# Subspace Clustering of CFS Data

NYU Bioinformatics

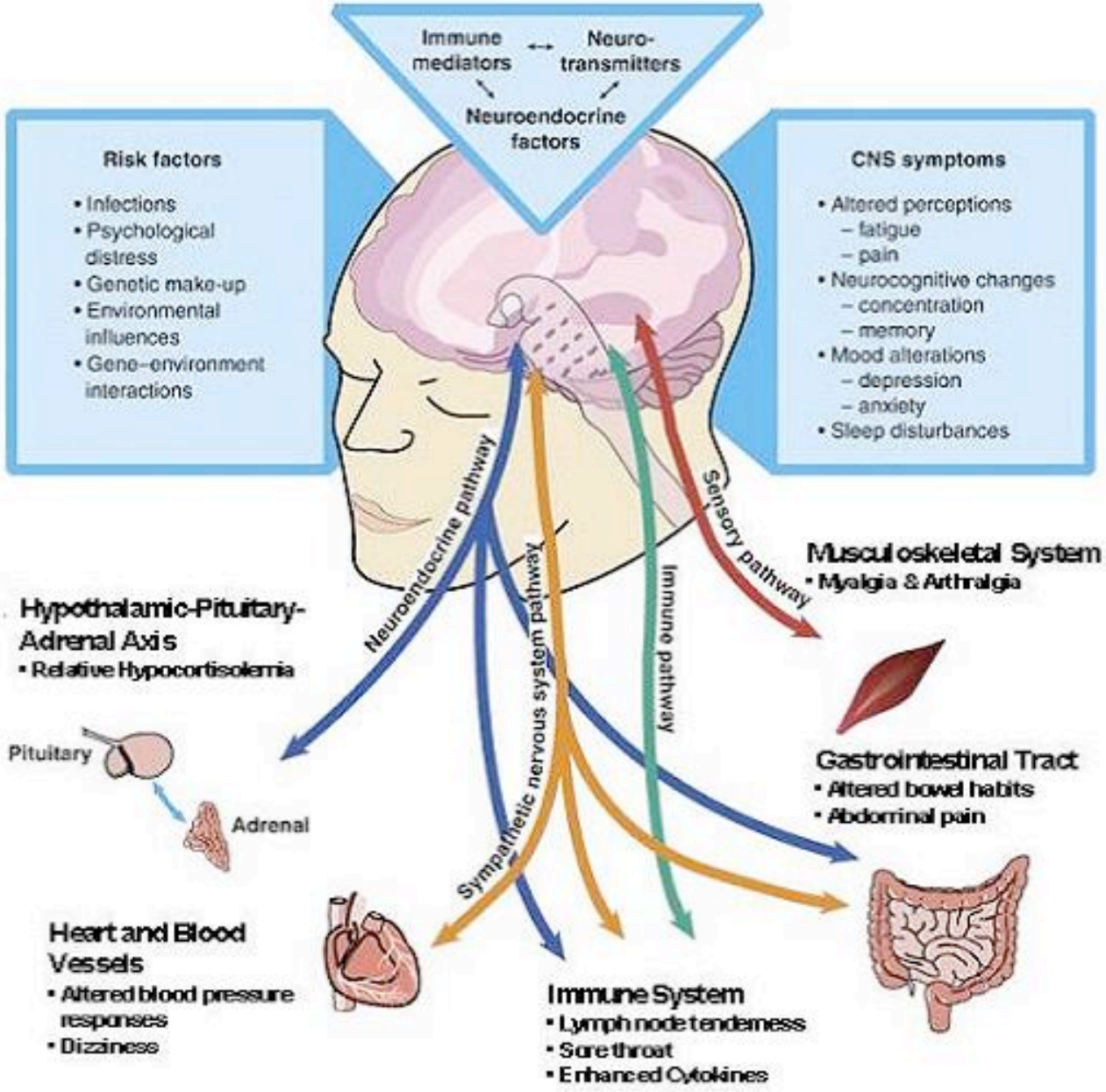
Oct 5, 2009

# The Epidemiology of CFS

- Clinic populations
  - Women
  - Primarily white
  - Middle upper socioeconomic status
  - Sudden onset (recovery more likely)
  - Illness duration 5 yrs
  - 22 physician visits/year

- General population
  - Women
  - All race/ethnicities
  - Low socioeconomic
  - Gradual onset (lower recovery rates)
  - Illness duration 5 yrs
  - 15% have been diagnosed/treated

Economic impact - $9 billion/year, $20,000 per household

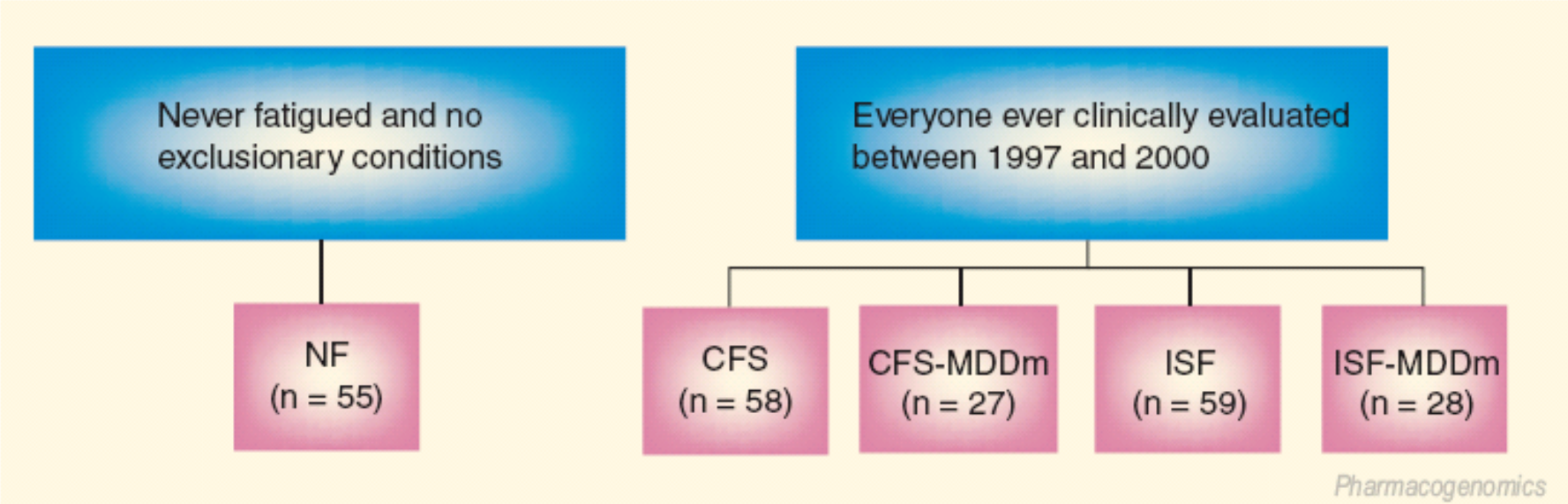

Immune mediators
Neuro-transmitters
Neuroendocrine factors
Risk factors
• Infections
• Psychological distress
• Genetic make-up
• Environmental influences
• Gene–environment interactions
CNS symptoms
• Altered perceptions
– fatigue
– pain
• Neurocognitive changes
– concentration
– memory
• Mood alterations
– depression
– anxiety
• Sleep disturbances
Neuroendocrine pathway
Sympathetic nervous system pathway
Immune pathway
Sensory pathway
Hypothalamic-Pituitary-Adrenal Axis
• Relative Hypocortisolemia
Pituitary
Adrenal
Musculoskeletal System
• Myalgia & Arthralgia
Gastrointestinal Tract
• Altered bowel habits
• Abdominal pain
Heart and Blood Vessels
• Altered blood pressure responses
• Dizziness
Immune System
• Lymph node tenderness
• Sore throat
• Enhanced Cytokines

# CFS Heterogeneity

- Regarding CFS heterogeneity.
- Here are some of Q's related to CFS heterogeneity…
  - Since we observe differing insults (viral or infection or stress, … or a combination of the these) preceding the CFS state, is there a difference in what's going wrong 'under the hood' physiologically, in people with CFS?
  - Since there are natural genetic variations between people, does this lead to variations in susceptibility to the disease from these insults?
  - What is an exhaustive list of disturbed biological pathways in chronic fatigued persons – does this list vary by person?
  - What are the biological sequence of events (in terms of biological pathway disruptions) leading to a CFS state, and does this sequence of events vary from person to person?
  - Are there common themes in the answer to the above questions? Can we define subgroups of people with CFS from these common themes?
- Can we use these subgroups to develop optimal, targeted treatments - custom tailored to address each subgroups specific characteristic problems?

# Study Subjects

**This in-hospital study enrolled people who were identified with CFS according to the 1994 CFS case definition as described by Fukuda and colleagues during the 4-year longitudinal study of CFS in Wichita, KS, USA.**




People with ISF were also enrolled, in addition to those with CFS and ISF with MDDm. NF controls were selected from people surveyed in the longitudinal study that did not report fatigue, medical and psychiatric exclusionary conditions and were similar in age, race and BMI to people with CFS and ISF.
BMI: Body mass index; CFS: Chronic fatigue syndrome; ISF: Insufficient symptoms and fatigue; MDDm: Major depressive disorder with melancholy; NF: Nonfatigued.

