Hierarchical Latent Dictionaries for Models of Brain Activation

Alona Fyshe*, Emily Fox, David Dunson and Tom Mitchell

Carnegie Mellon CNBC Center for the Neural Basis of Cognition

ML MACHINE LEARNING DEPARTMENT

Wharton University of Pennsylvania

Duke University

afyshe@cs.cmu.edu

Multi-sensor time series

- Multiple noisy recordings of the same process
  - Many sensors (high dimension)
  - Changing sensor correlation

Multi-sensor time series

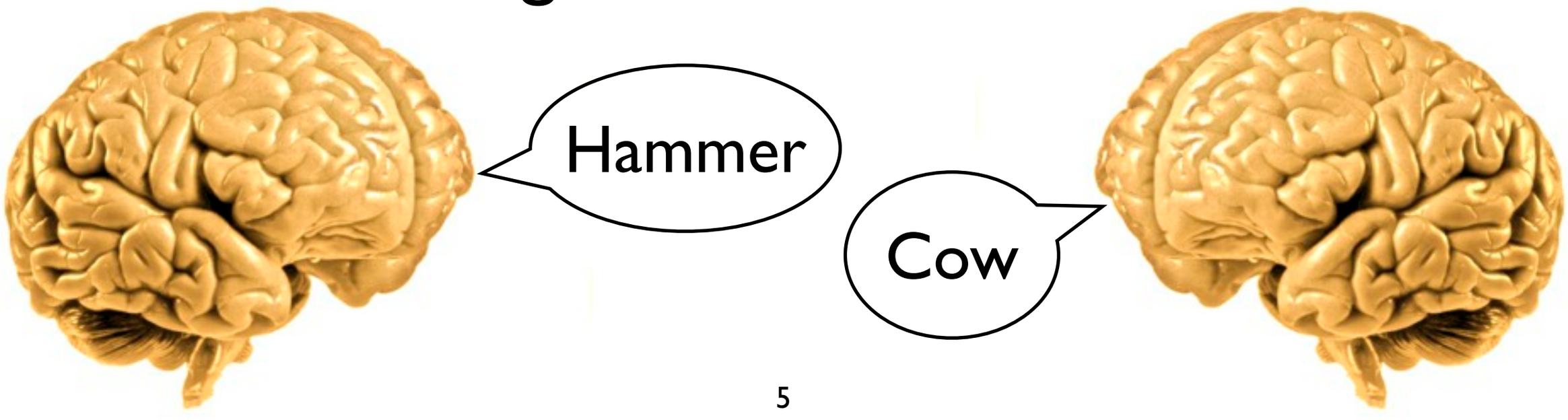
- Seismic data
- Weather sensors
- Physiological recordings
  - EKG, Ecog, MEG

