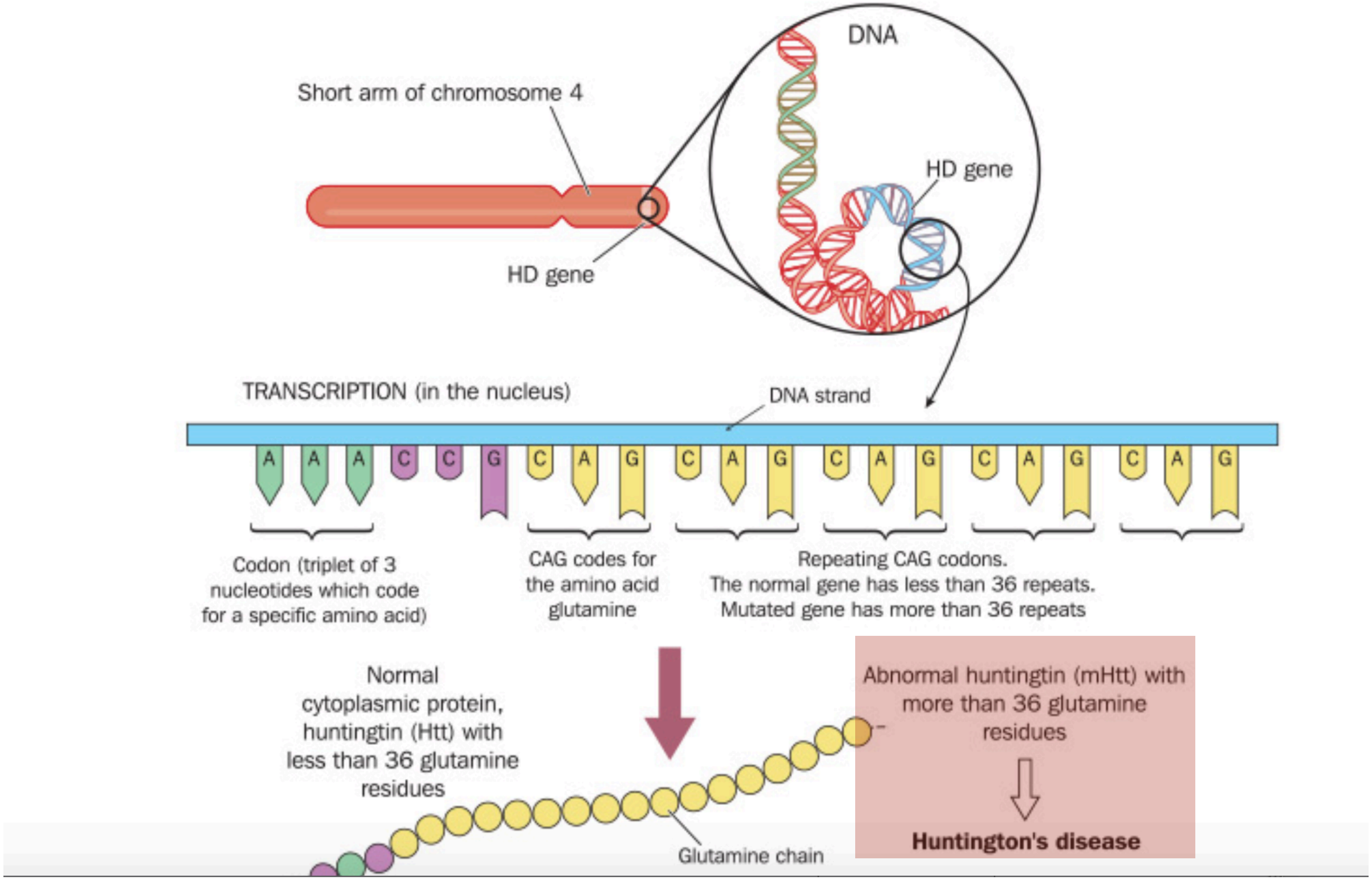
An Exploration of Latent Structure in Observational Huntington’s Disease Studies

Soumya Ghosh[1], *Zhaonan Sun*[1], *Ying Li*[1], *Yu Cheng*[1], *Amrita Mohan*[2], *Cristina Sampaio*[2], *and Jianying Hu*[1],

[1] *Center for Computational Health, IBM Research*

[2] *CHDI Foundation*

Huntington’s Disease (HD)



Codon (triplet of 3 nucleotides which code for a specific amino acid)

CAG codes for the amino acid glutamine

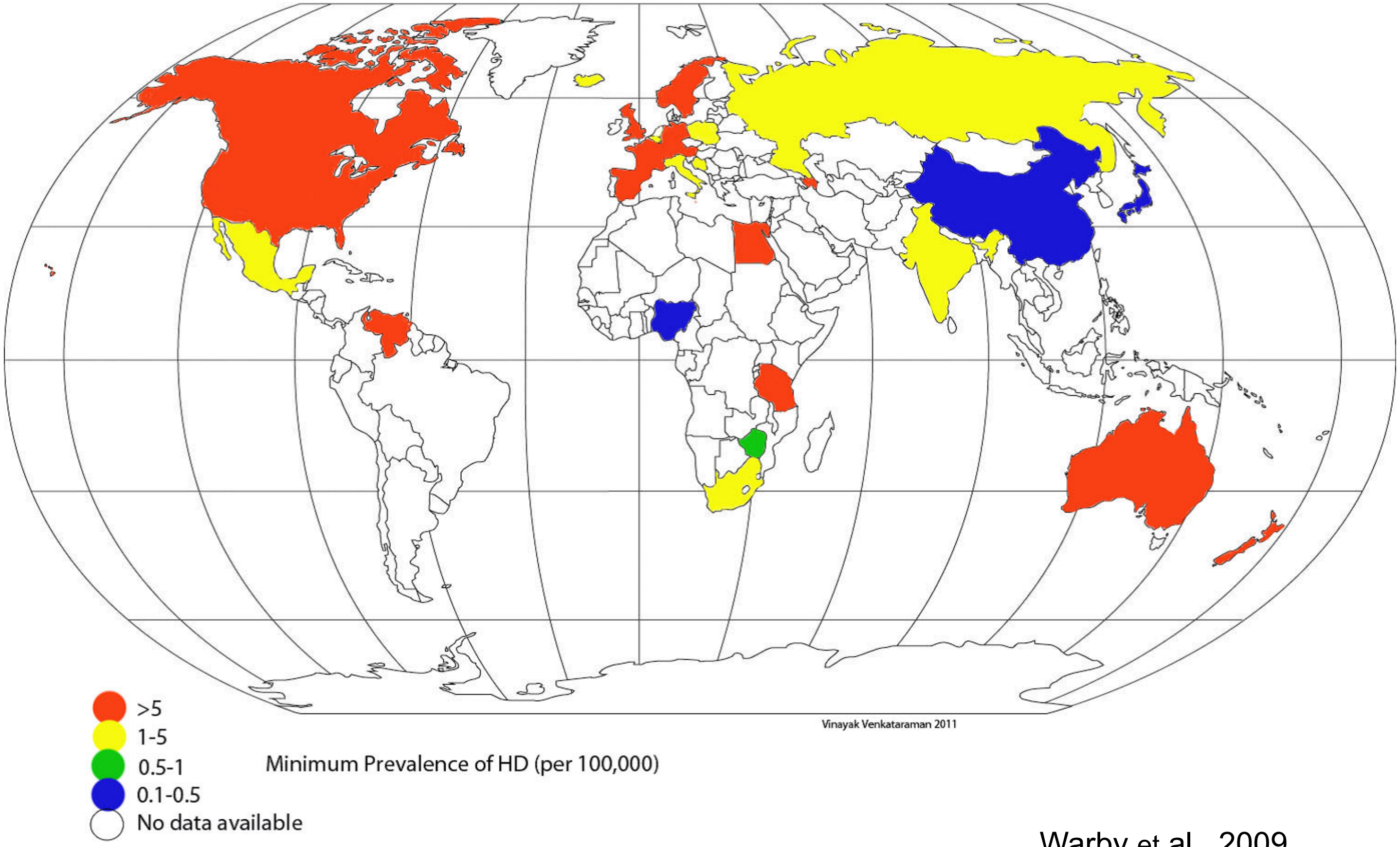
Repeating CAG codons.
The normal gene has less than 36 repeats.
Mutated gene has more than 36 repeats