# Data Integration to Identify Biomarkers

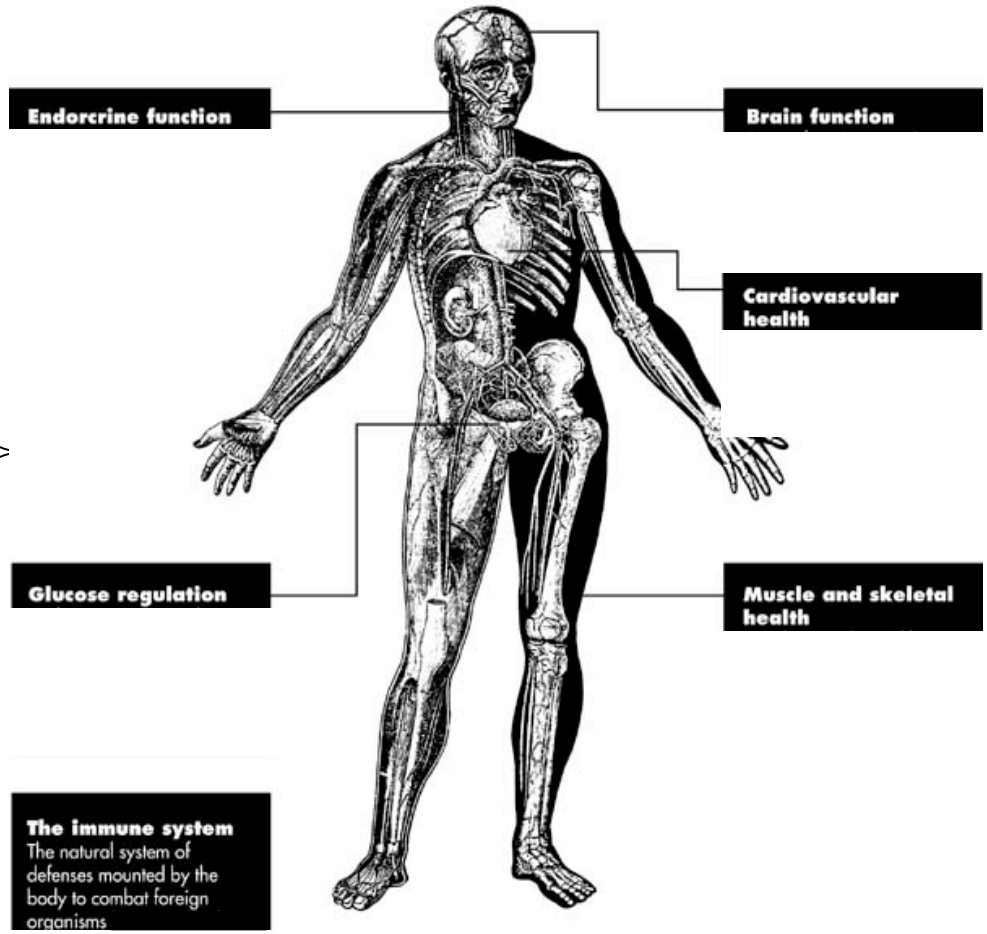
Descriptive Data
- Body – physical and clinical
- Instruments to describe symptom domains (SF-36, MFI, CDCSI)
- Psychology/psychiatric

Neuroendocrine/Immune
- HPA, HPG, HPT
- ANS
- Immune system

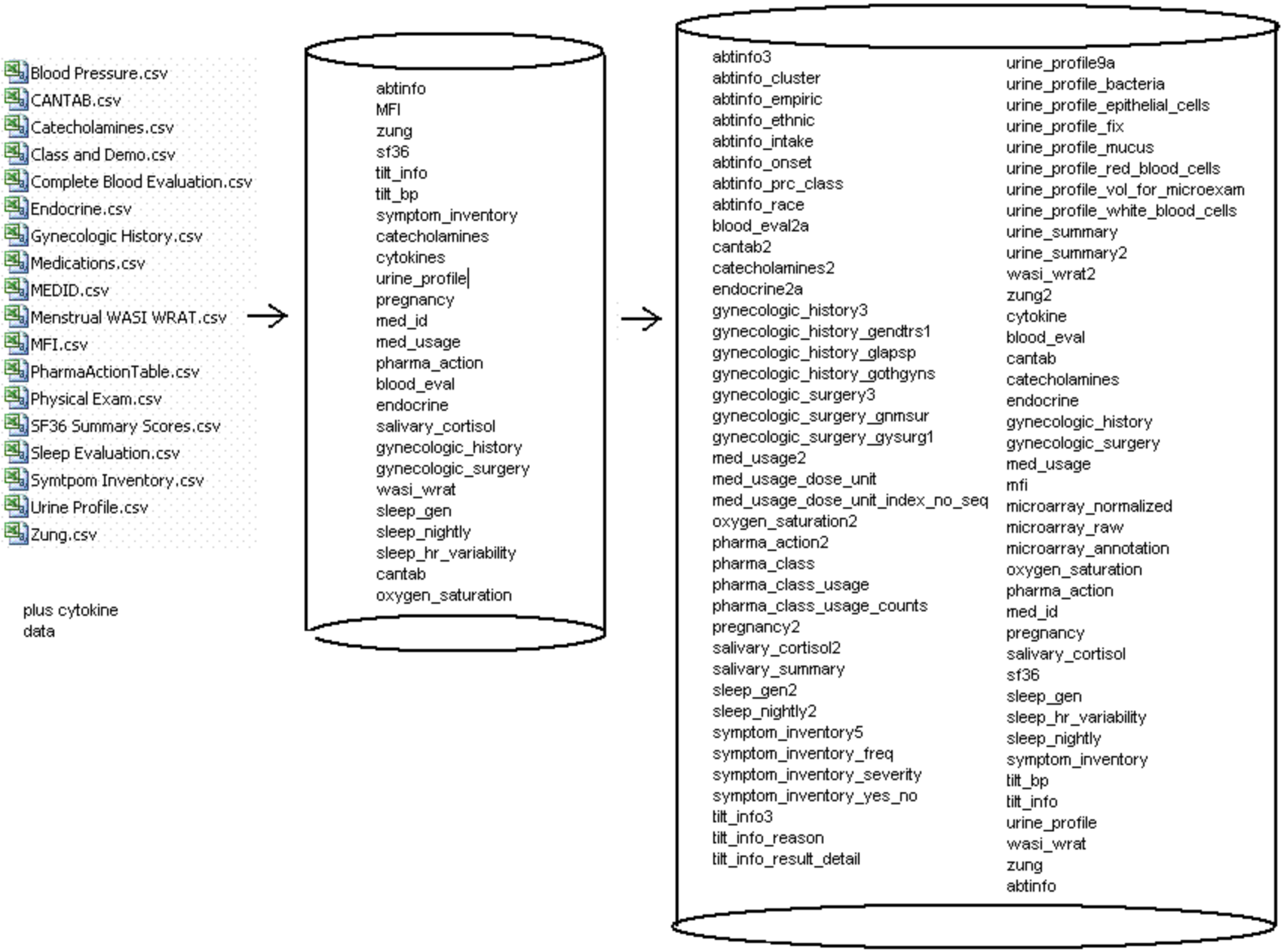
Sleep
- brain and muscle

Cognition
- CANTAB

Targeted Genetics
- single nucleotide polymorphisms of genes involved in HPA axis function



