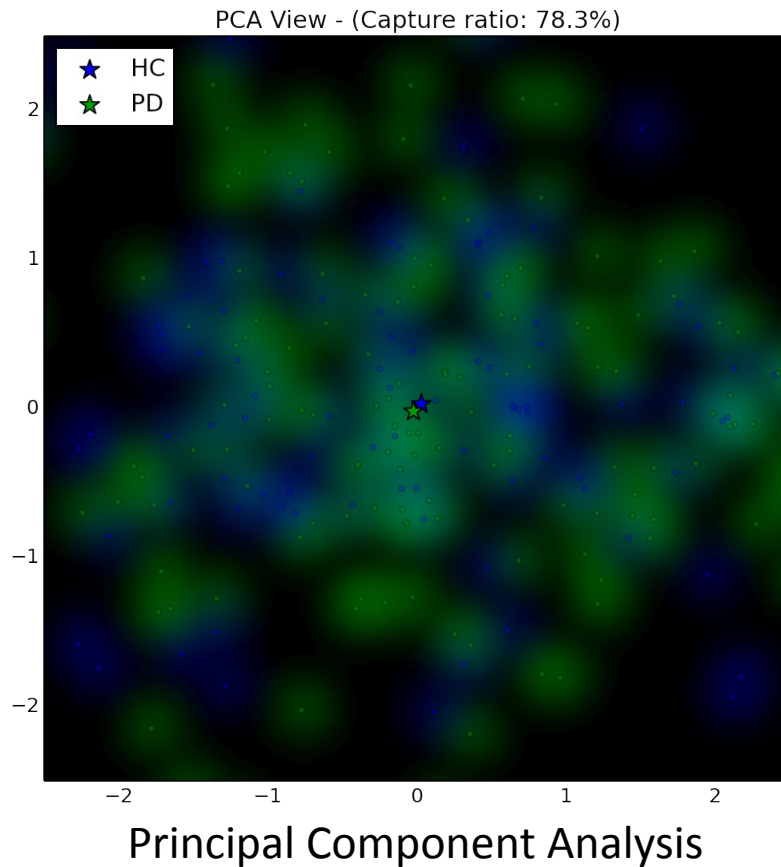
No correlation between blood lipids and PD cohort found. (*Note*: SWEDD data missing.)
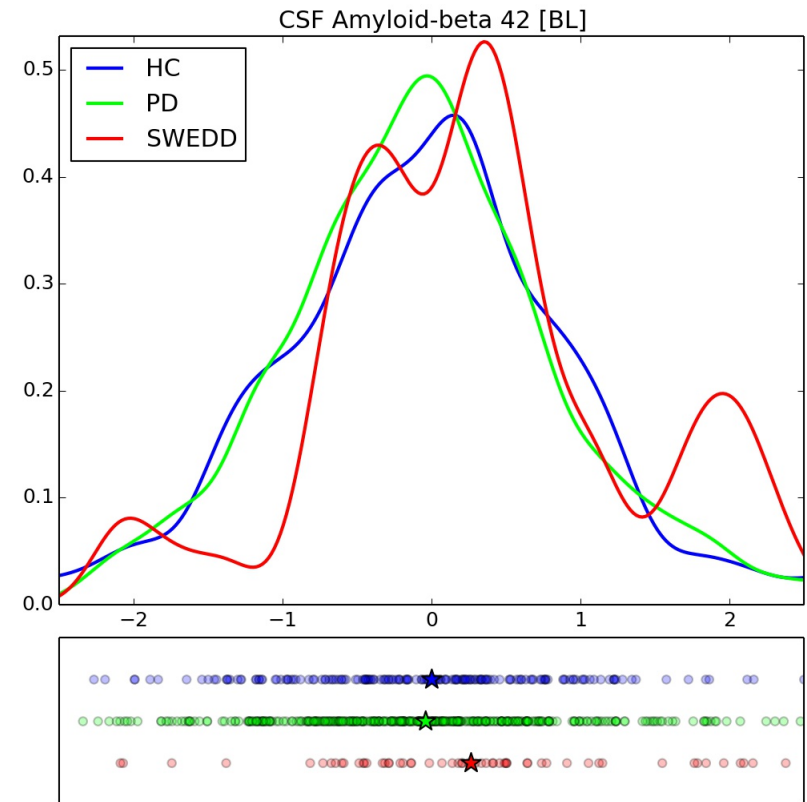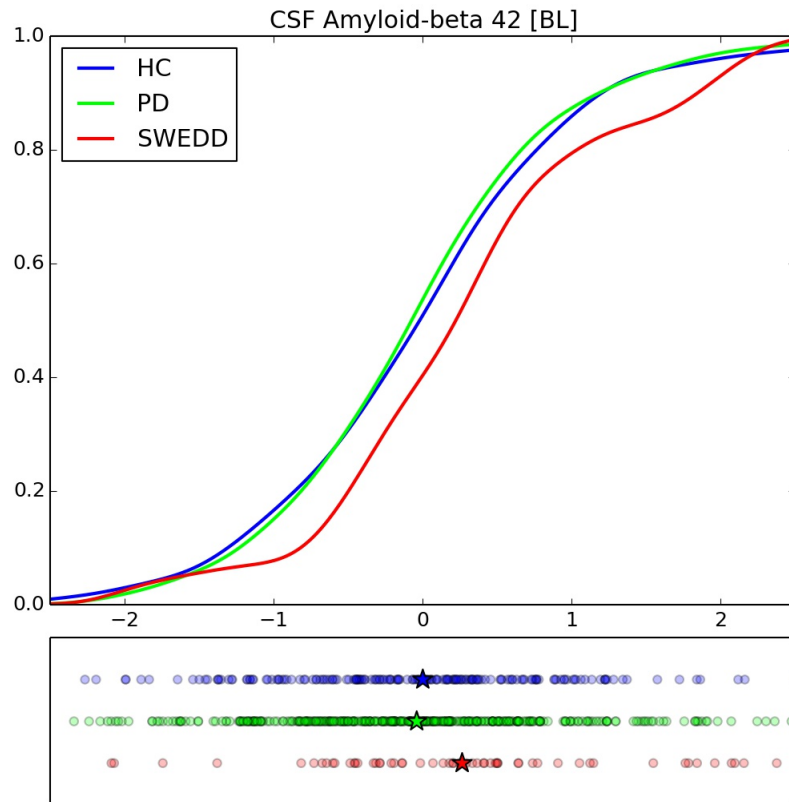
CSF Profiles: α-Synuclein