Normal cytoplasmic protein, huntingtin (Htt) with less than 36 glutamine residues

Abnormal huntingtin (mHtt) with more than 36 glutamine residues

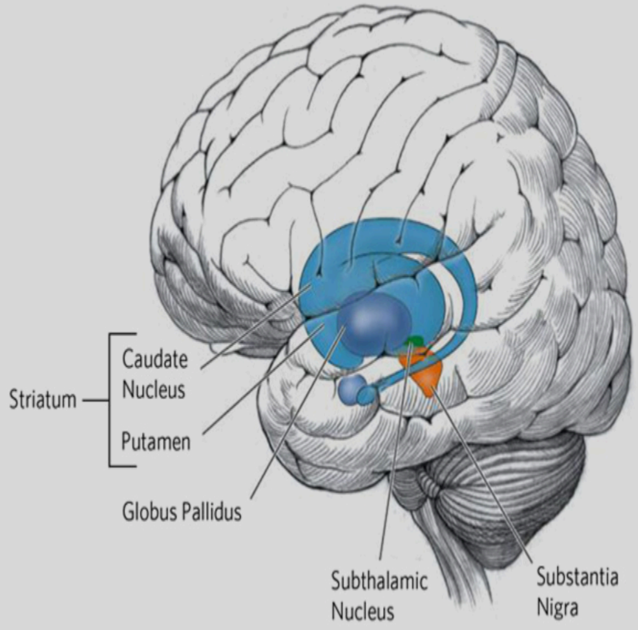
source: hdsa.org

HD prevalence


Vinayak Venkataraman 2011
🔴 >5
🟡 1-5
🟢 0.5-1
🔵 0.1-0.5
○ No data available
Minimum Prevalence of HD (per 100,000)

Warby et al., 2009

HD symptoms

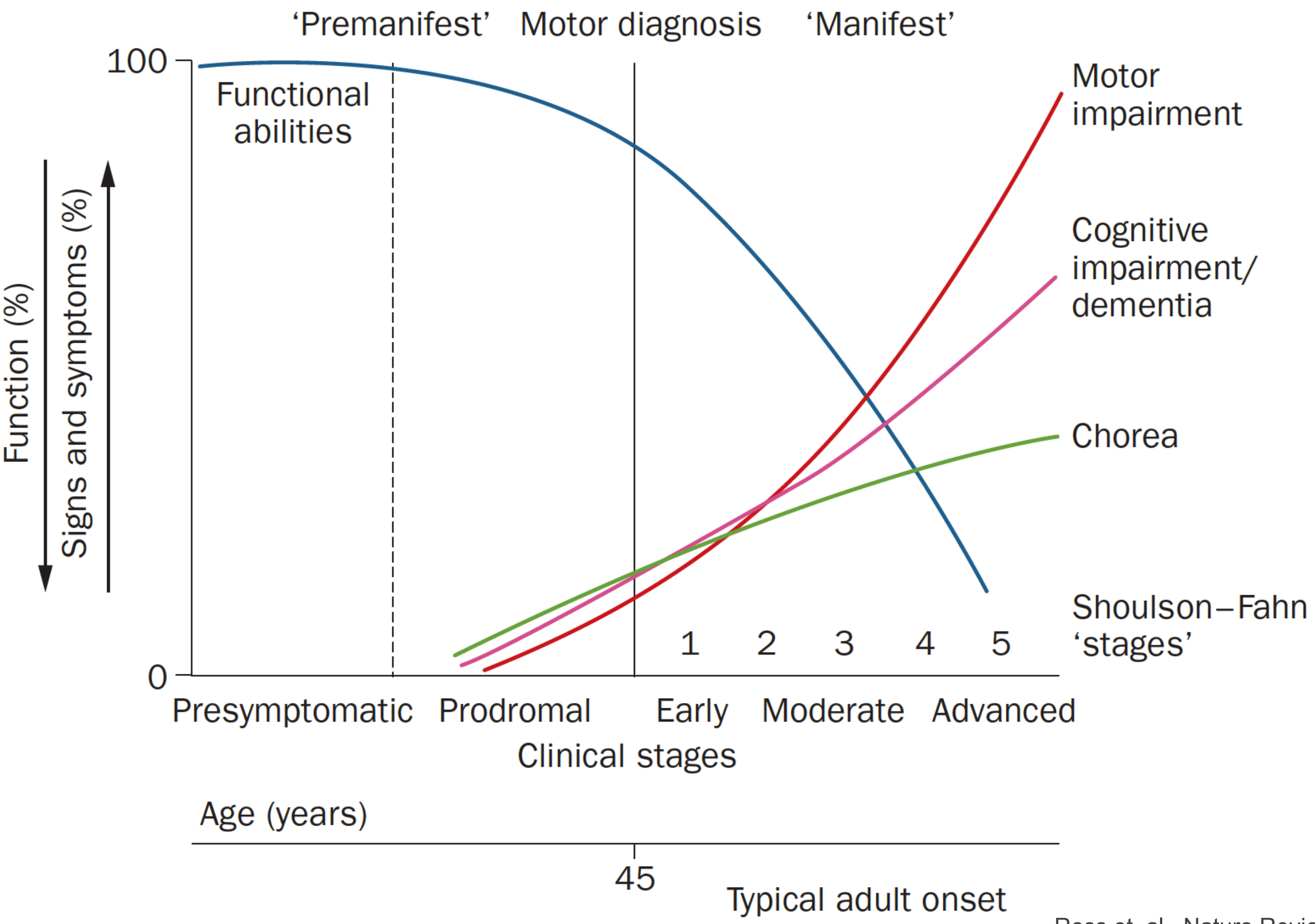


HD affects the whole brain, but certain areas are more vulnerable than others. Pictured above are the basal ganglia - a group of nerves cell clusters, called nuclei. These nuclei play a key role in movement and behavior control and are the parts of the brain most prominently affected in early HD.