# Initial Dataset (Table List)



# Initial Dataset

| abtid | intake_index | onset_index | prc_class_in | empiric_inde | cluster_inde | years_ill | sex_bin | age | race_oid |
|---|---|---|---|---|---|---|---|---|---|
| 10043905 | 3 | 2 | 5 | 8 | 2 | 26.3 | 0 | 40 | 3 |
| 10081101 | 1 | 2 | 5 | 2 | 3 | 18.2 | 0 | 58 | 6 |
| 10103103 | 3 | 2 | 12 | 8 | 3 | 12.1 | 0 | 50 | 6 |
| 10193601 | 3 | 2 | 5 | 8 | 3 | 8 | 0 | 57 | 6 |
| 10203401 | 1 | 2 | 3 | 5 | 1 | 9.8 | 0 | 54 | 6 |
| 10215301 | 5 | 1 | 11 | 13 | 2 | | 0 | 57 | 6 |
| 10215901 | 5 | 1 | 15 | 15 | 1 | | 0 | 35 | 6 |
| 10240402 | 1 | 2 | 2 | 8 | 3 | 8.7 | 1 | 43 | 6 |
| 10243501 | 5 | 1 | 11 | 8 | 2 | | 1 | 41 | 6 |
| 10261501 | 5 | 1 | 11 | 13 | 2 | | 0 | 53 | 6 |
| 10268605 | 2 | 2 | 5 | 2 | 4 | 9.3 | 0 | 54 | 6 |
| 10323401 | 5 | 1 | 11 | 13 | 2 | | 0 | 58 | 6 |
| 10372601 | 1 | 2 | 3 | 5 | 1 | 10.7 | 0 | 54 | 6 |
| 10528403 | 1 | 2 | 1 | 1 | 1 | 10.7 | 0 | 45 | 6 |
| 10689003 | 1 | 2 | 5 | 2 | 3 | 7.6 | 0 | 47 | 6 |
| 10803801 | 3 | 2 | 5 | 8 | 3 | 8.6 | 1 | 59 | 6 |
| 10860201 | 5 | 1 | 11 | 13 | 3 | | 1 | 39 | 6 |
| 20052705 | 3 | 2 | 6 | 11 | 1 | 12.5 | 0 | 43 | 6 |
| 20077904 | 1 | 2 | 6 | 15 | 1 | 31.3 | 1 | 37 | 6 |
| 20082302 | 1 | 2 | 5 | 8 | 3 | 5.3 | 0 | 53 | 6 |
| 20086501 | 2 | 2 | 9 | 3 | 1 | 12.3 | 0 | 61 | 6 |
| 20129103 | 5 | 1 | 15 | 15 | 1 | | 0 | 65 | 3 |
| 20366001 | 3 | 2 | 6 | 5 | 1 | 8.8 | 0 | 48 | 6 |
| 20416901 | 2 | 1 | 3 | 11 | 1 | 14.3 | 0 | 62 | 6 |

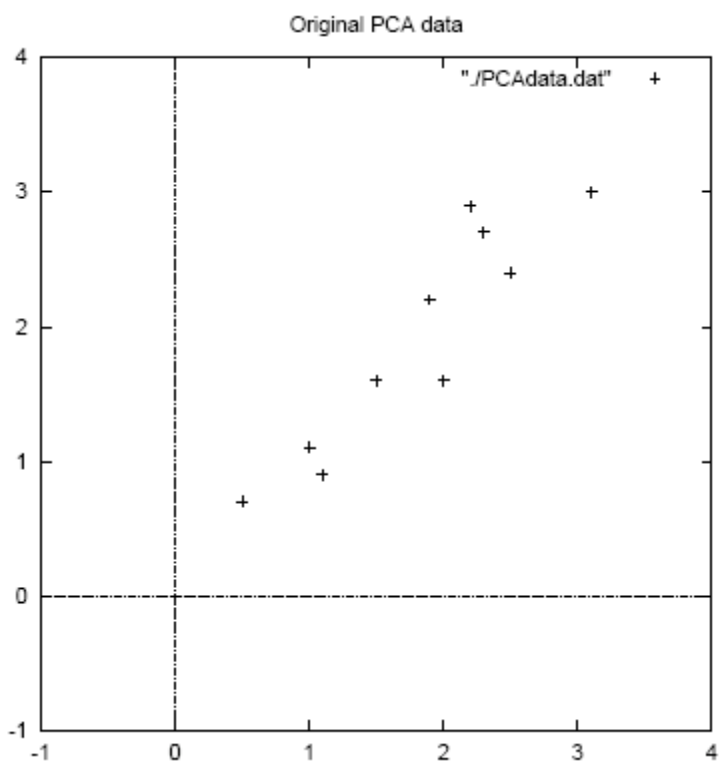
# 2 Approaches

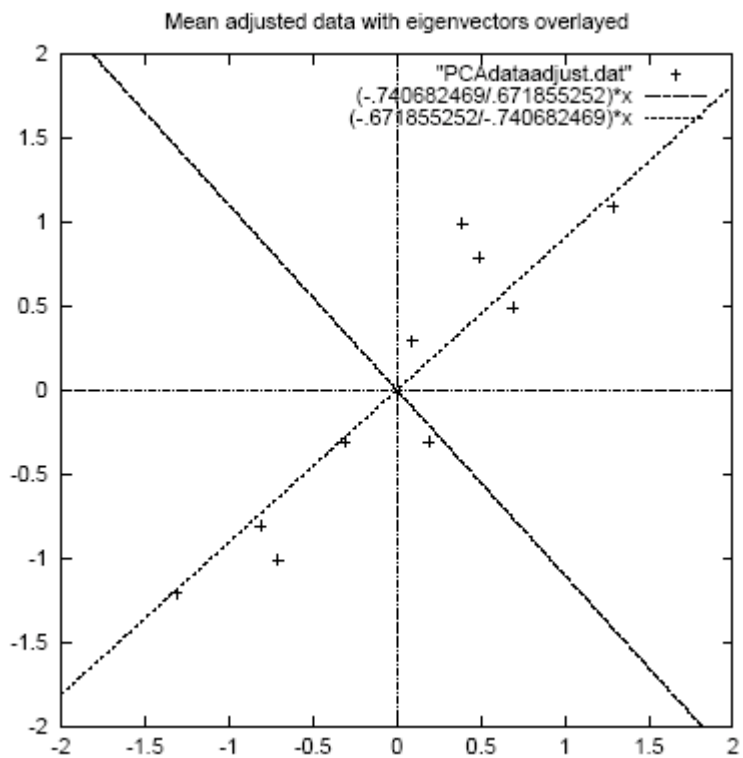
- PCA/LCA
  - Top Down

- Subspace Clustering
  - Bottoms Up

# PCA

- Utilized correlation coefficient based PCA
- Reduces dimensionality while preserving full variation
- Helps to establish a variable set which spans variations in the dataset



# PCA

| PV1 | PV1 | PV2 | PV2 | PV3 | PV3 | PV4 | PV4 |
|---|---|---|---|---|---|---|---|
| Eigenvalue | 28.49034 | Eigenvalue | 14.94629 | Eigenvalue | 14.19392 | Eigenvalue | 12.98039 |
| Percent | 6.475078 | Percent | 3.396883 | Percent | 3.225891 | Percent | 2.950088 |
| Cum Percer | 6.475078 | Cum Percer | 9.871961 | Cum Percer | 13.09785 | Cum Percent | 16.04794 |

| PV1 | PV1 | PV2 | PV2 | PV3 | PV3 | PV4 | PV4 |
|---|---|---|---|---|---|---|---|
| mfi.generalfa | -0.12066 | wasi_wrat2.\ | 0.240773 | gynecologic_ | 0.167289 | oxygen_saturation2.o2lowp | -0.14164 |
| sf36.vitality | 0.123514 | wasi_wrat2.\ | 0.240769 | gynecologic_ | 0.167064 | gynecologic_history3.gprdbld | -0.14111 |
| sf36.social_fi | -0.10374 | wasi_wrat2.\ | 0.235238 | gynecologic_ | 0.142115 | abtinfo3.waist_circum | 0.140854 |
| abtinfo3.emp | -0.1406 | wasi_wrat2.\ | 0.212085 | gynecologic_ | 0.140183 | gynecologic_history3.gnrmyr | -0.13837 |
| symptom_in | 0.039662 | wasi_wrat2.\ | 0.21114 | gynecologic_ | -0.1398 | gynecologic_history3.gbfryr | -0.13837 |
| zung2.sdsin | 0.012426 | wasi_wrat2.\ | 0.209433 | gynecologic_ | 0.138873 | gynecologic_history3.gbfrdy | -0.13635 |
| symptom_in | -0.02687 | wasi_wrat2.\ | 0.209421 | gynecologic_ | -0.13223 | gynecologic_history3.gnrmdy | -0.13272 |
| mfi.physicalfa | 0.022401 | wasi_wrat2.\ | 0.208413 | gynecologic_ | 0.127921 | gynecologic_history3.gbfrmn | -0.13098 |
| symptom_in | 0.002692 | wasi_wrat2.\ | 0.205281 | pharma_clas | 0.125068 | gynecologic_history3.gprdlst | -0.13093 |
| sf36.bodily_p | -0.01031 | wasi_wrat2.\ | 0.197316 | gynecologic_ | 0.124138 | gynecologic_history3.gprdapt | -0.12986 |
| sf36.role_phy | 0.015727 | wasi_wrat2.\ | 0.196524 | gynecologic_ | 0.124031 | abtinfo3.weight_lbs | 0.12968 |
| mfi.activityre | 0.006969 | wasi_wrat2.\ | 0.194707 | pharma_clas | 0.123418 | gynecologic_history3.gnrmmn | -0.12956 |
| sf36.gnrl_hlt | 0.005082 | wasi_wrat2.\ | 0.19301 | gynecologic_ | 0.123342 | pharma_class_usage.antihypertensive | 0.125211 |
| symptom_in | -0.00025 | wasi_wrat2.\ | 0.185406 | gynecologic_ | -0.123 | sleep_nightly2.rdi | 0.124787 |
| mfi.motivatio | -0.01941 | wasi_wrat2.\ | 0.184849 | gynecologic_ | -0.12271 | abtinfo3.bmi | 0.123688 |
| mfi.mentalfat | 0.001438 | wasi_wrat2.\ | 0.171769 | gynecologic_ | 0.122031 | sleep_nightly2.apnea_hyponea_total | 0.122246 |
| abtinfo3.ons | 0.018855 | wasi_wrat2.\ | 0.161124 | gynecologic_ | 0.122001 | sleep_nightly2.hyponea | 0.119969 |
| sf36.phys_fu | -0.01398 | sleep_hr_var | -0.09508 | gynecologic_ | 0.121537 | oxygen_saturation2.o2avgp | -0.11684 |