Significant decrease of α-syn in PD cohort (compared to HC, SWEDD) on average:
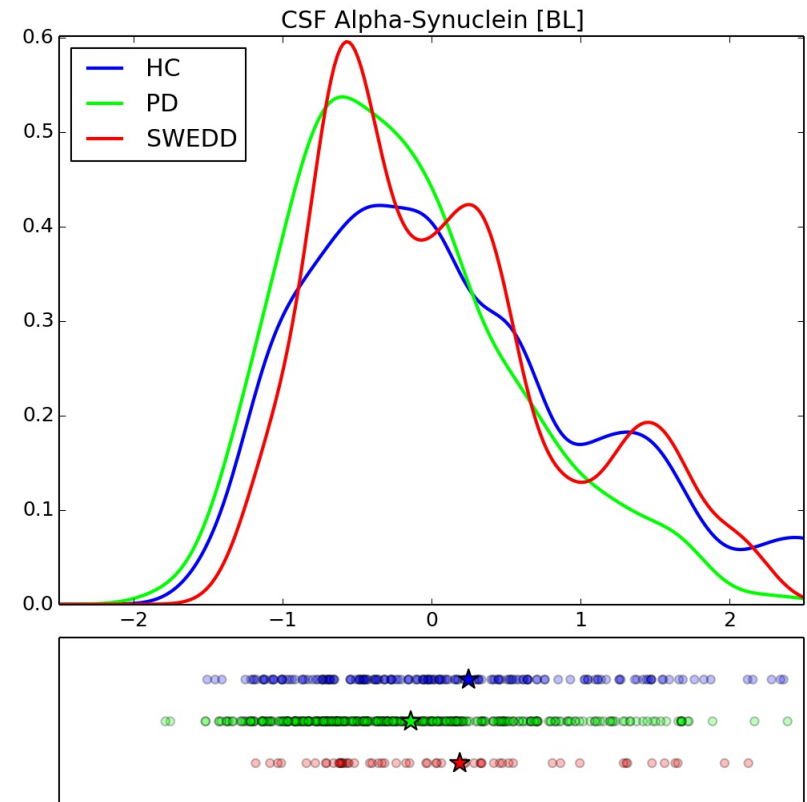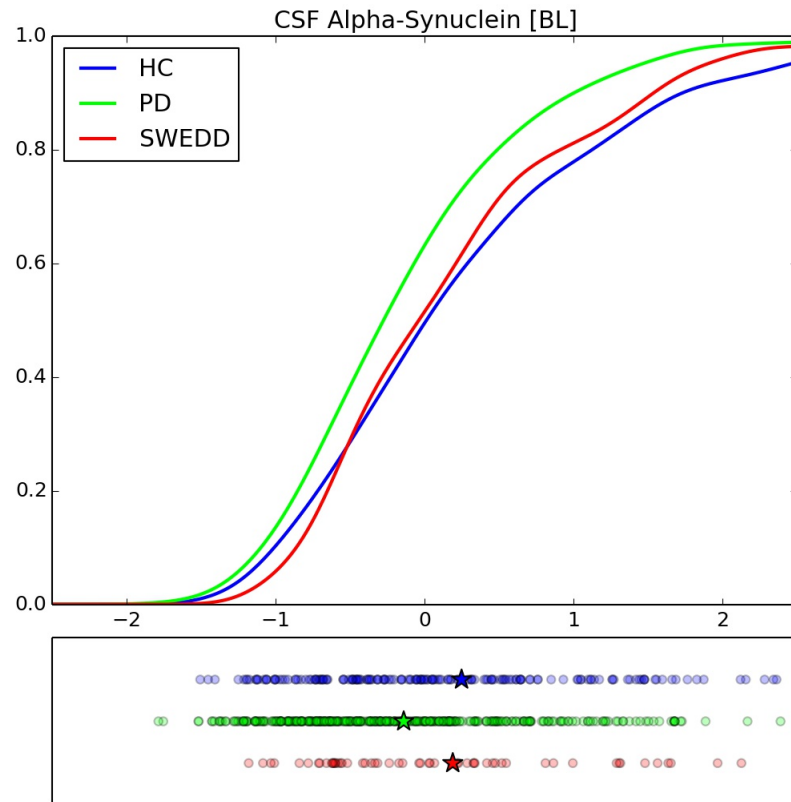
Cumulative probability (left) and probability