source: hdsa.org

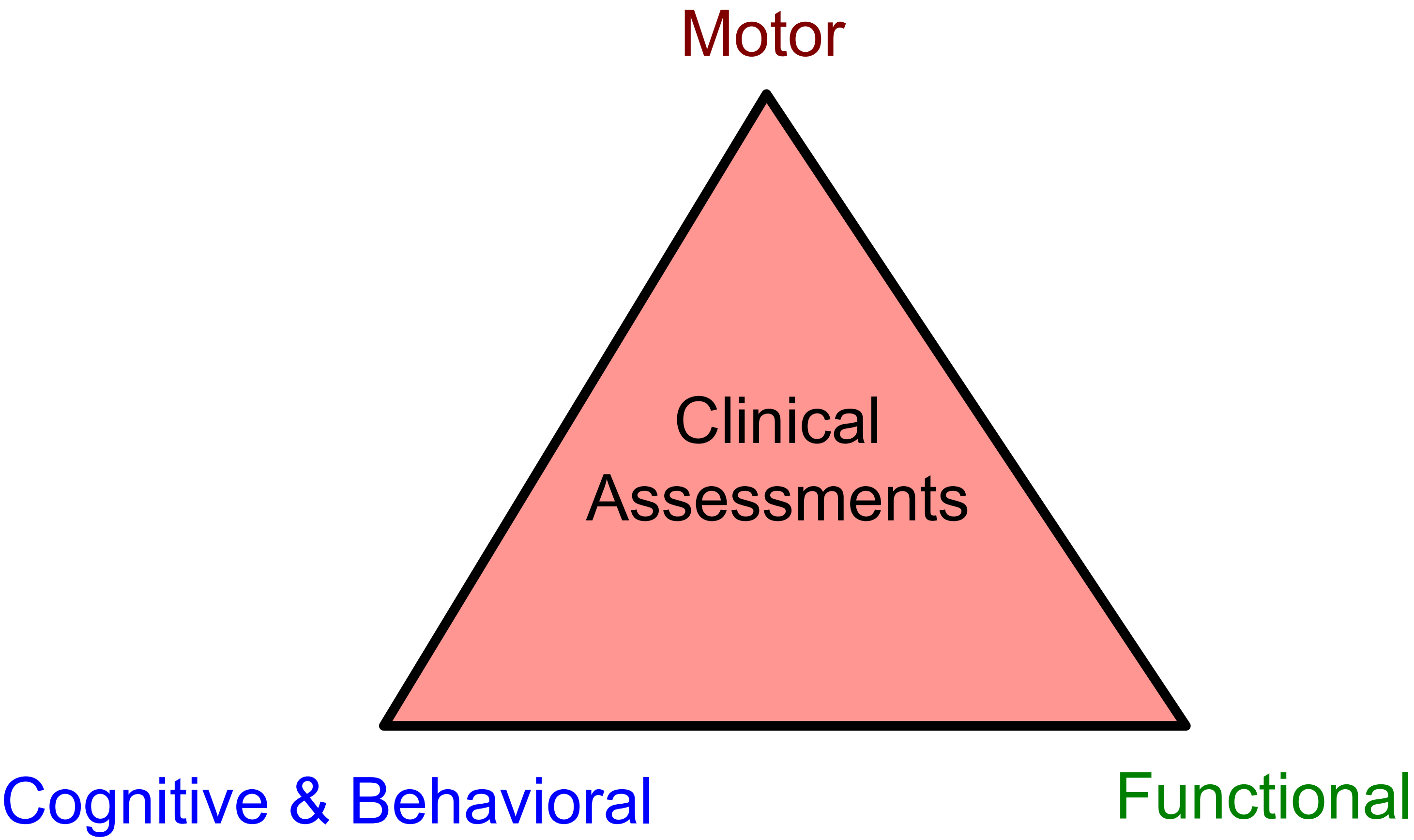
- Unsteady gait & involuntary movements (chorea)
- Slurred speech, difficulty in swallowing
- Forgetfulness & impaired judgment
- Personality changes, mood swings & depression
- Activities of daily living severely hampered