# Attribute Reduction

| Dataset Name | Notes |
|---|---|
| all_data_outer | all data - with null data elements (outer join) |
| all_data_inner | all data - no null data elements (inner join) |
| bio_data_outer | contains only the non-survey (biological) data, remove sf36, mfi, gynecologic_history, symptom_inventory |
| select_data_outer | remove blood_eval2 cols, cantab cols, endocrine2A cols, gynacoloic_history3 cols, mfi cols, sf36 cols, sleep_hr_variability cols, symptom_inventory cols, tilt_bp cols, tilt_info3 cols, uring_profile9A cols,wasi_wrat2 cols |
| select2_data_outer | remove pharma_class_usage |
| select3_data_outer | remove sleep related variables |
| select5_data_outer | remove redundant measures such as 3 measures of testo |
| … | |
| select9_data_outer | remove waist_circum, cytokine cols, some endocrine cols (androstenedione), mfi cols, sf36 cols, sleep_nightly2 cols |
| select10_data_outer | remove med/psych exclusions |
| select11_data_outer | remove men |
| select12_data_outer | remove DHEA |
| select13_data_outer | remove gabdpn, years_ill_index, potassium. Wbc, rti_five_choice_reaction_time_index, soc_num_problems_solved_in_min_num_moves_index swm_num_total_errors_index, metanephrine, epinephrine, thyroxin_t4_index, lm_index, specific_gravity |

# Class attributes in order of significance to the solution

| Class attributes | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 |
|---|---|---|---|---|---|---|
| % unrefreshing_sleep | 98 | 13 | 79 | 78 | 95 | 100 |
| high_sensitivity_crp | 5.3 | 4.2 | 2.1 | 0.9 | 1 | 5.1 |
| % sleep_problems | 100 | 29 | 75 | 96 | 91 | 91 |
| $O_2$lowp | 85 | 88 | 86 | 90 | 89 | 89 |
| %sore_throat | 29 | 8 | 17 | 13 | 59 | 73 |
| % photophobia | 56 | 11 | 38 | 22 | 64 | 73 |
| % abdominal_pain | 6 | 5 | 4 | 35 | 68 | 45 |
| % concentration | 27 | 0 | 21 | 4 | 86 | 45 |
| % joint_pain | 63 | 32 | 54 | 30 | 91 | 82 |
| rdi_index | 3.8 | 1 | 6 | 0.4 | 0.8 | 1.1 |
| insulin_serum | 10 | 7 | 10 | 5 | 7 | 5 |
| epworth | 11 | 6 | 7 | 8 | 10 | 9 |
| % short_breath | 32 | 5 | 8 | 4 | 50 | 45 |
| progesterone | 14 | 12 | 17 | 35 | 37 | 0 |
| % post_exertional_fatique | 78 | 0 | 33 | 43 | 82 | 91 |
| % fever | 10 | 5 | 4 | 13 | 50 | 18 |
| % muscle_pain | 93 | 47 | 54 | 70 | 100 | 91 |
| BMI | 32 | 30 | 30 | 24 | 27 | 26 |
| free_t3 | 3 | 2 | 2 | 2 | 2 | 2 |
| sleep_period_time_spt | 471 | 452 | 467 | 456 | 448 | 439 |
| il_6 | 68 | 56 | 45 | 32 | 66 | 50 |
| arousals_total_index_psg | 18.5 | 11.2 | 13.4 | 13.4 | 9.9 | 17.1 |
| neck_circum | 37 | 35 | 35 | 33 | 35 | 33 |
| hgb | 14.2 | 13.8 | 13.6 | 13.5 | 13.6 | 12.8 |
| sleep_HR variability(sdnn) | 60 | 62 | 38 | 50 | 54 | 37 |
| pct_free_testo_ratio | 2.6 | 1.6 | 2 | 2.5 | 2.6 | 0.8 |
| % headache | 66 | 42 | 50 | 48 | 91 | 64 |
| latency_to_sleep_onset | 8 | 9 | 10 | 5 | 14 | 10 |
| urine_free_cortisol_24h | 17 | 19 | 10 | 21 | 20 | 14 |
| ZUNG | 50 | 35 | 51 | 48 | 54 | 55 |

Well

Class 1: OBESE, HYPNOEIC, FATIGUED & PAINED

Class 2: WELL

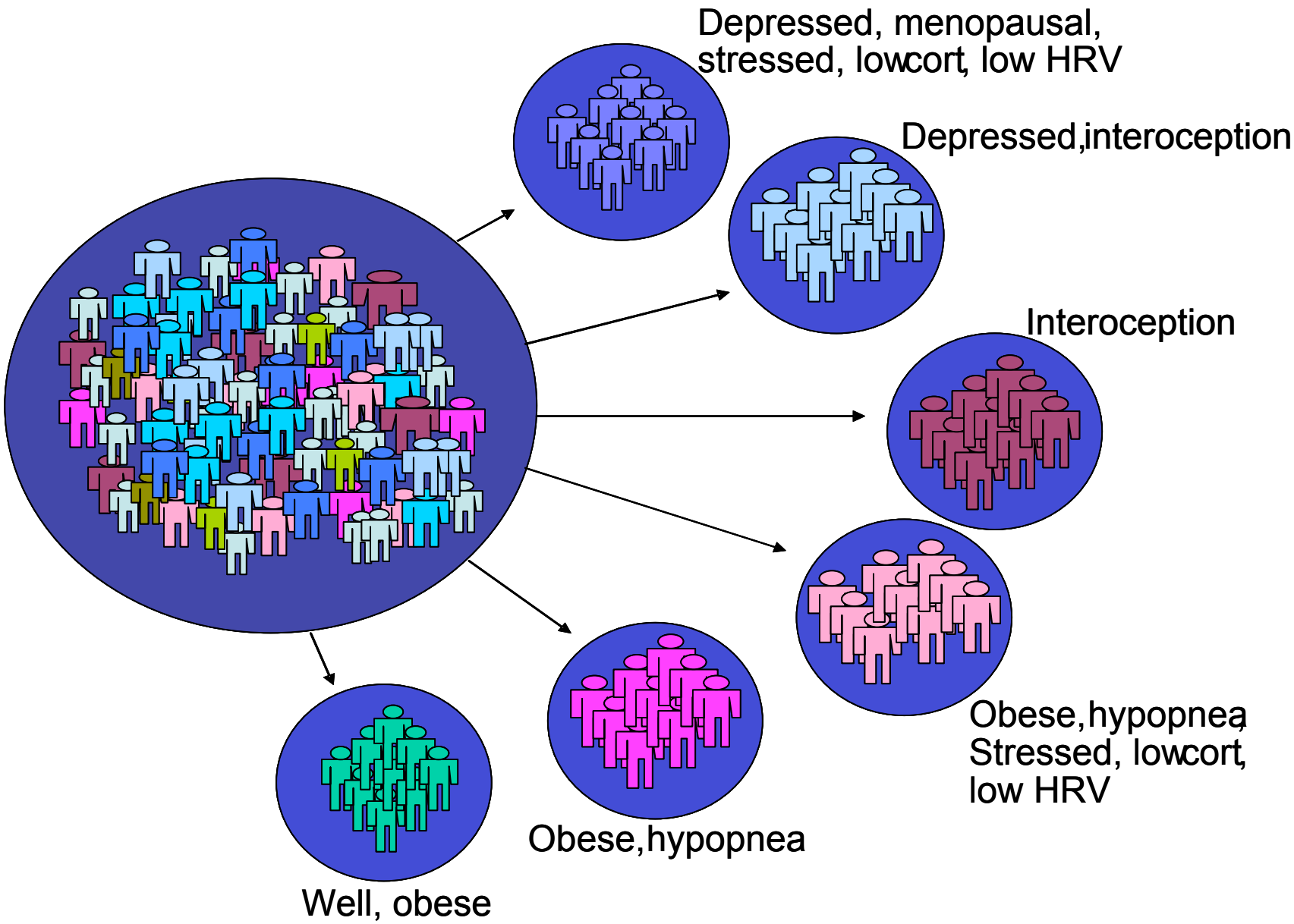
Class3: HYPNOEIC, OBESE, STRESSED - LOW CORTISOL