density (right) distributions with Gauss filter.
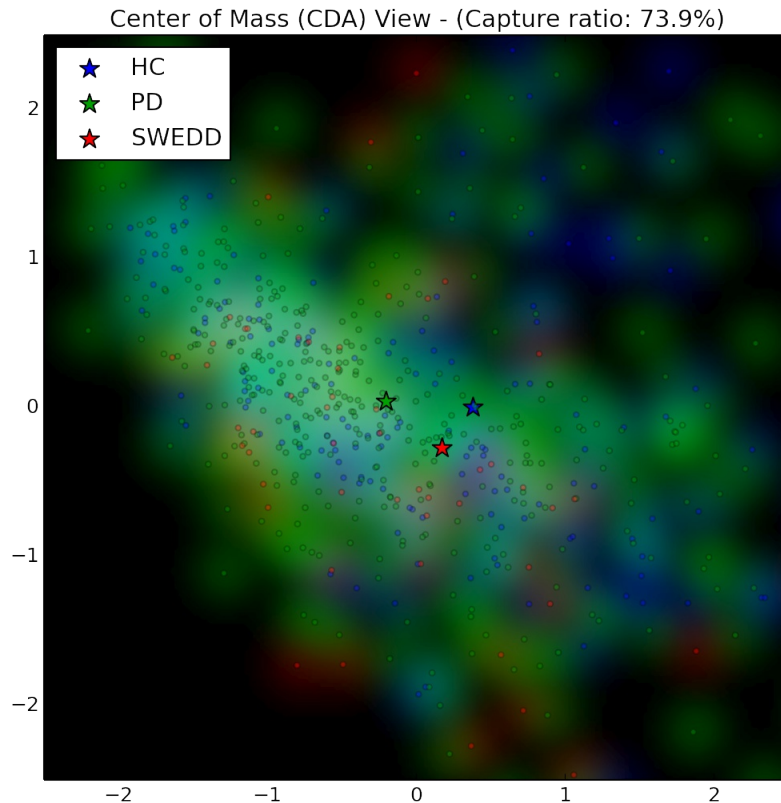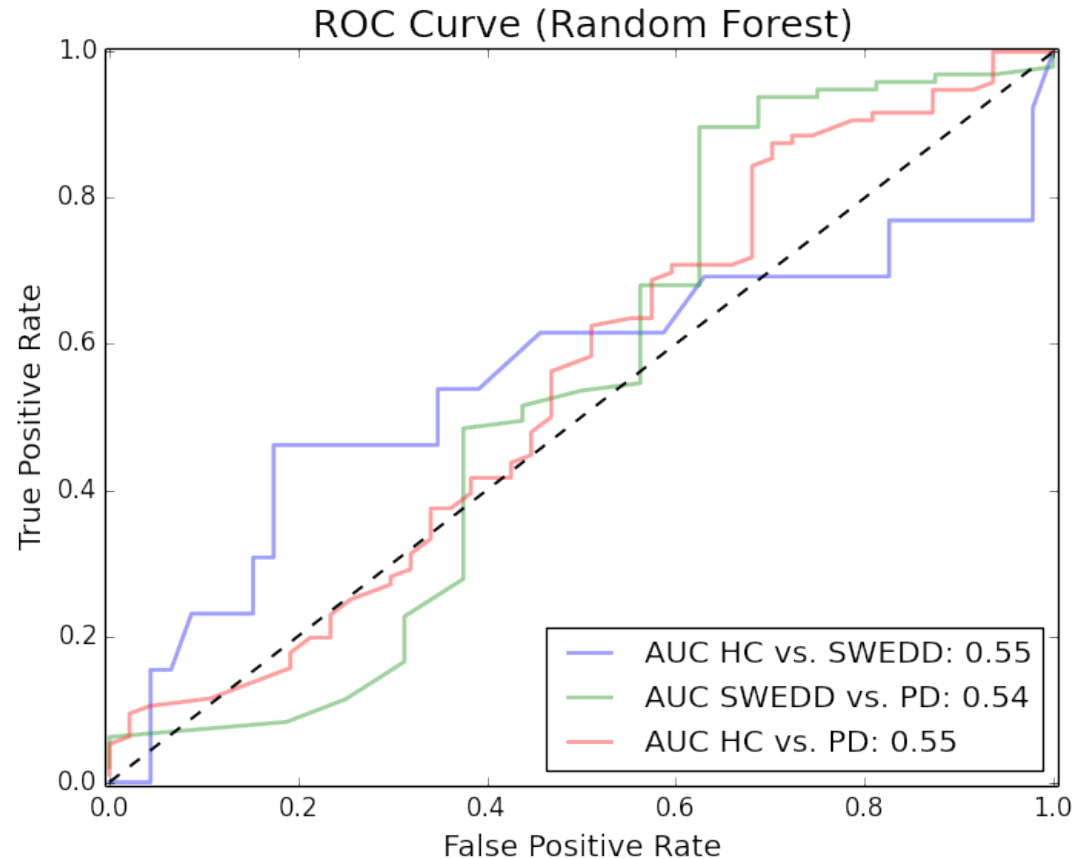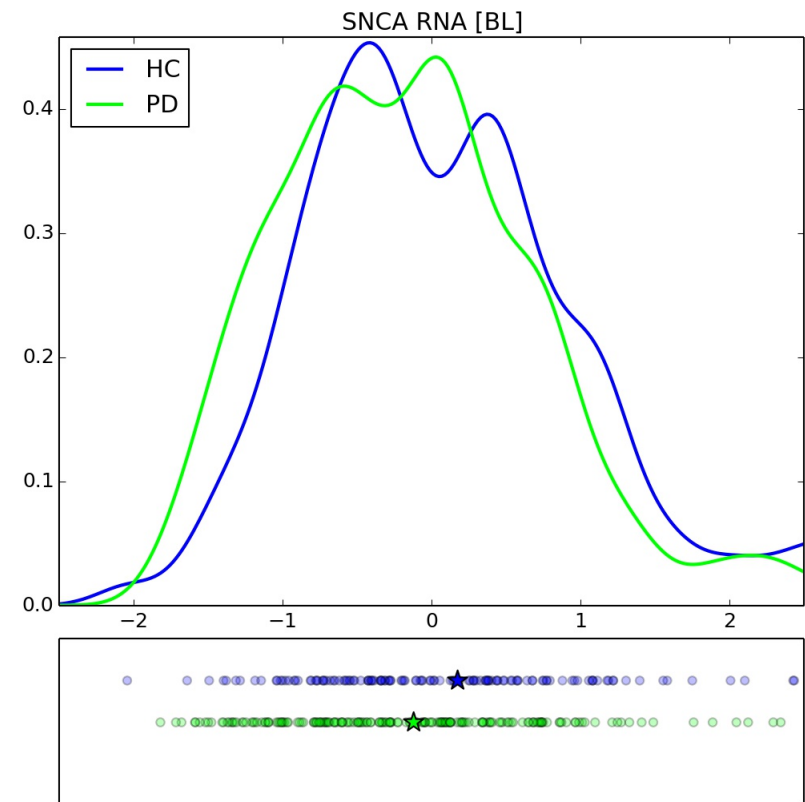Bottom: Raw data and cohort averages