HD natural history



Ross et. al., Nature Reviews, 2004

Clinical assessments



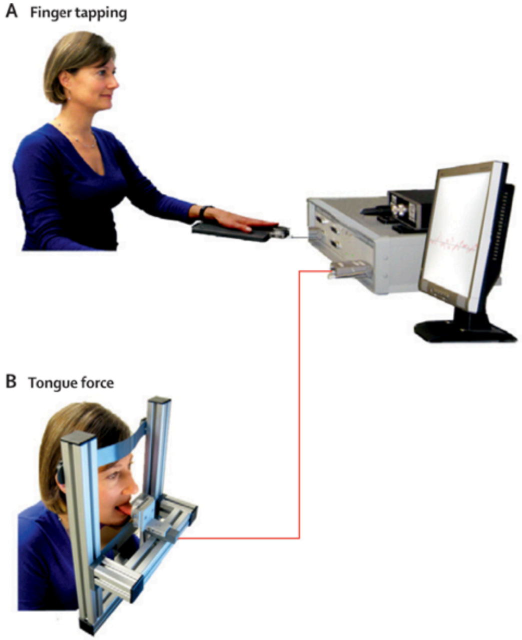
Motor

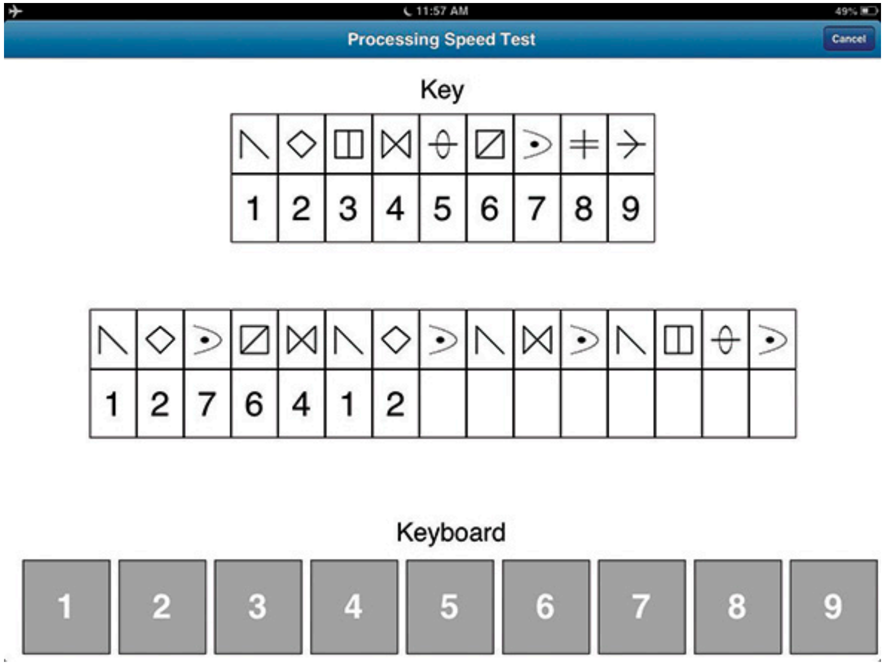
Clinical Assessments

Cognitive & Behavioral

Functional

Clinical assessment examples





Finger Tapping and Tongue Protrusion

Symbol Digits Modalities Test (SDMT)