Class 4: YOUNG, normal BMI, fewer symptoms besides SLEEP

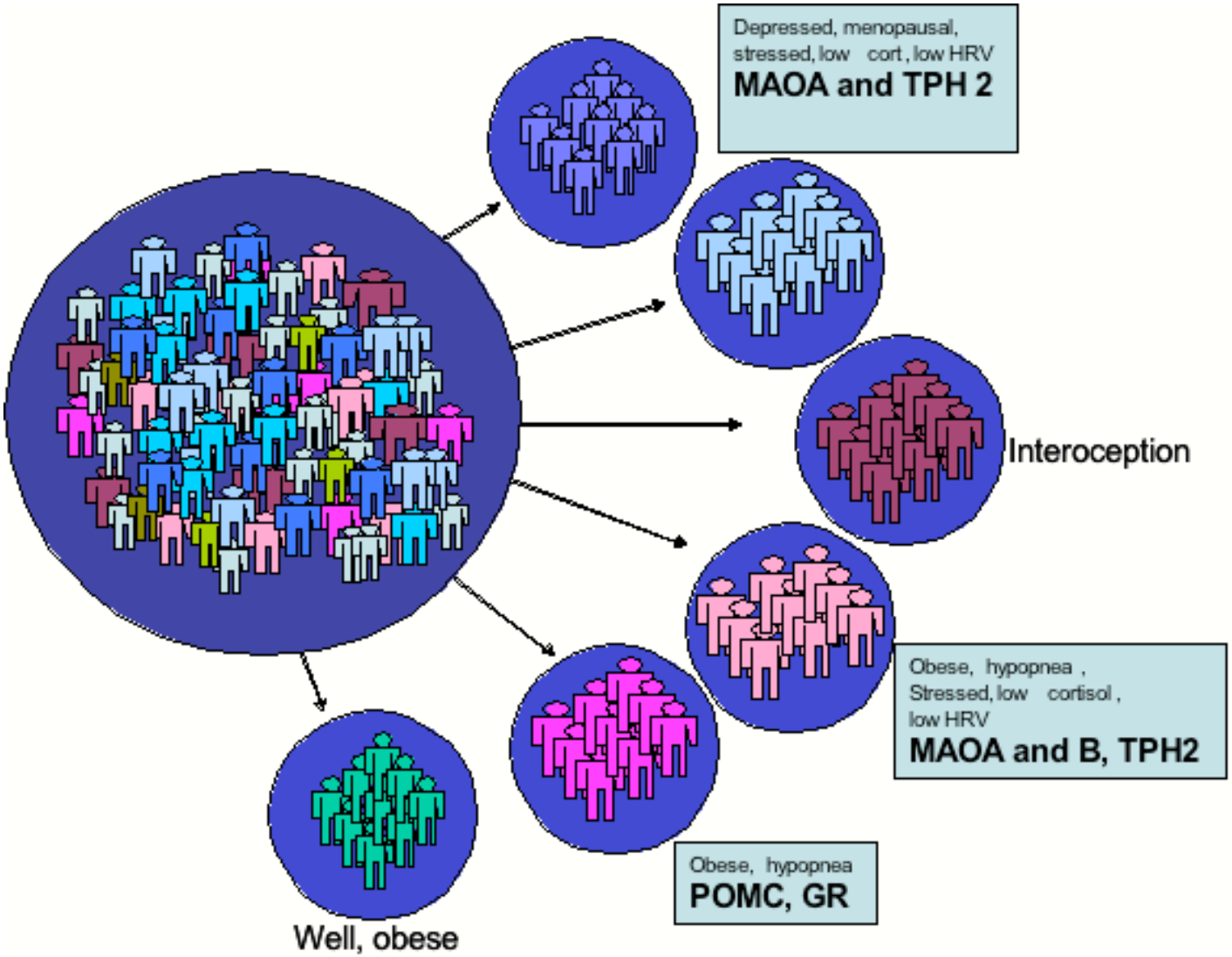
Class 5: INTEROCEPTIVE, depressed & symptomatic

Class 6: SYMPTOMATIC, OLDer, THINer, DEPRESSED, STRESSED, poor sleep, NO SEX,

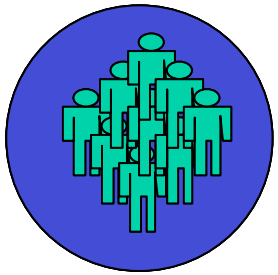
# CFS is Heterogeneous



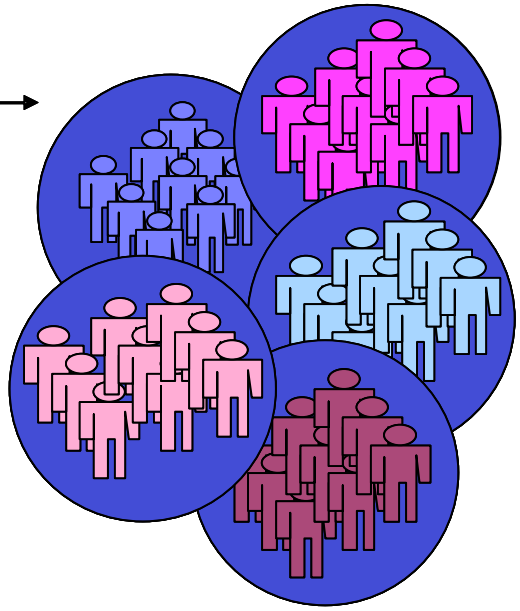
# The Genomics of CFS



# Gene Expression Differentiates Fatigue



1. **Vac14**: regulates phosphatidylinositol kinases (stress response and membrane trafficking)
2. **SLC1A6**: an excitatory aa transporter (glutatmate/aspartate)
3. **Fbxo7:** Fbxo7 has been characterised as a selective enhancer of cdk6 activity (regulate major cell cycle transitions)
4. **ZNF350**: crucial roles in ubiquitination events involved in diverse cellular processes including signal transduction (MAPK), differentiation and apoptosis



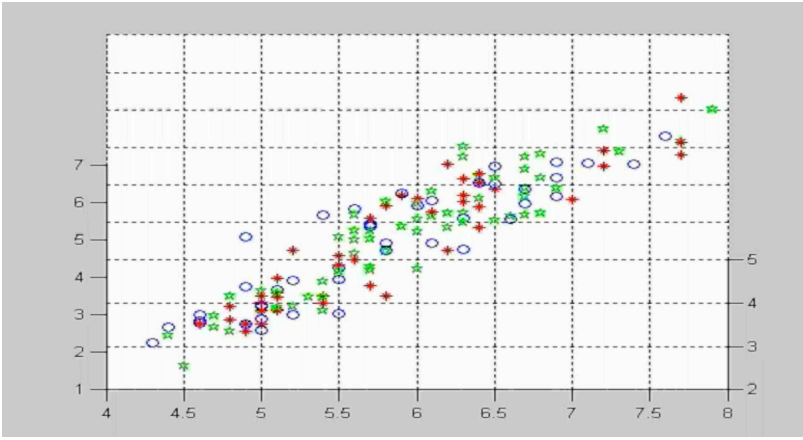
1. **PTCH2**: receptor for shh signaling which is active in T cell growth and differentiation and proliferation
2. **TCL1A:** TCL1A regulates the growth and survival of peripheral T cells

# CFS Heterogeneity

- Problems with PCA/LCA approach.
  - Are we studying illness end-stage symptomatology rather than more helpful subgroups?
- More helpful subgroups may answer these Q's related to CFS heterogeneity…
  - Since we observe differing insults (viral or infection or stress, … or a combination of the these) preceding the CFS state, is there a difference in what's going wrong 'under the hood' physiologically, in people with CFS?
  - Since there are natural genetic variations between people, does this lead to variations in susceptibility to the disease from these insults?
  - What is an exhaustive list of disturbed biological pathways in chronic fatigued persons – does this list vary by person?
  - What are the biological sequence of events (in terms of biological pathway disruptions) leading to a CFS state, and does this sequence of events vary from person to person?
  - Are there common themes in the answer to the above questions? Can we define subgroups of people with CFS from these common themes?
- Can we use these subgroups to develop optimal, targeted treatments - custom tailored to address each subgroups specific characteristic problems?