Understanding Cognition

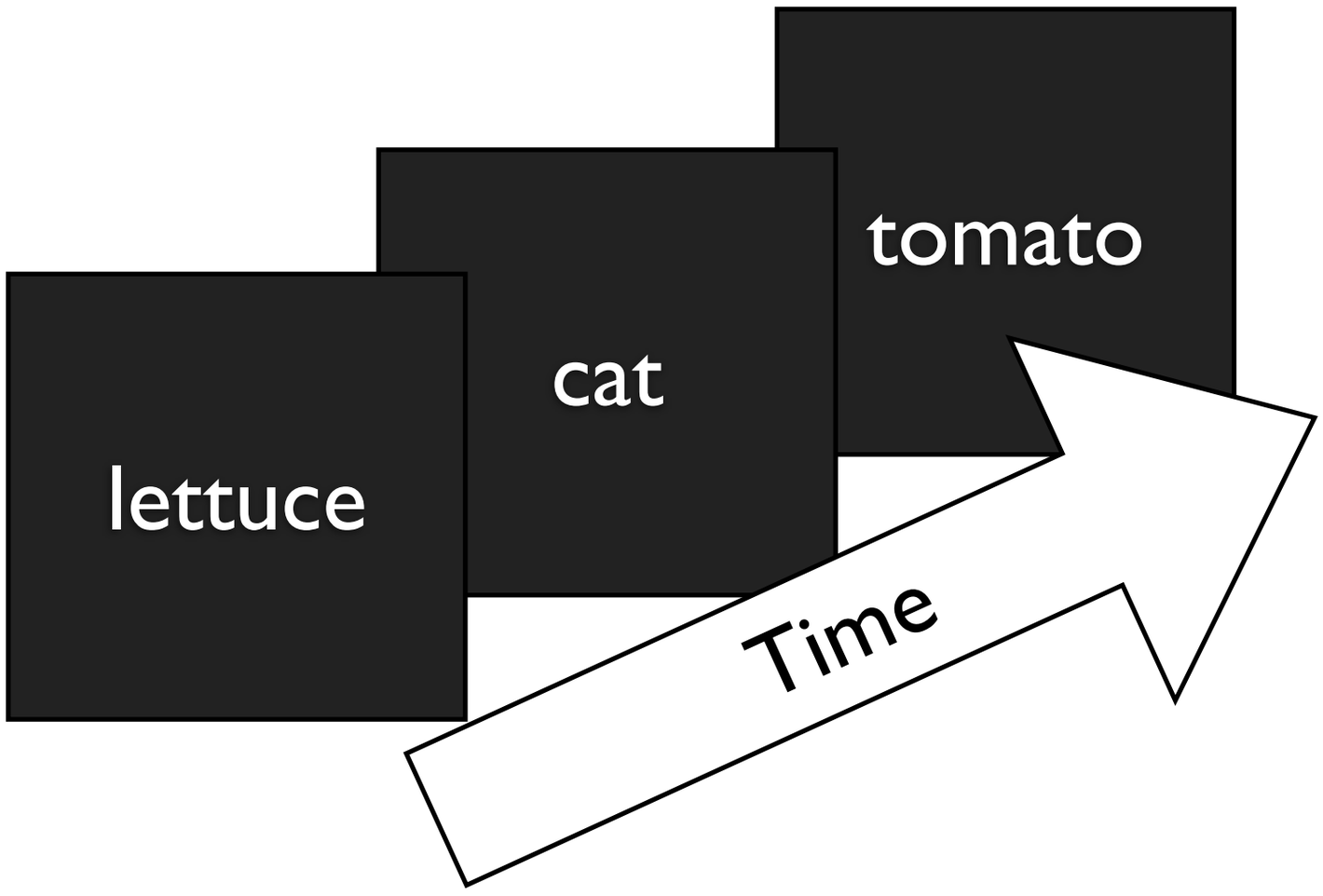
- How does the brain code concepts?
    - e.g. animals, food...
- How does brain activity change over time?

Understanding Cognition

- Show words to subjects
- Record & model brain activity
- Distinguish between word categories
  - e.g. animals vs tools?