source: Weir et. al., Lancet, 2011

source: clevelandclinic.org

Observational studies

TRACK HD

PREDICT-HD

An observational study of the earliest signs of Huntington disease

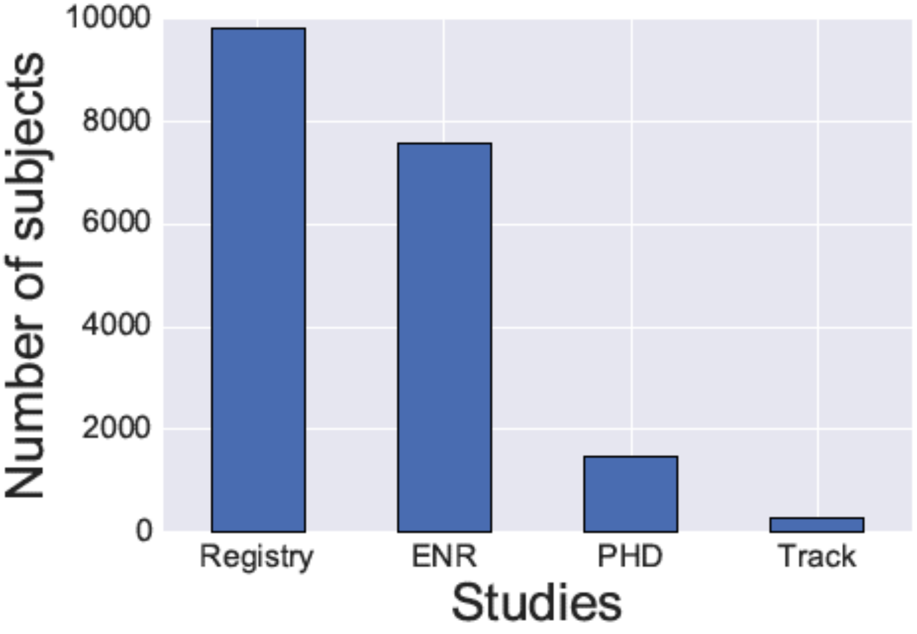
Enroll-HD

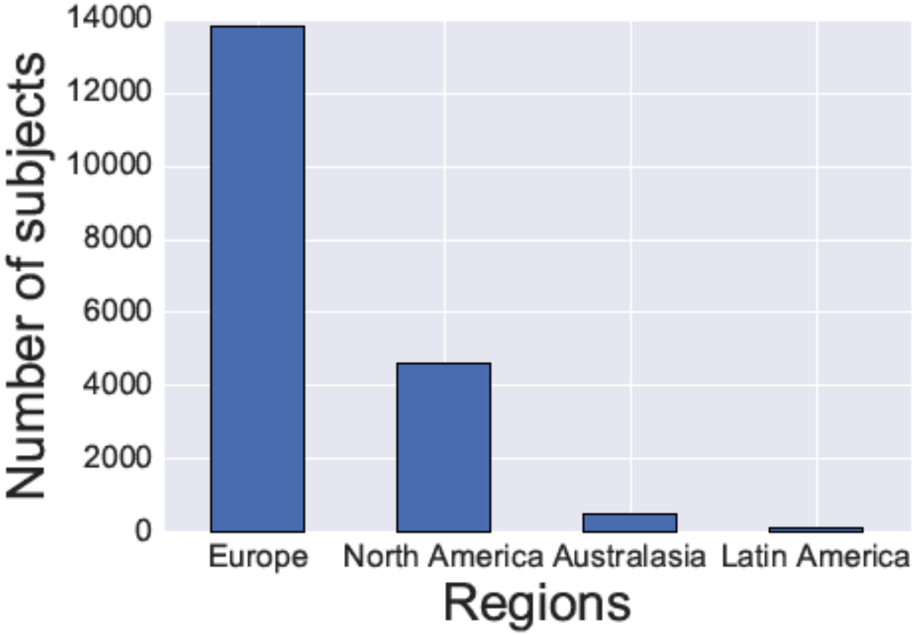
HD

EUROPEAN **HUNTINGTON'S DISEASE** NETWORK

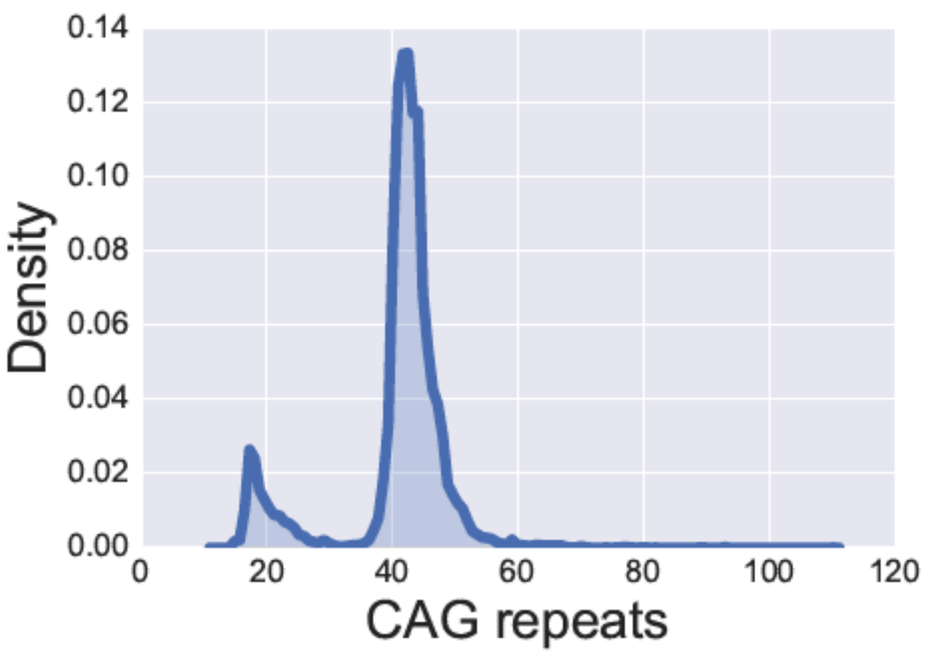
Combined dataset

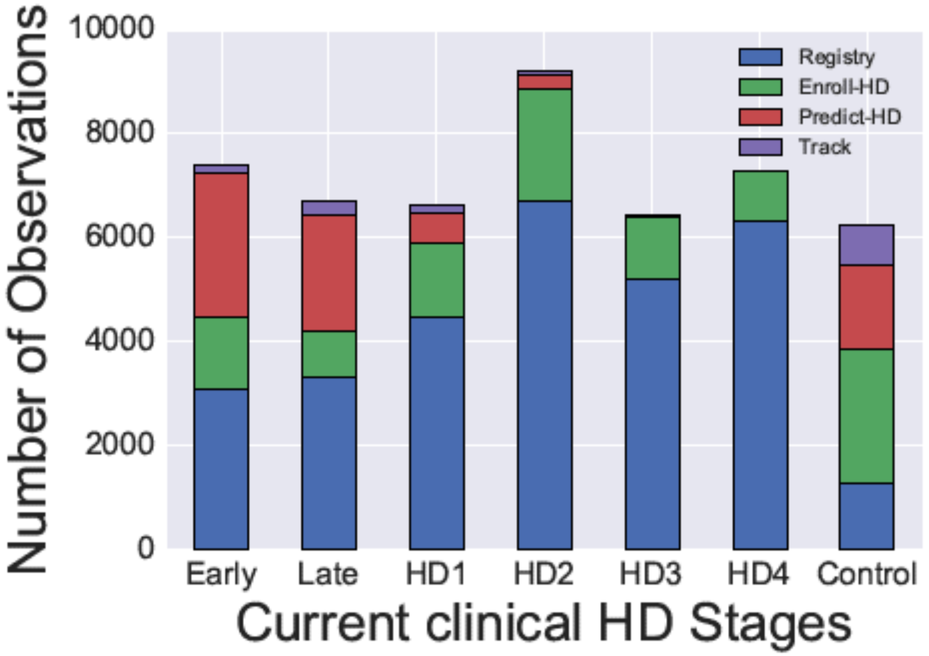
- Largest HD dataset studied to-date,
  - 16,553 HD subjects and 2,716 Controls
  - ~ 2000 Assessments





Combined dataset





Assessment selection

- All put together there are ~ 2000 assessments.
  - Not all are available in all studies or even between centers in a study
  - Not all are stable under repeated measurements
  - Some are more noisy than others
- We selected a subset based on *clinical feedback* and,
  - *Correlation* with surrogate measures of HD progression
  - Ability to *discriminate* between clinical HD stages and controls
  - 57 assessments

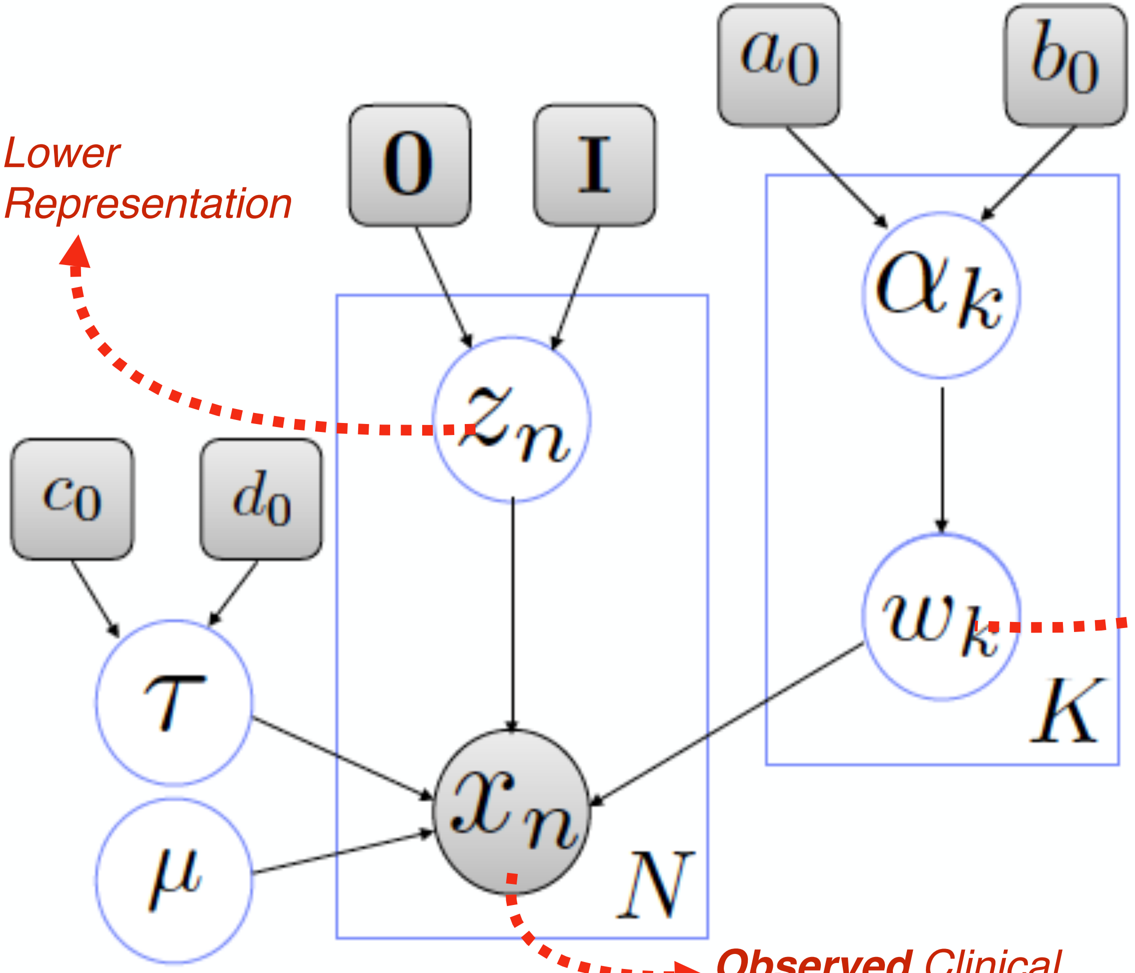
Lower dimensional embedding

- Assessments are high dimensional, but clearly not independent.

- We posit that there is a hidden lower dimensional structure underlying the assessments.