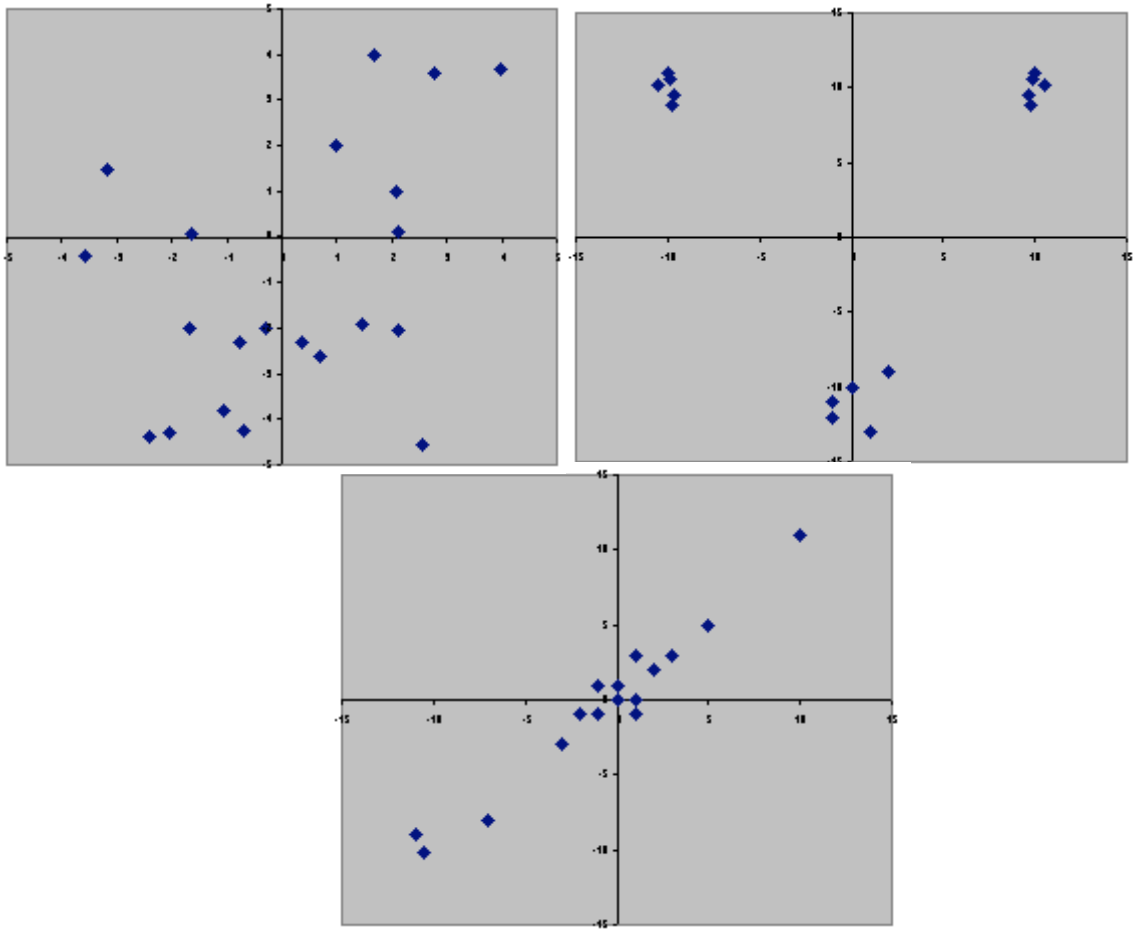
# Clustering

- Clustering algorithms find associations between sets of variables and subjects by identifying regions or clusters of closely spaced values



# Why Clustering?

- Differs from variance based calculations
- Can elucidates trends obscured through heterogeneous data

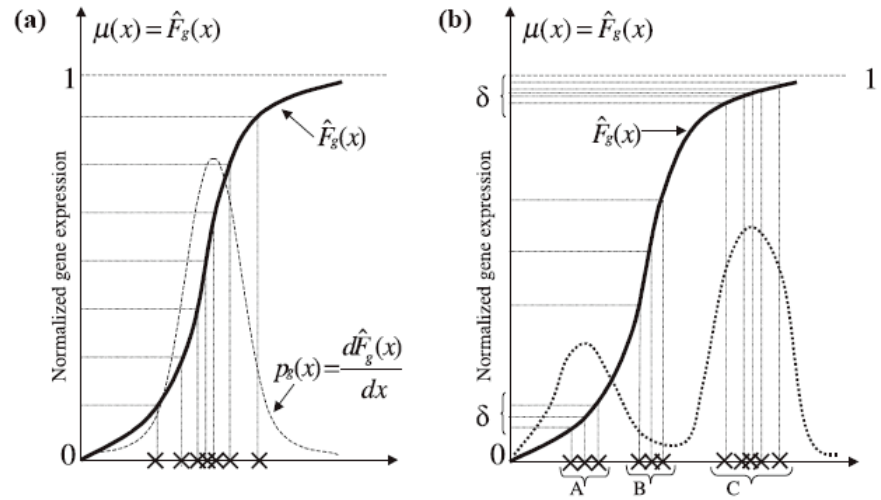


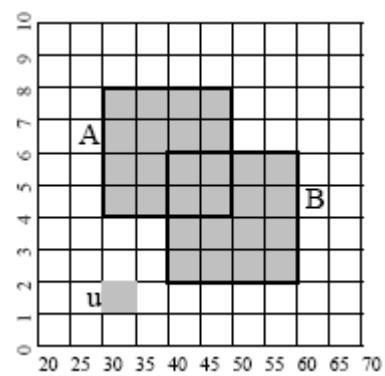
# Why Subspace Clustering?

Relative high dimensionality of dataset (3686 variables) compared to the number of participants (227) favors subspace clustering algorithms

# Subspace Clustering

- Subspace clustering considers all combinations of subsets of the original space
- We utilize portions of IBM's Genes@Work tool (clique) which identifies clusters under a user specified p-value
- the algorithm identifies maximally sized clusters

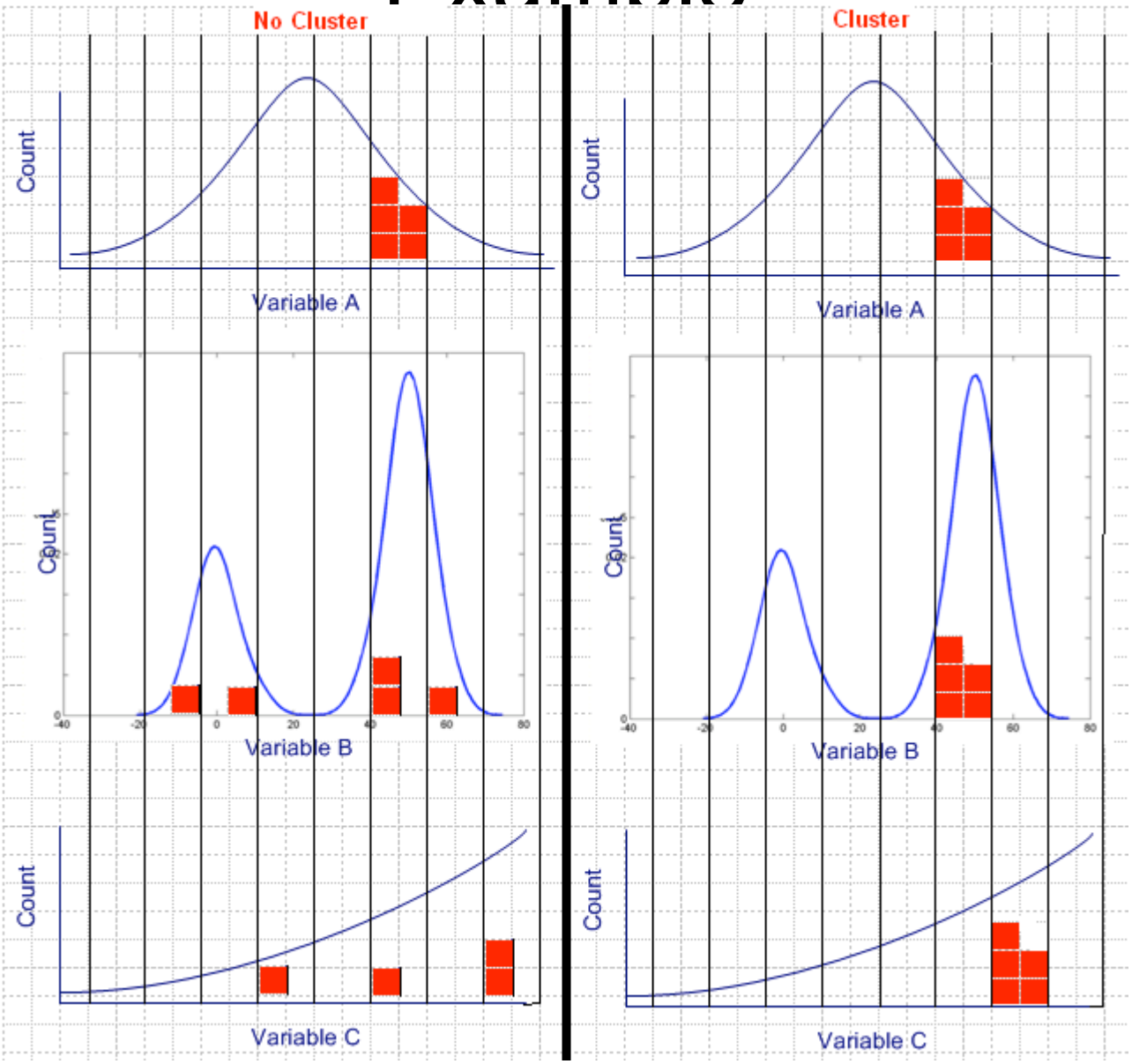



Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, Agrawal et. al. (1998)