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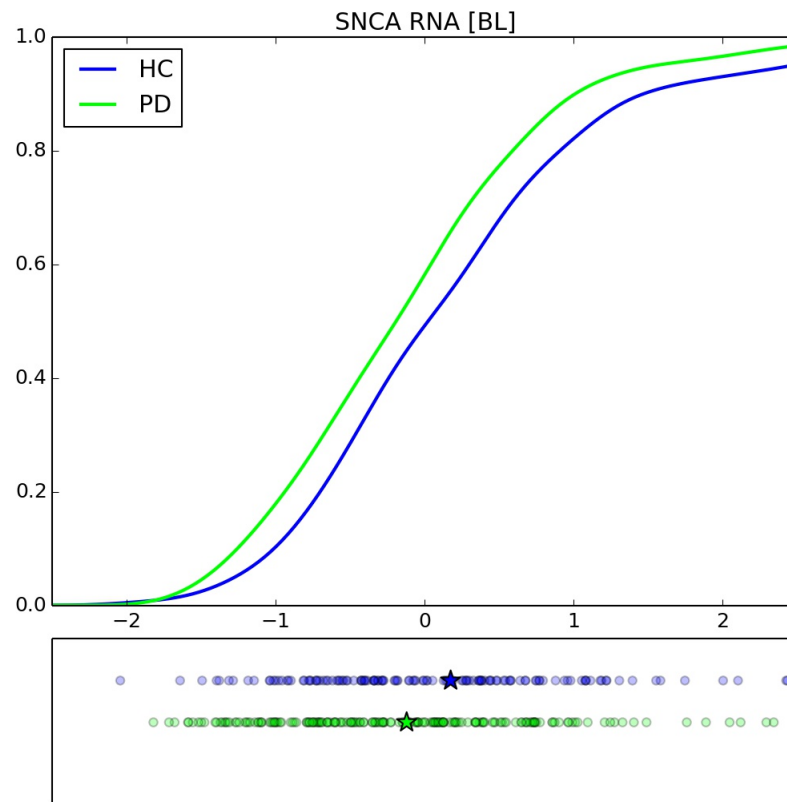
Significant increase of Amyloid-β_{42} in SWEDD cohort (compared to HC, PD) **on average:**