- Discovering this structure is challenging,
  - Noisy
  - High dimensional
  - Missing values

Robust Probabilistic PCA

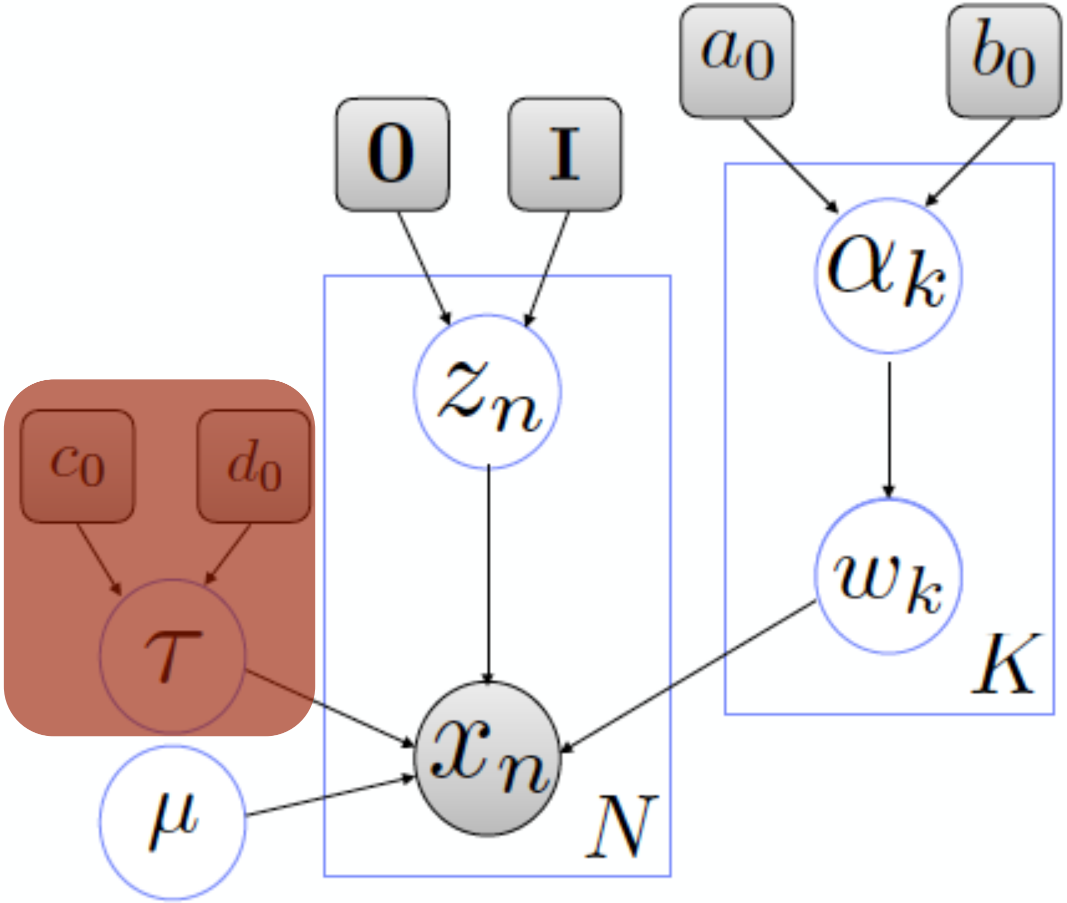


***Unobserved** Lower Dimensional Representation*

***Unobserved** Lower Dimensional Basis*

***Observed** Clinical Assessments*

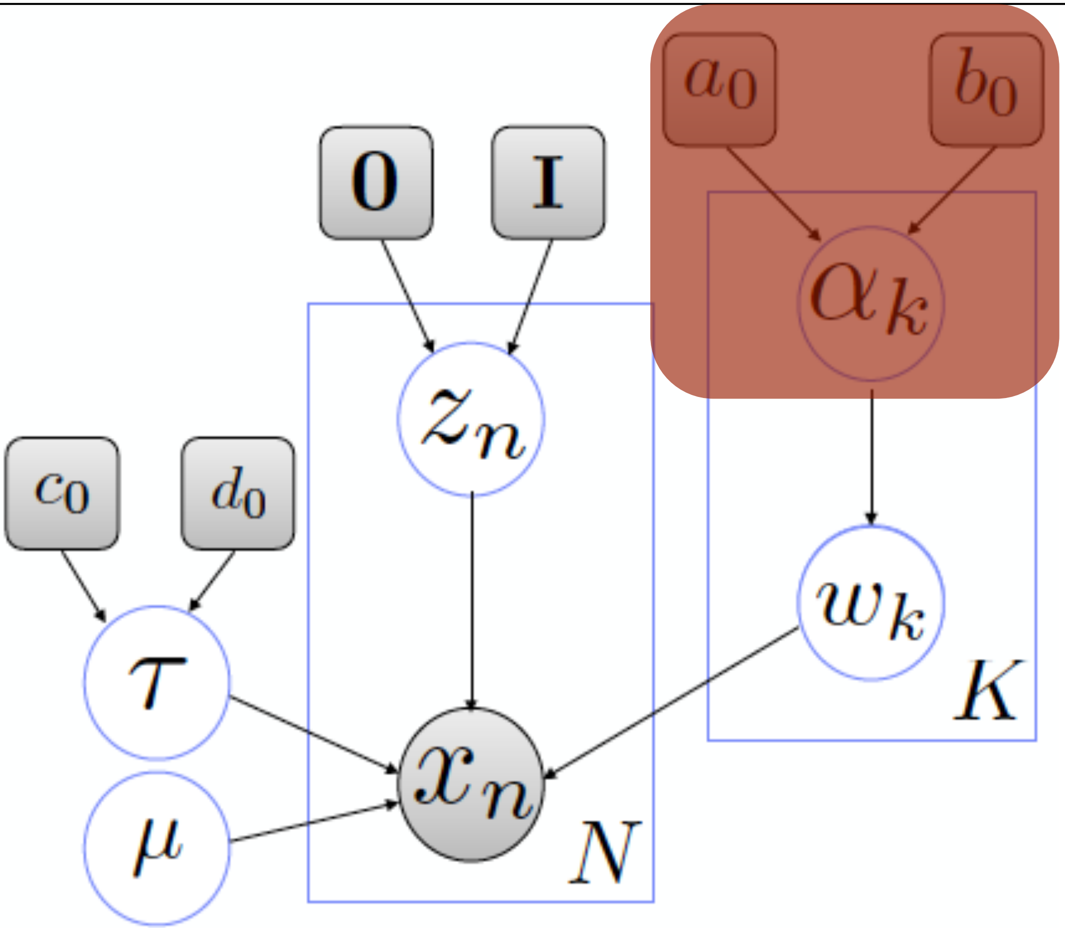
Robustness to outliers



Robust Likelihoods:

$$x_n \mid W, \mu, z_n, \tau \sim \mathcal{N}(W z_n + \mu, \tau^{-1}\mathbf{I})$$
$$\tau \sim \mathrm{Gamma}(c_0, d_0)$$