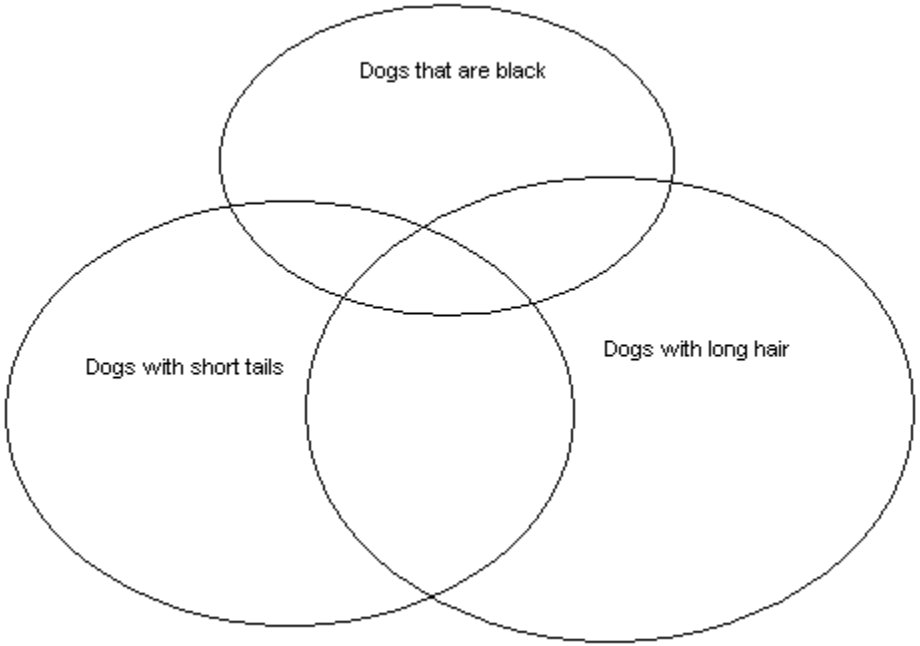
Genes@Work: an efficient algorithm for pattern discovery and multivariate feature selection in gene expression data, Lepre, J. et. al. Bioinformatics 2004

# Example


No Cluster
Cluster
Count
Variable A
Count
Variable B
Count
Variable C
Count
Variable A
Count
Variable B
Count
Variable C

# Subspace Clustering Interpretation

- We utilize VENN diagrams to visualize the information in the clusters

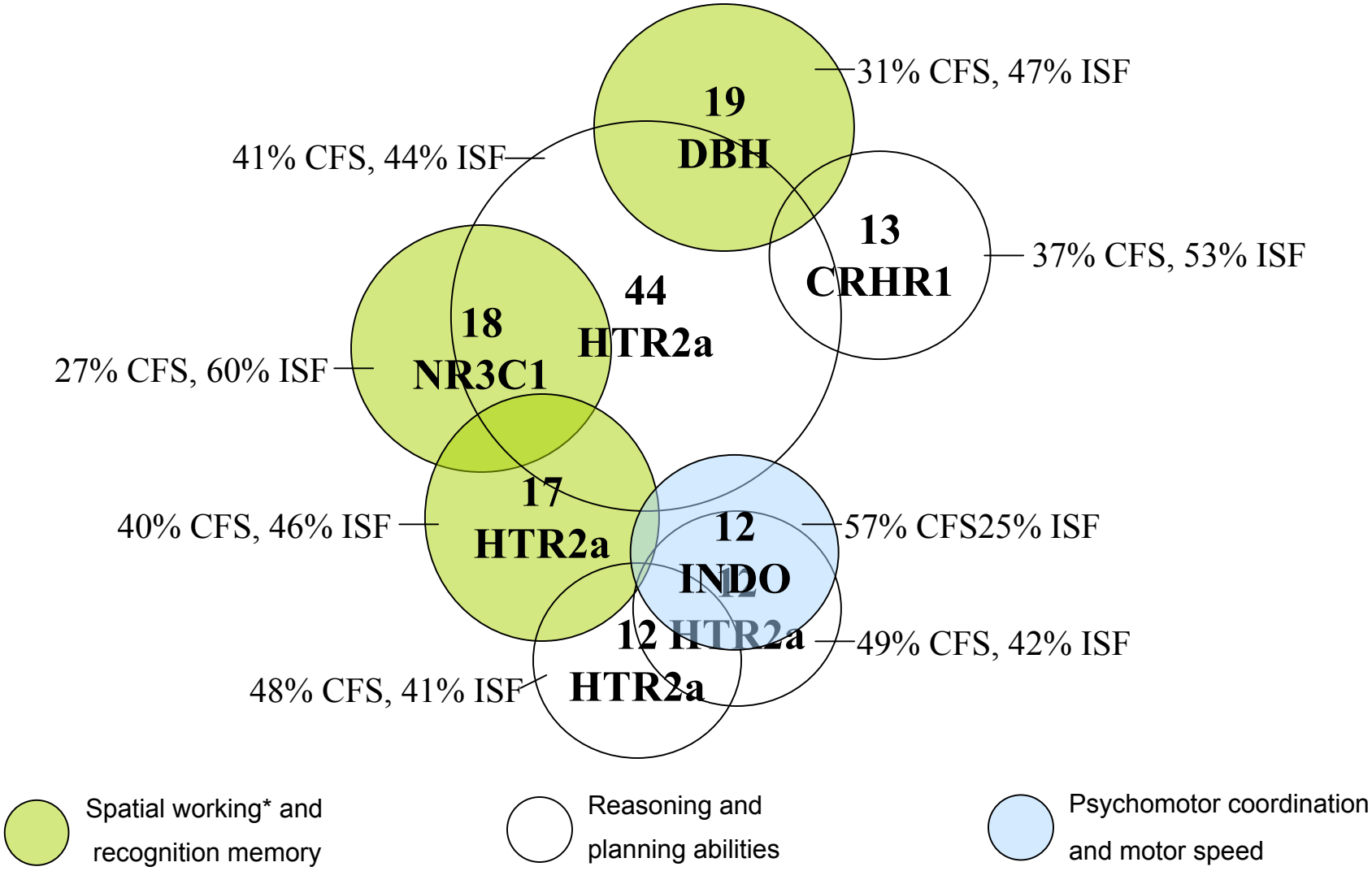


# Results

- Stringent selection of significant clusters
  - 183 of 1,042 (18%) significant clusters with (p-value = 0.005) were free of null values and analyzed further
  - Three major “themes” were identified
    - Cognition (29 clusters)
    - Sleep (31 clusters)
    - Allostatic load (7 clusters)

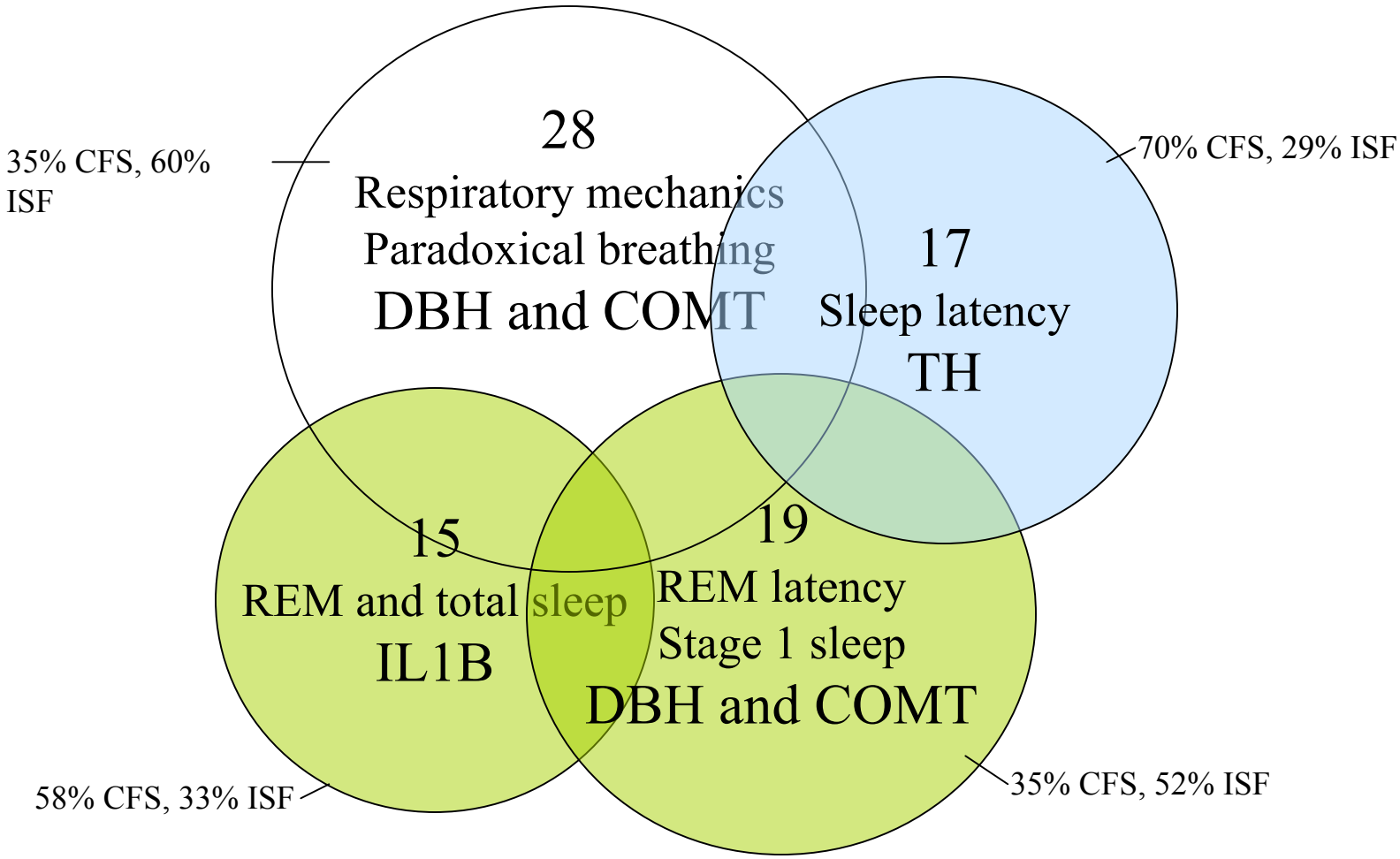
# Cognition

The CANTAB was used to assess function in the basal ganglia (especially psychomotor retardation) and/or frontal cortex (attention shift and problem solving/planning abilities).