Understanding Cognition



Challenging Data

- High noise & high dimension
- Need to predict with single (noisy) trials

Data Collection

- MEG recordings
- 2 subjects
- 20 words

Data Collection

- 4 word categories

Animals

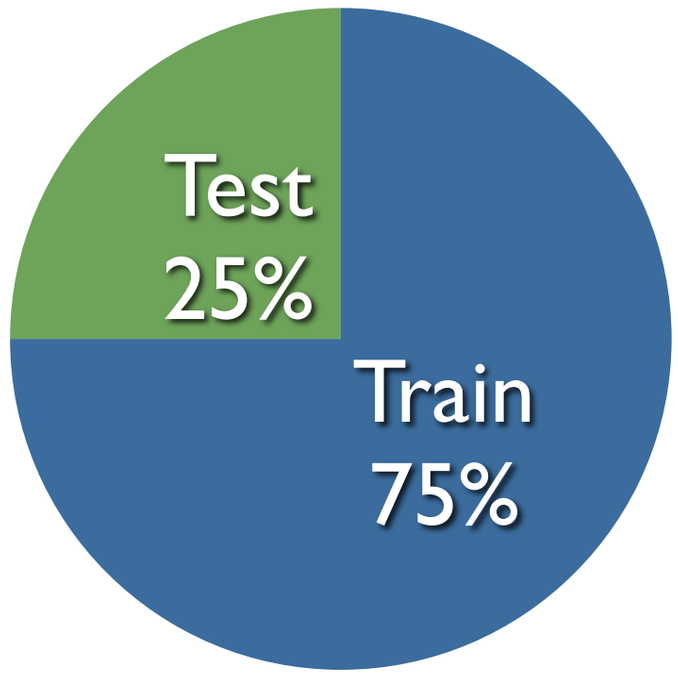
Food

Tools

Buildings

Data Collection

- 20 repetitions per word (400 total)
  - 15 train/word (300 total)
  - 5 test/word (100 total)

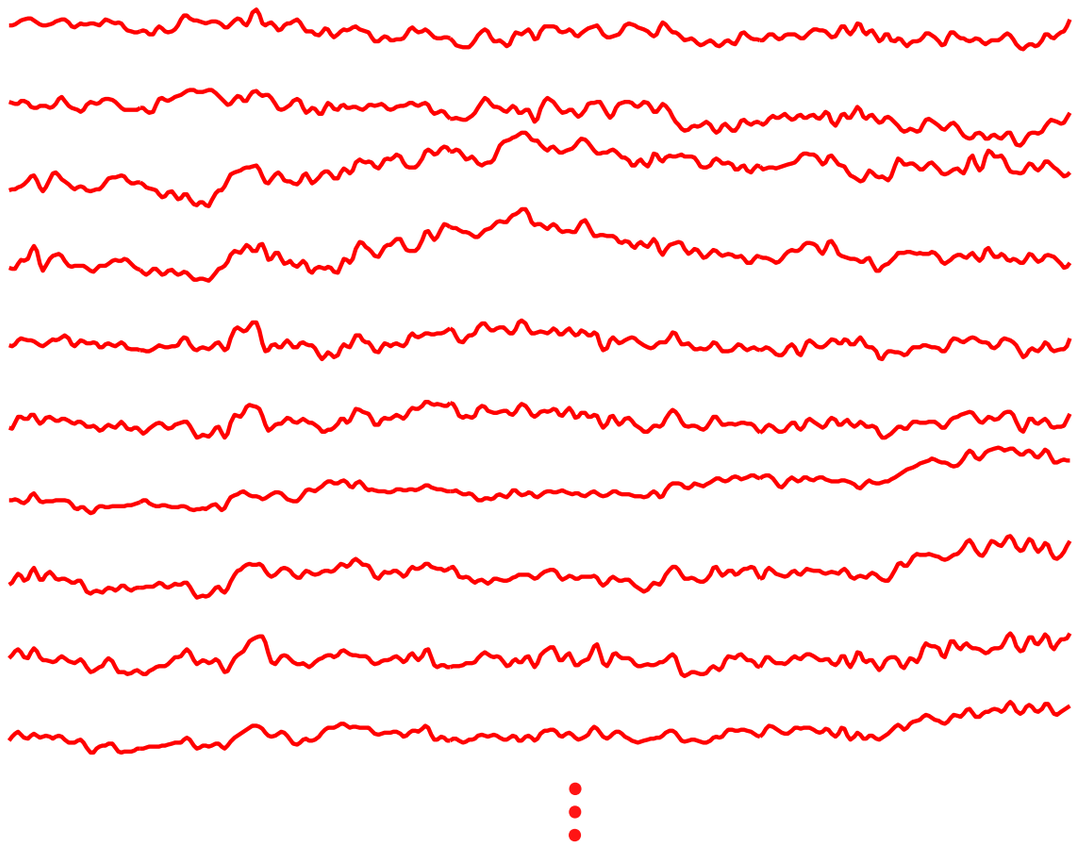


Brain Imaging

- fMRI
  - Measures blood oxygenation
  - Great spatial resolution
  - Poor time resolution
- MEG
  - Measures neuron's magnetic field
  - Great temporal resolution
  - Reduced spatial resolution