Automatic Model Selection



$$\alpha_k \sim \text{Gamma}(a_0, b_0)$$
$$w_k \mid \alpha_k \sim \mathcal{N}(0, \alpha_k^{-1}\mathbf{I})$$

Automatic Relevance Determination priors:
Sparsity promoting; turns off additional bases

Learning

- We learn the model by maximizing the marginal likelihood of the data.

$$p(\mathbf{x};\theta) \geq \mathcal{L}(W,\mathbf{z};\theta)$$

↑

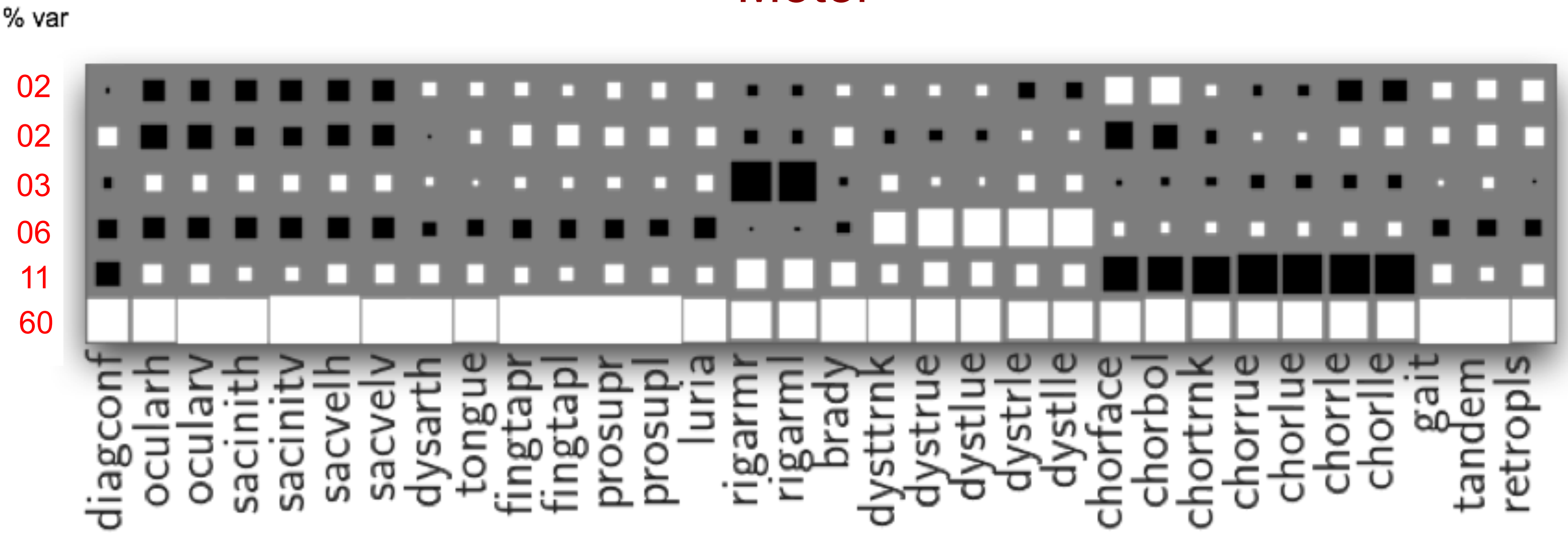
- This is intractable. We maximize a *lower bound* to the marginal likelihood (*variational inference*)

- Generalization of EM; involves cycling over fixed point updates.

Inferred Embedding and Bases

Discovered Bases E[W | x]