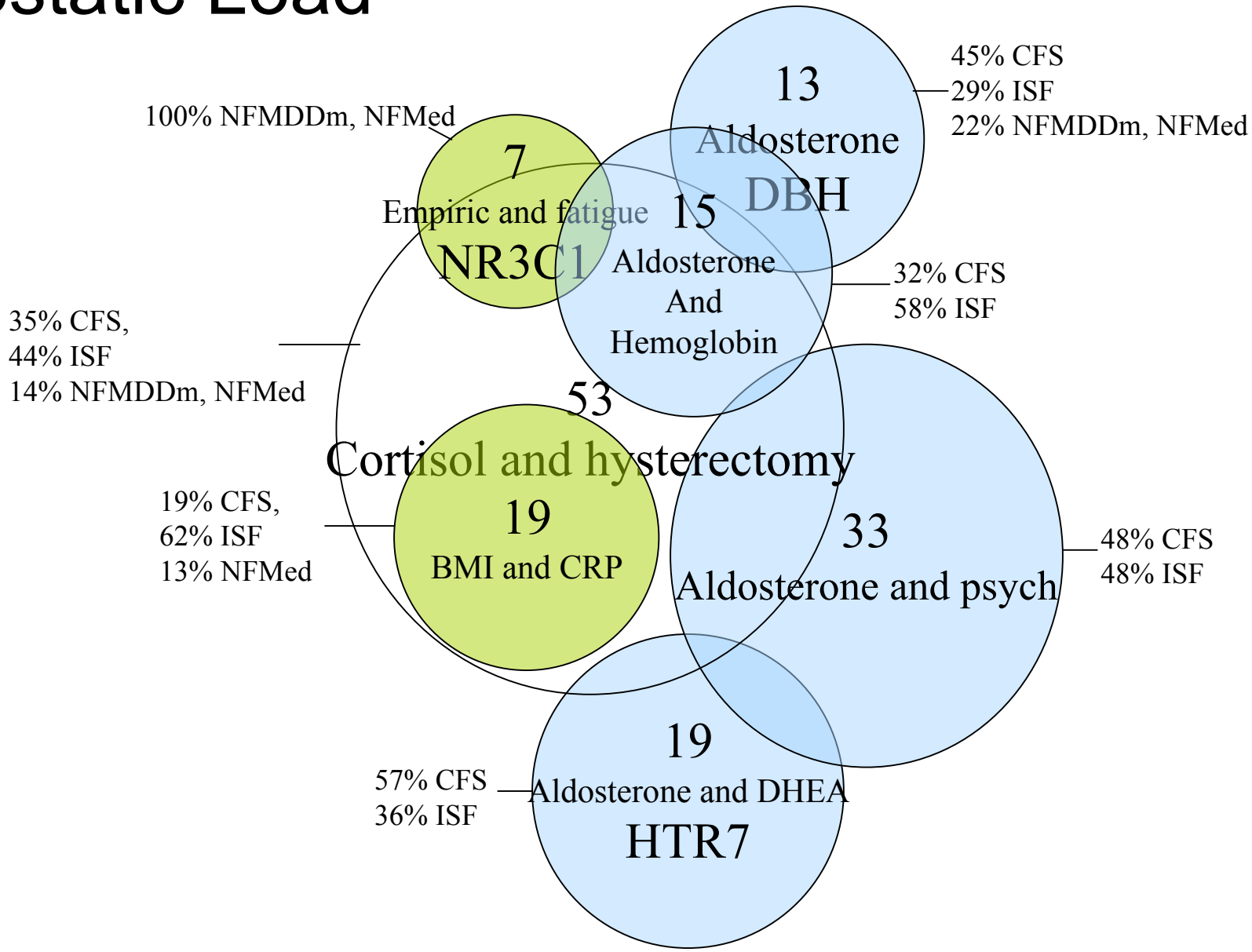


# Sleep

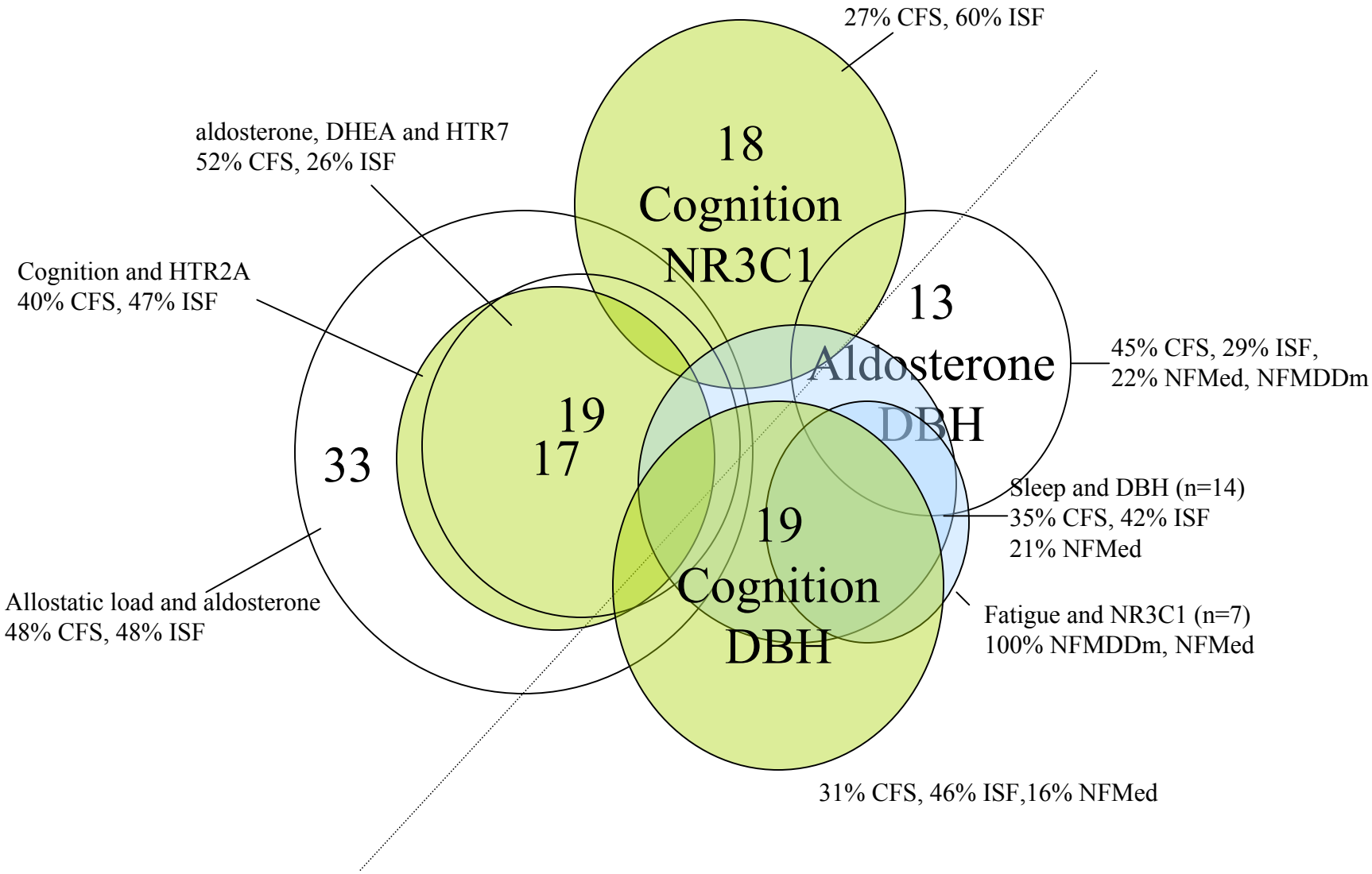
Polysomnographic monitoring and multiple sleep latency tests were used to diagnose sleep pathology.



# Allostatic Load



# Important Clusters



# Summary

- Heterogeneity at a biological level has been demonstrated through PCA/LCA and subspace clustering techniques
- subspace clustering technique validated existing results and revealed hidden associations
- Predominant themes are cognition, sleep and allostatic load

# Heterogeneity Revealed