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

Analysis: CSF Proteins

Center of Mass (CDA) View - (Capture ratio: 73.9%)

Canonical Discriminant Analysis

ROC Curve (Random Forest)

AUC HC vs. SWEDD: 0.55
AUC SWEDD vs. PD: 0.54
AUC HC vs. PD: 0.55

Cohorts strongly overlap, weak predictive power of CSF protein data

Transcription: RNA Counts SNCA (all types)

Transcription of SNCA is overall suppressed in PD cohort (no SWEDD data available):

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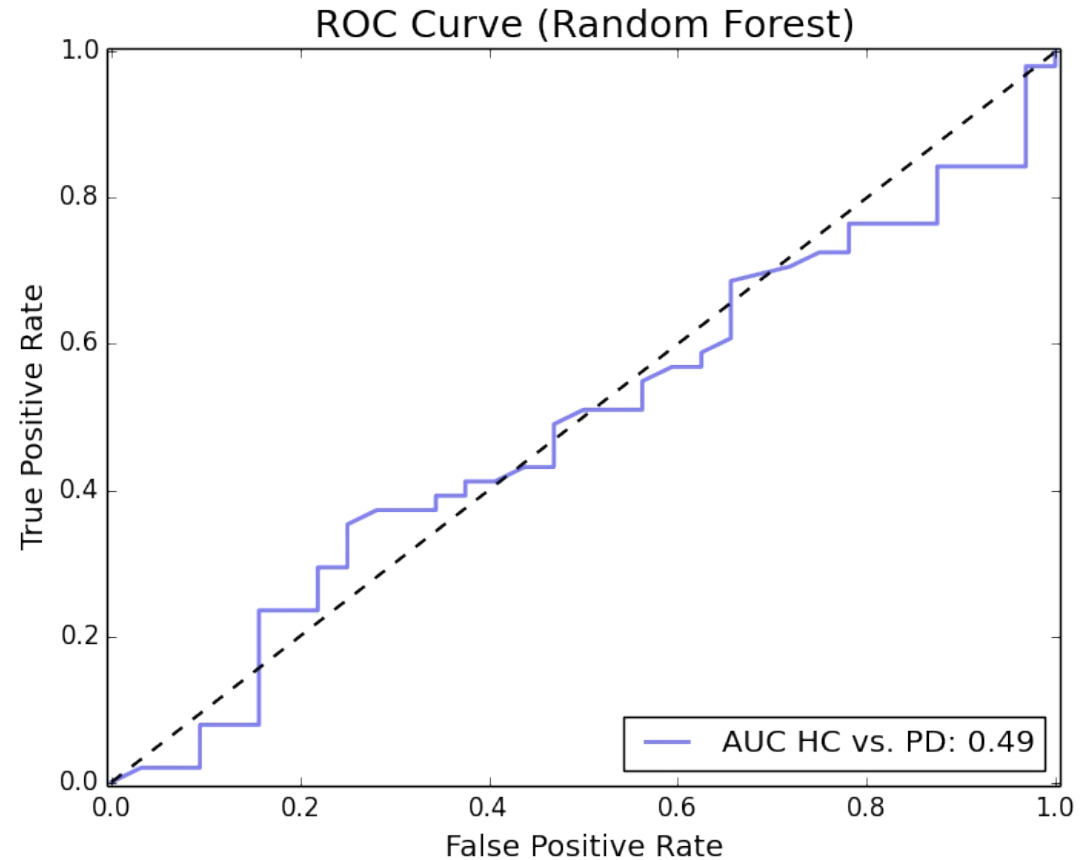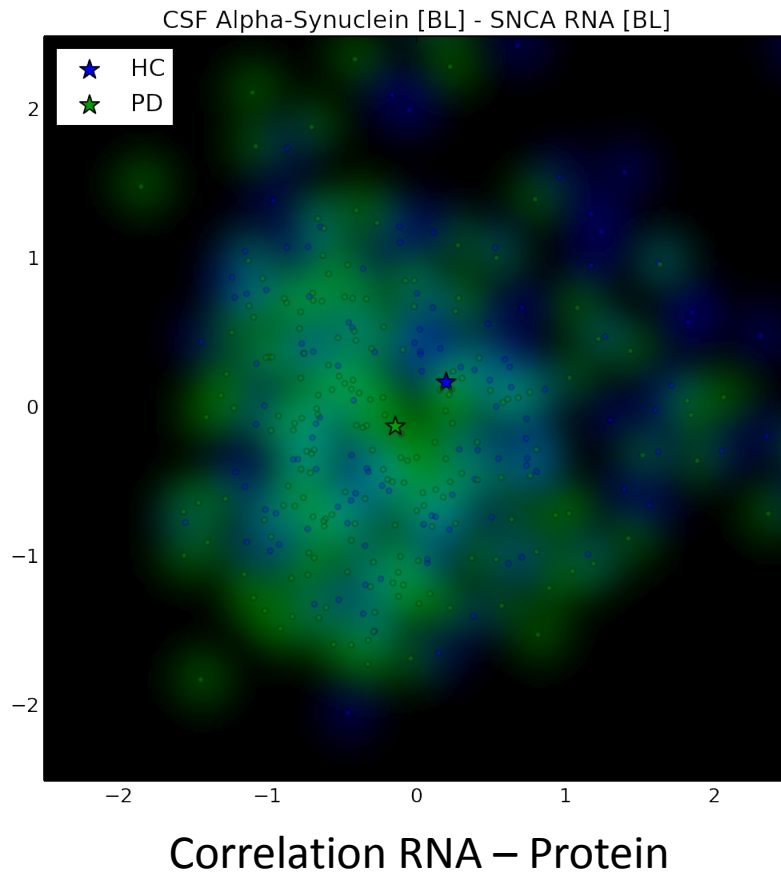Bottom: Raw data and cohort averages

Transcription: RNA Counts FBXO7 (all types)