Motor



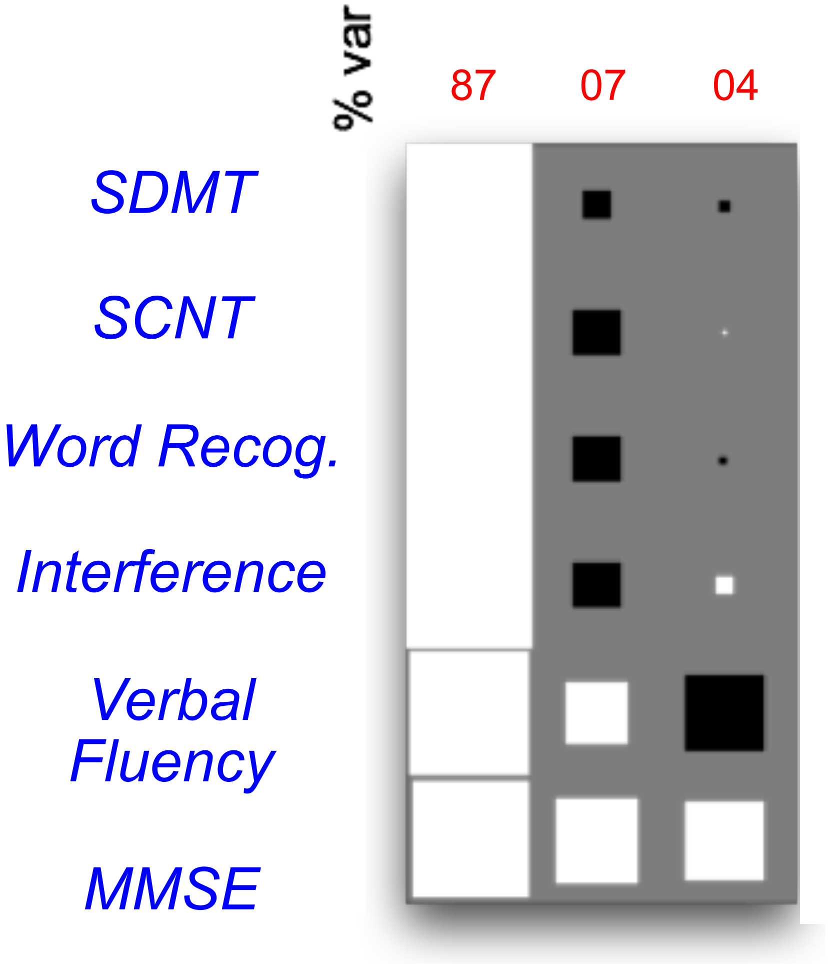
Occulo-motor

Dystonia

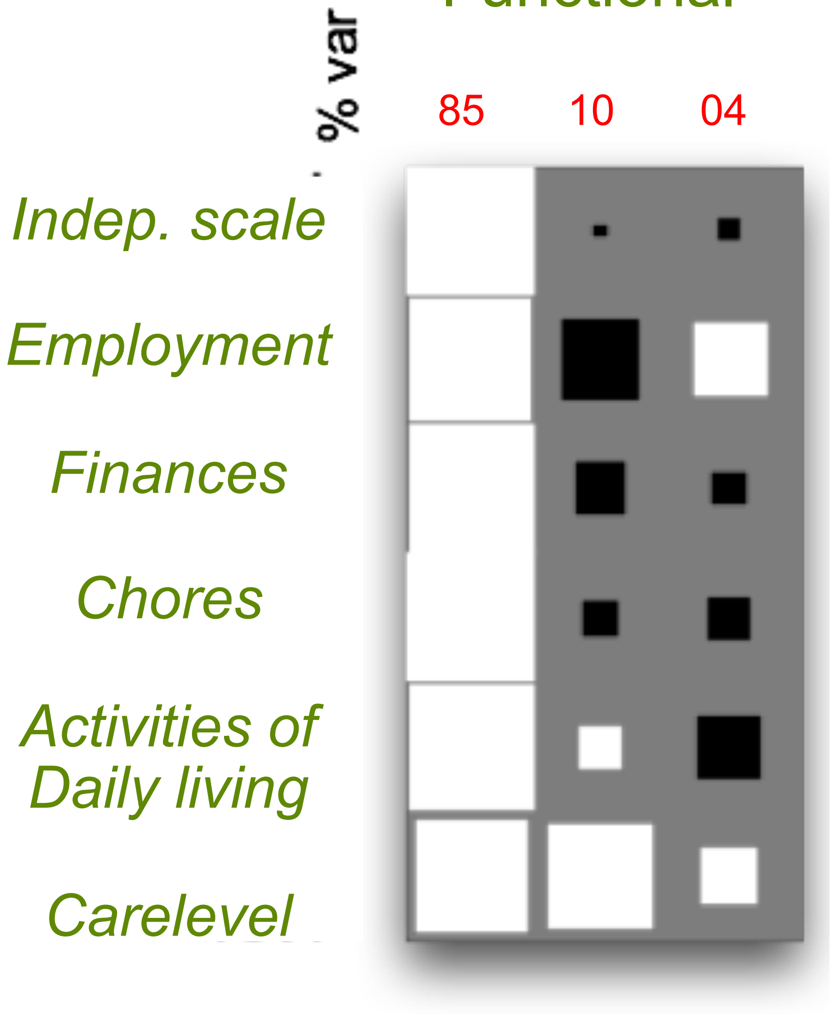
Chorea

Discovered Bases E[W | x]

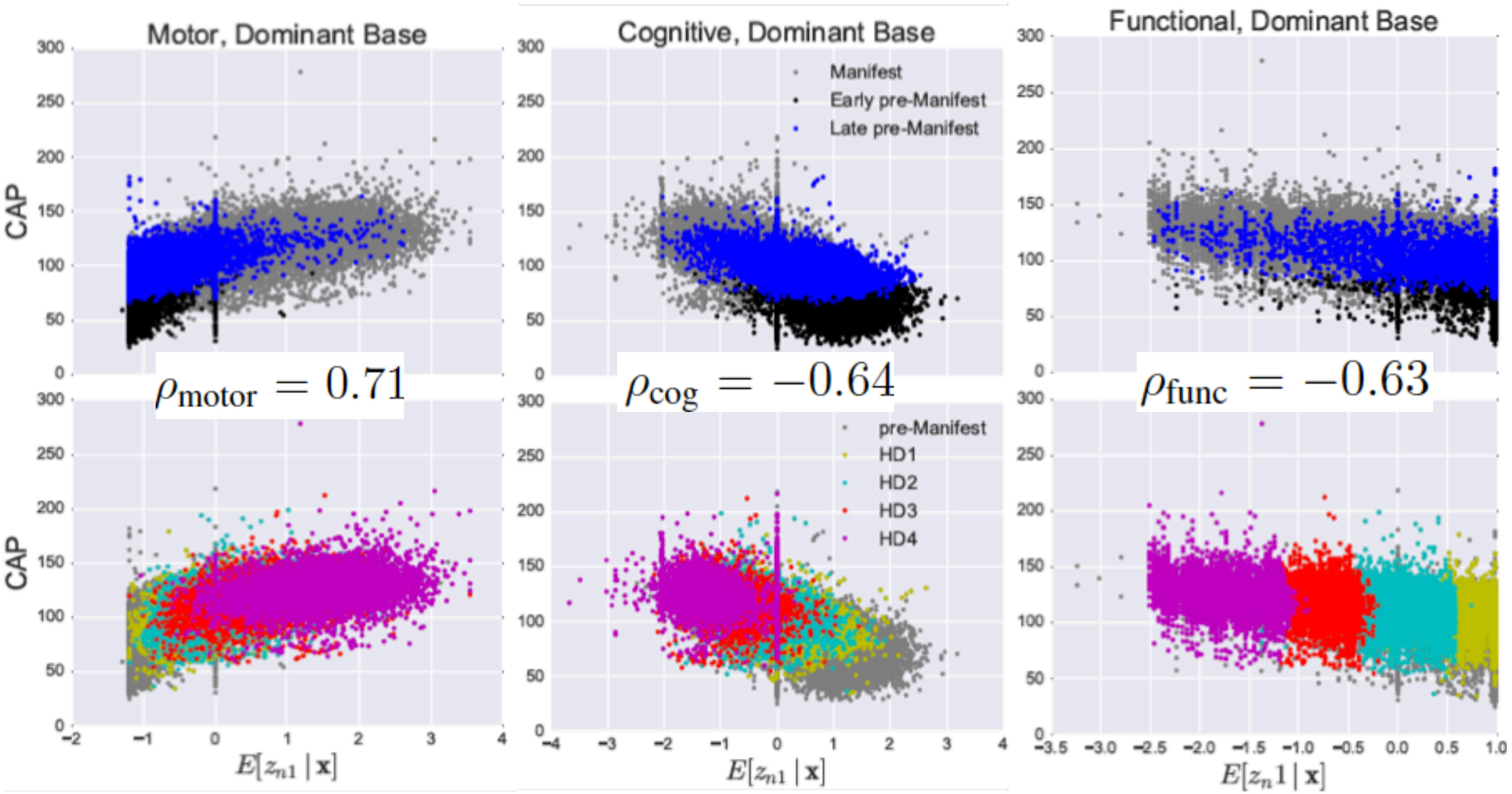
Cognitive



Functional



HD severity vs embedding



CAP = cag repeats x age. A coarse measure of disease progression.
Higher CAP → More advanced HD

Summary

- Curated the largest observational HD dataset to date

- Robust probabilistic latent variable analysis
  - Generates lower dimensional embeddings that track well with surrogate measures of HD progression.
  - Discovers interesting latent structure
    - Dominant base tracks well with CAP, subsequent bases don't.
    - Non-negligible unexplained variance.
    - Behavior assessments appear less reliable.

- Follow up preliminary work using these embeddings has resulted in exciting new data driven HD stages.

Questions

ghoshso@us.ibm.com