In comparison, there is little change in transcription of FBXO7 and other PD-related genes:
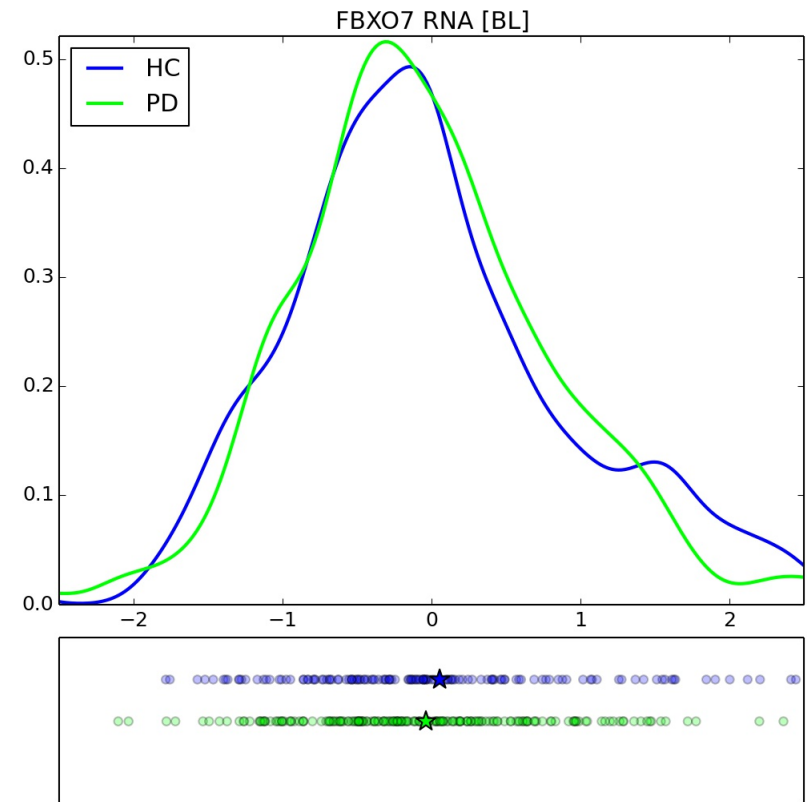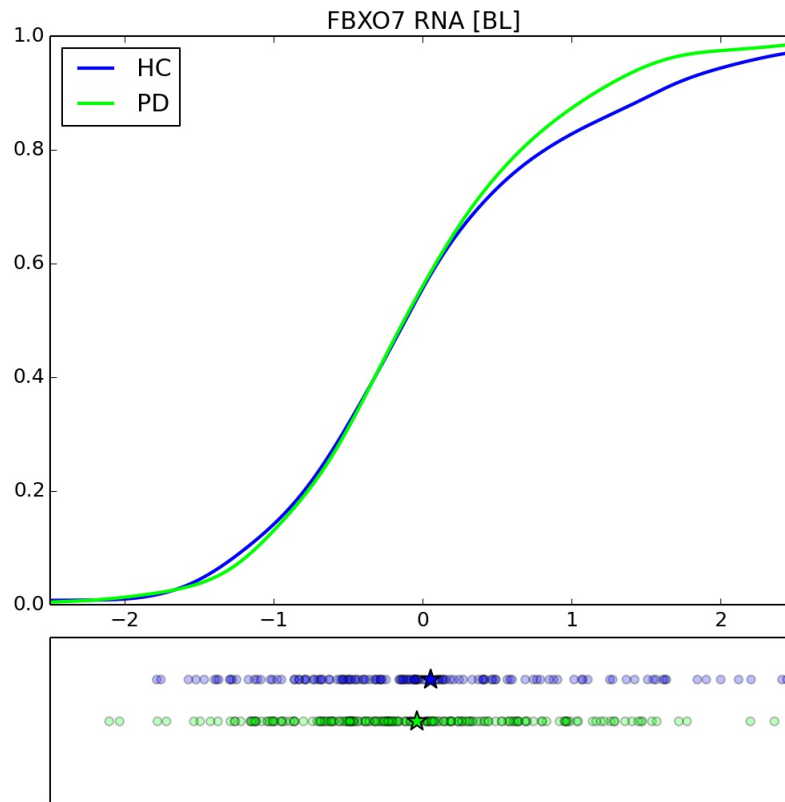
Cumulative probability (left) and probability density (right) distributions with Gauss filter.
Bottom: Raw data and cohort averages

Are SNCA RNA, α-Syn Protein Correlated?

CSF Alpha-Synuclein [BL] - SNCA RNA [BL]

Correlation RNA – Protein

ROC Curve (Random Forest)

AUC HC vs. PD: 0.49

Transcription of *SNCA* RNA and CSF α-Syn level are both reduced in PD patients on average.
Although the *SNCA* gene translates into α-Syn protein, their levels display little correlation.

Leveraging the Web for Parkinson's Research

Smell-REM-Motor classifier contains self-reported data:

- Use social media to raise PD awareness

- Gather large-scale population data from online community

- Outreach via websites, apps, gamification?

Smell-REM-Motor testing requires no specialist knowledge:

- Define "best practices" in PD diagnosis and differential testing

- Provide tools for family practitioners to identify likely PD sufferers

Online Self-Assessment of Motor Skills
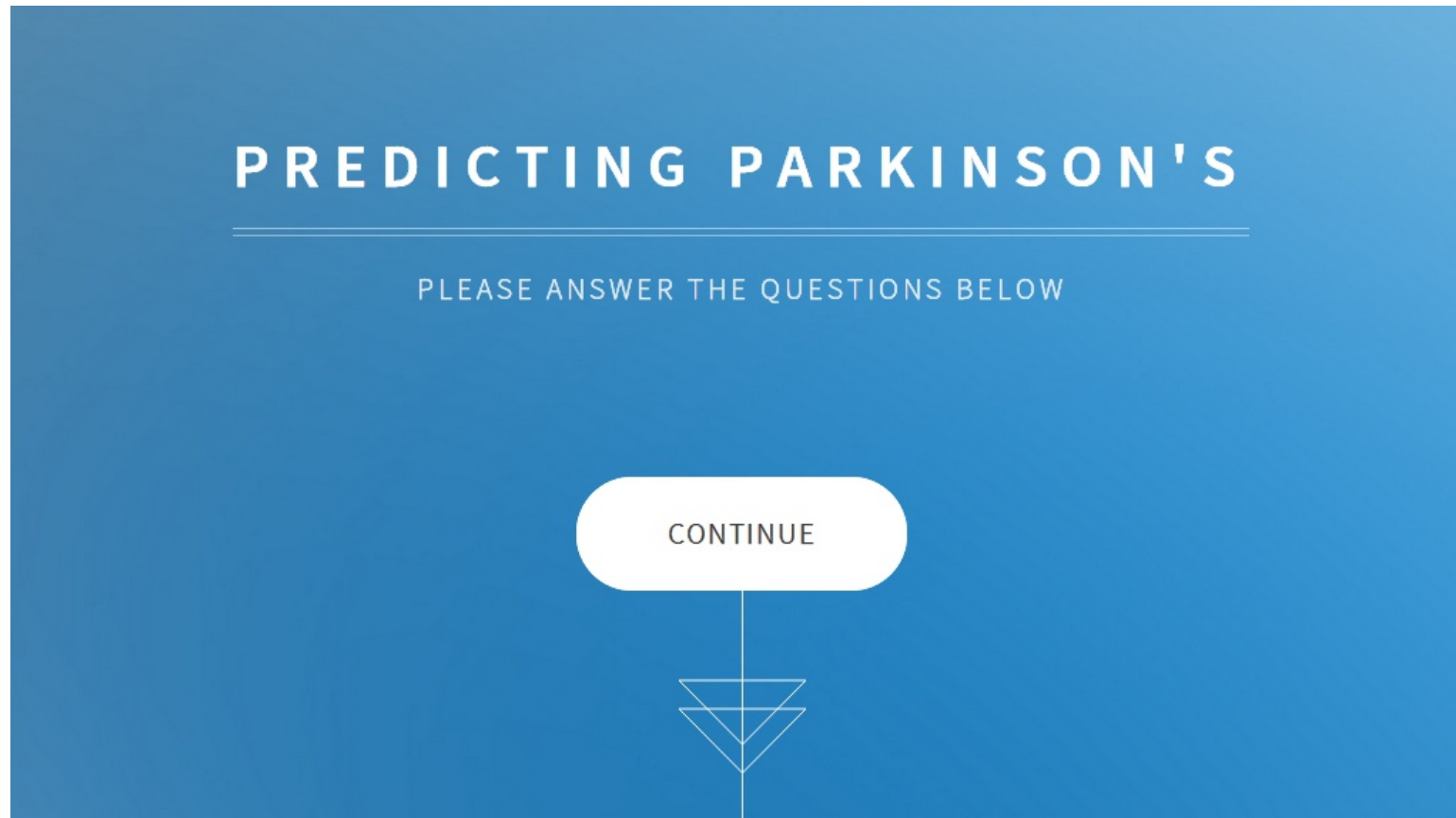
"Predicting Parkinson's" — Screenshot of landing page. Online at: www.mgomezm.com

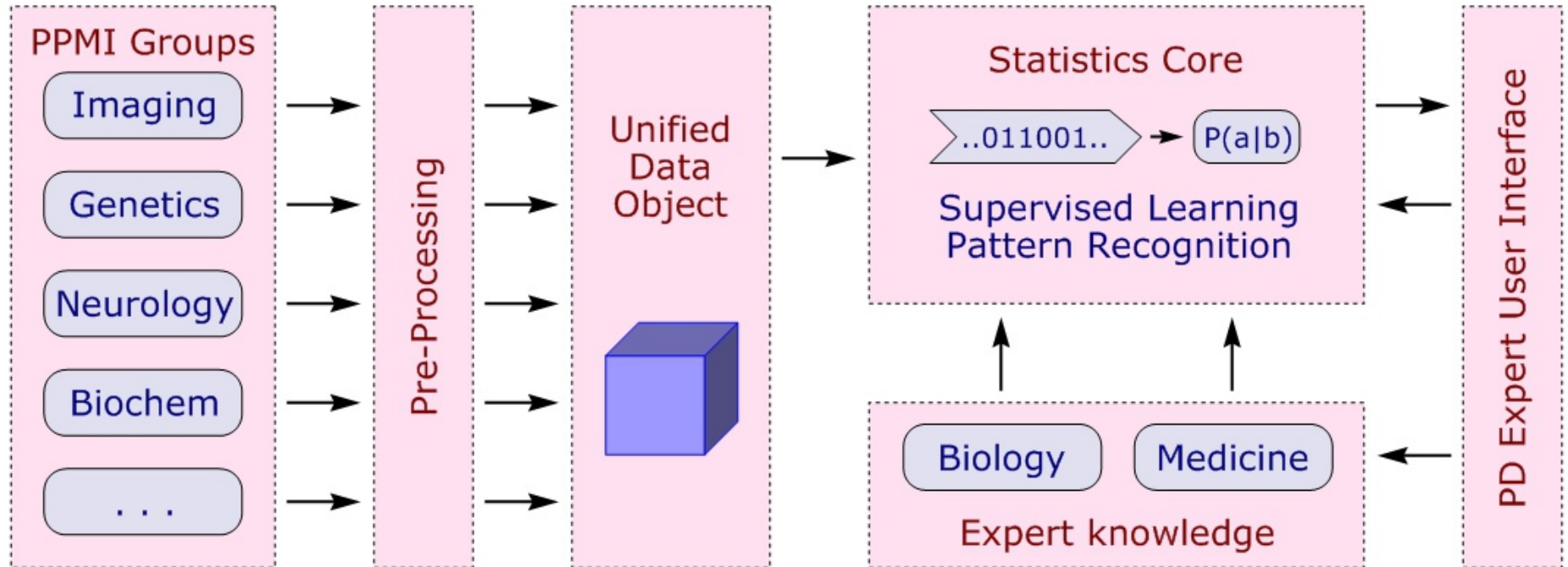# A Machine Learning Application for PPMI

Background:

- Machine learning is a promising tool for analyzing medical data, but acceptance is low:

  - Few biologists and physicians are familiar with advanced statistics

  - Few data scientists have a background in the life sciences

- Idea: Provide a utility that makes data mining techniques accessible to PPMI research community

A Machine Learning Application for PPMI

Specifications:

- Utility should be easy to operate
  - Intuitive approach: "Select and click"
  - Reduce numerical information, rely on graphical display instead

- PPMI is an international study
  - World-wide access: Employ website as user interface

- Medical data is privileged, sensitive information:
  - Keep PPMI data, machine learning core on central server
  - Transmit only aggregate data to address privacy concerns

Machine Learning Utility Framework

PPMI Groups
- Imaging
- Genetics
- Neurology
- Biochem
- ...

Pre-Processing

Unified Data Object

Statistics Core

..011001.. → P(a|b)

Supervised Learning Pattern Recognition

Biology   Medicine

Expert knowledge

PD Expert User Interface

PPMI Data

ETL Stage:
Load & Assemble Data

Statistics & ML Core:
Data Selection & Graphics
Communication with UI

WWW

PD-LEARN User Interface

PD-LEARN — Statistical Analysis with PPMI Data

THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH

BAYES IMPACT

Step #1: Select Data Sets

These data sets are available for analysis. Mark the sets you'd like to include, then click the **Continue** button:

	SC	BL	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12
UPSIT Olfactory Score	0	655	0	0	0	0	0	0	0	0	0	0	0	0
DaTSCAN Left Putamen	649	0	0	0	0	323	0	130	0	0	0	2	0	0
DaTSCAN Right Putamen	649	0	0	0	0	323	0	130	0	0	0	2	0	0
DaTSCAN Left Caudate	649	0	0	0	0	323	0	130	0	0	0	2	0	0
DaTSCAN Right Caudate	649	0	0	0	0	323	0	130	0	0	0	2	0	0
CSF Amyloid-beta 42	0	635	0	263	0	300	0	0	0	0	0	0	0	0
CSF Alpha-Synuclein	0	635	0	262	0	300	0	0	0	0	0	0	0	0
CSF Total Tau-Protein	0	629	0	262	0	299	0	0	0	0	0	0	0	0
CSF Phosphorylated Tau-Protein	0	633	0	263	0	299	0	0	0	0	0	0	0	0

[Tip]   [Continue]

Version 0.01 (mock-up) - Created August 11, 2014 - Author: Christian Bracher

Summary: Assessing Disease Markers

- PPMI is a large-scale, multi-feature study of Parkinson's Disease.

- We investigated variables for their value in diagnosing PD.

- Combining olfactory ability, REM sleep abnormalities, and perceived motor disability, we found a superior predictor for PD.

- Unlike DaTSCAN, our test reliably identifies SWEDD patients.

- We applied machine learning methods to PPMI data to test markers suggested in the PD literature. We found that the amount of disease-related proteins in the CSF, and translation of genes implied in Parkinson's disease are not good predictors for PD.

Summary: Using the Internet for PD

- The olfactory – REM sleep – motor method is simple, noninvasive, and affordable, and suitable for non-clinical settings.

- We built a WebApp, based on a self-assessment of motor difficulties, to identify people at risk of Parkinson's.

- Advanced statistical methods and machine learning are underutilized in medical research.

- We are developing a web-based utility that extracts PPMI data, and gives non-experts access to ML tools through a graphical interface.

Further Investigations

- Using machine learning on longitudinal data in the PPMI study: How do PD patients' symptoms progress over time?

- PPMI has a plethora of genetic information that should be studied.

- We did not have full access to the rich PPMI brain imaging data. Further analysis may improve on the classification of PD.

- Use Topological Analysis and unsupervised learning to identify clusters in the data linked to subtypes of PD (like SWEDD).

- Use Natural Language Processing (NLP) on patients diaries/journals and doctors' notes to obtain additional features.

- Use social media (Twitter, ...) for PD awareness and early detection.

Contact Information

- Christian Bracher

  cbracher69@gmail.com

  http://www.christianbracher.net/

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

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

- Mijail Gomez

  mijail.gomez.munoz@gmail.com

  https://www.linkedin.com/in/mijailgomez

  https://github.com/mgomezm/Predicting-Parkinsons