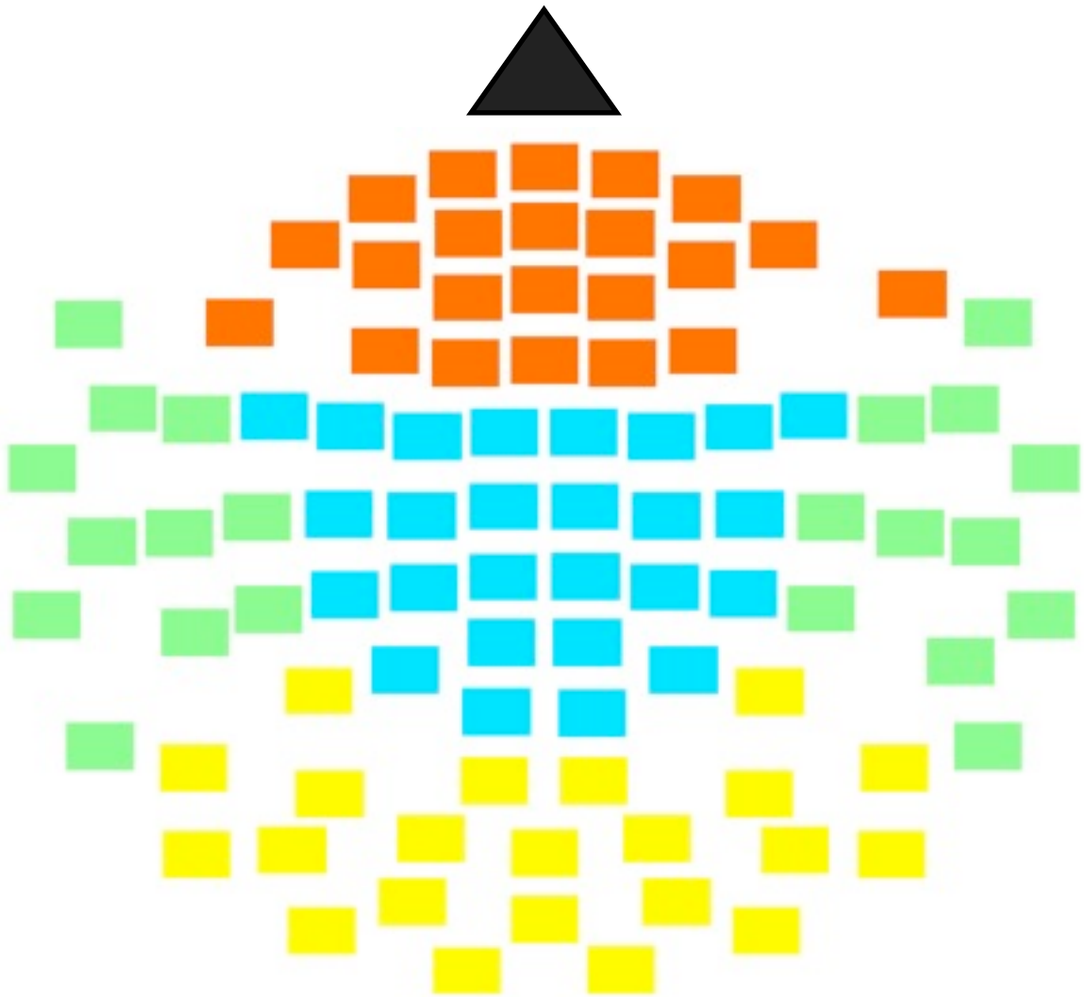
MEG machine

- 102 sensors in a helmet



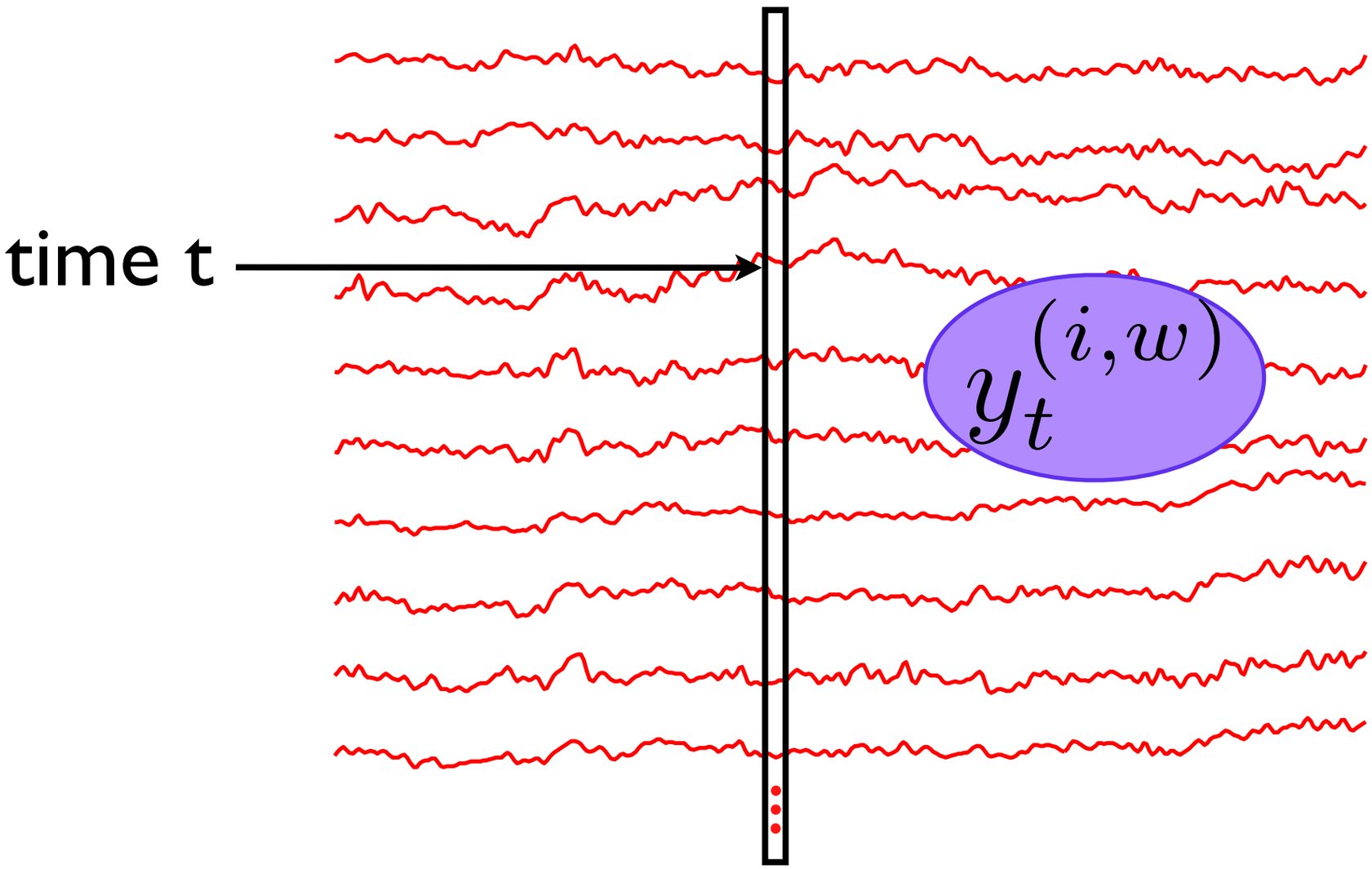
12

MEG Machine

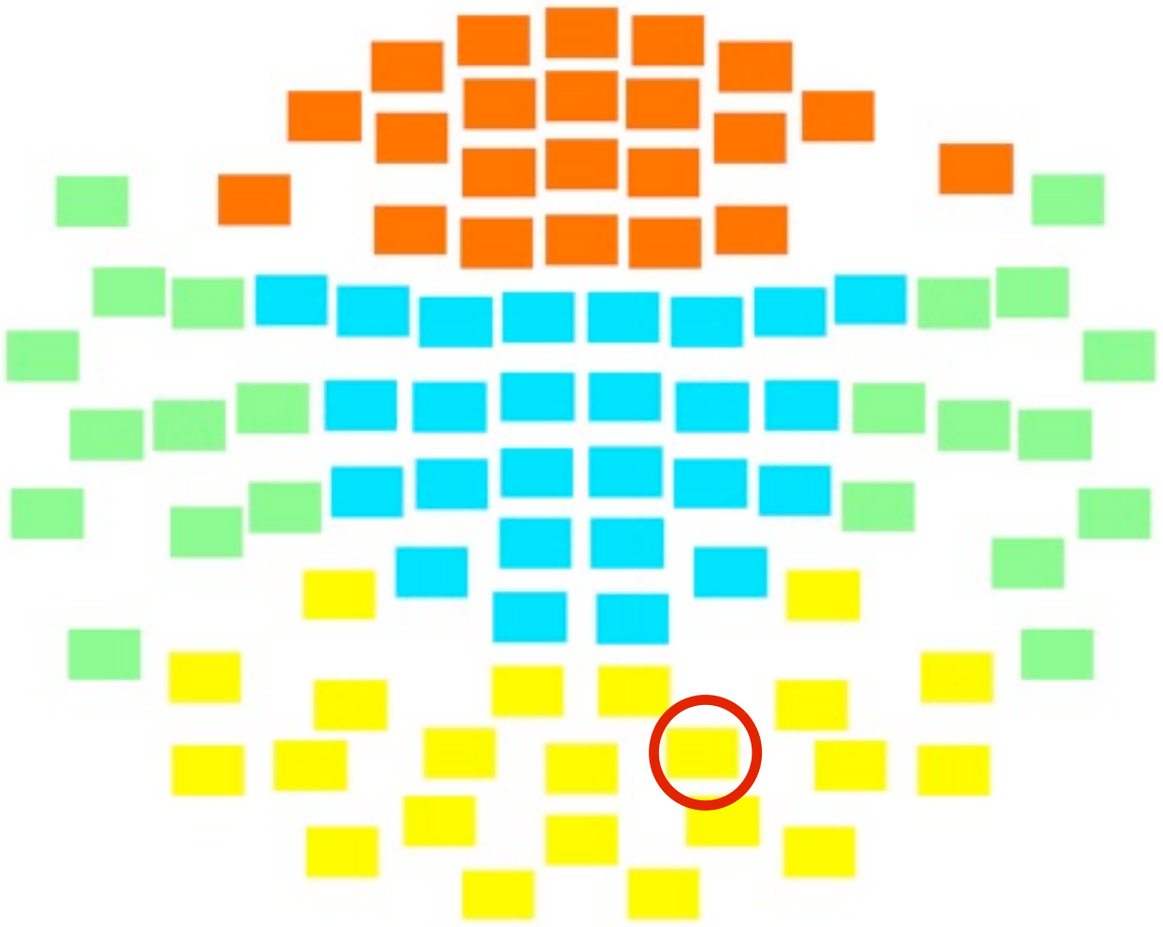


Terminology

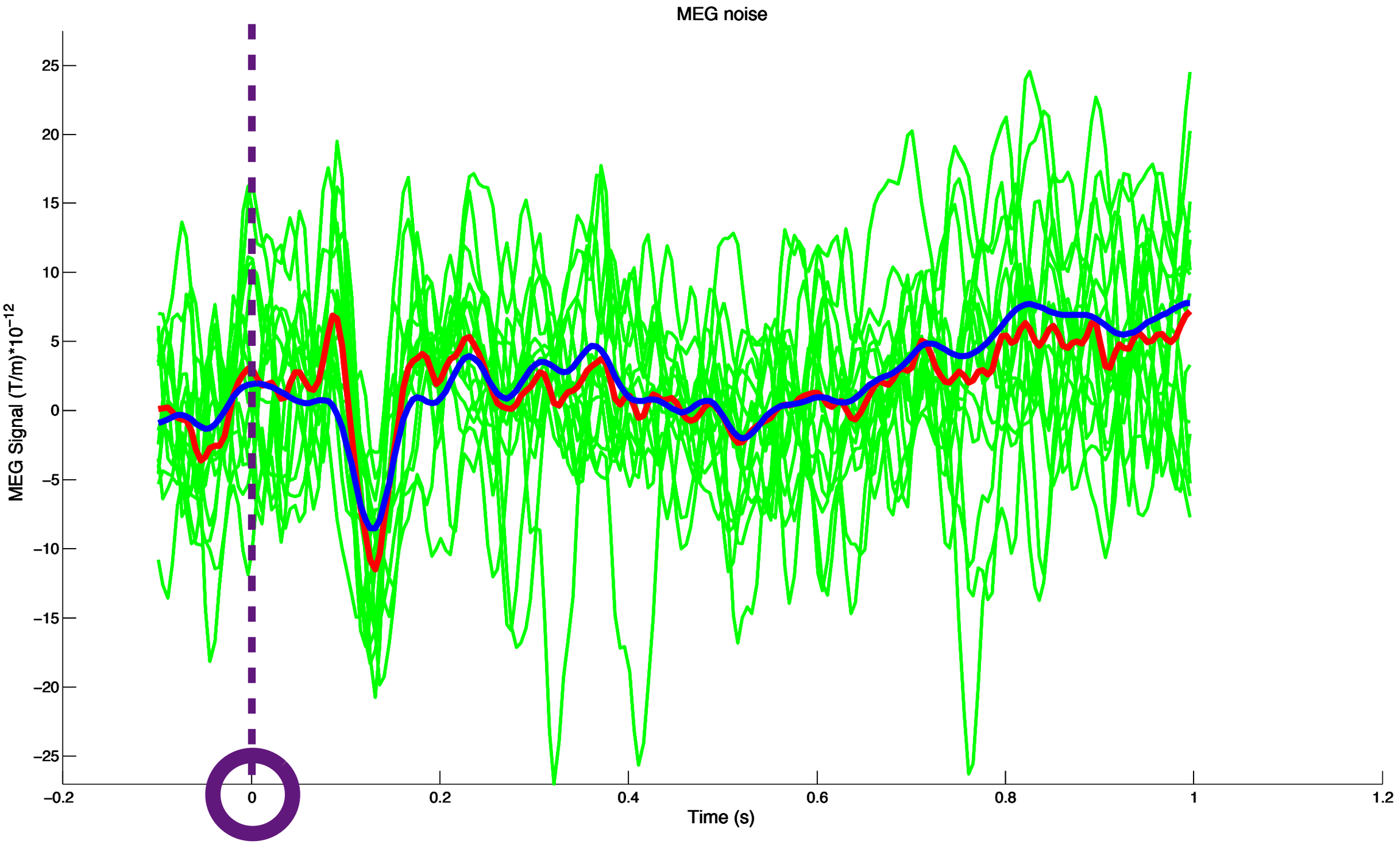
MEG recording for word w, trial i



MEG Noise



MEG Noise


MEG noise
MEG Signal (T/m)*10^-12
25
20
15
10
5
0
−5
−10
−15
−20
−25
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
Time (s)

Challenges

1) Noise

2) High dimension

3) Covariance changes with time

- Functional connectivity

4) Trial-to-trial variation

Challenge 1:
MEG data is noisy

MEG data is noisy

- Traditionally:
  - Average repetitions (MLE)
    - assume channels are independent
- Can we do better than average?

MEG data is noisy

Observation: underlying neural signal is smooth

Underlying signal is smooth

- Model with a Gaussian process (GP)
  - f ~ GP(m,c)
    - m: mean function
    - c: kernel function (covariance)
      - squared exponential

$$c_i(t, t') = d_i \exp(-\kappa||t - t'||_2^2)$$

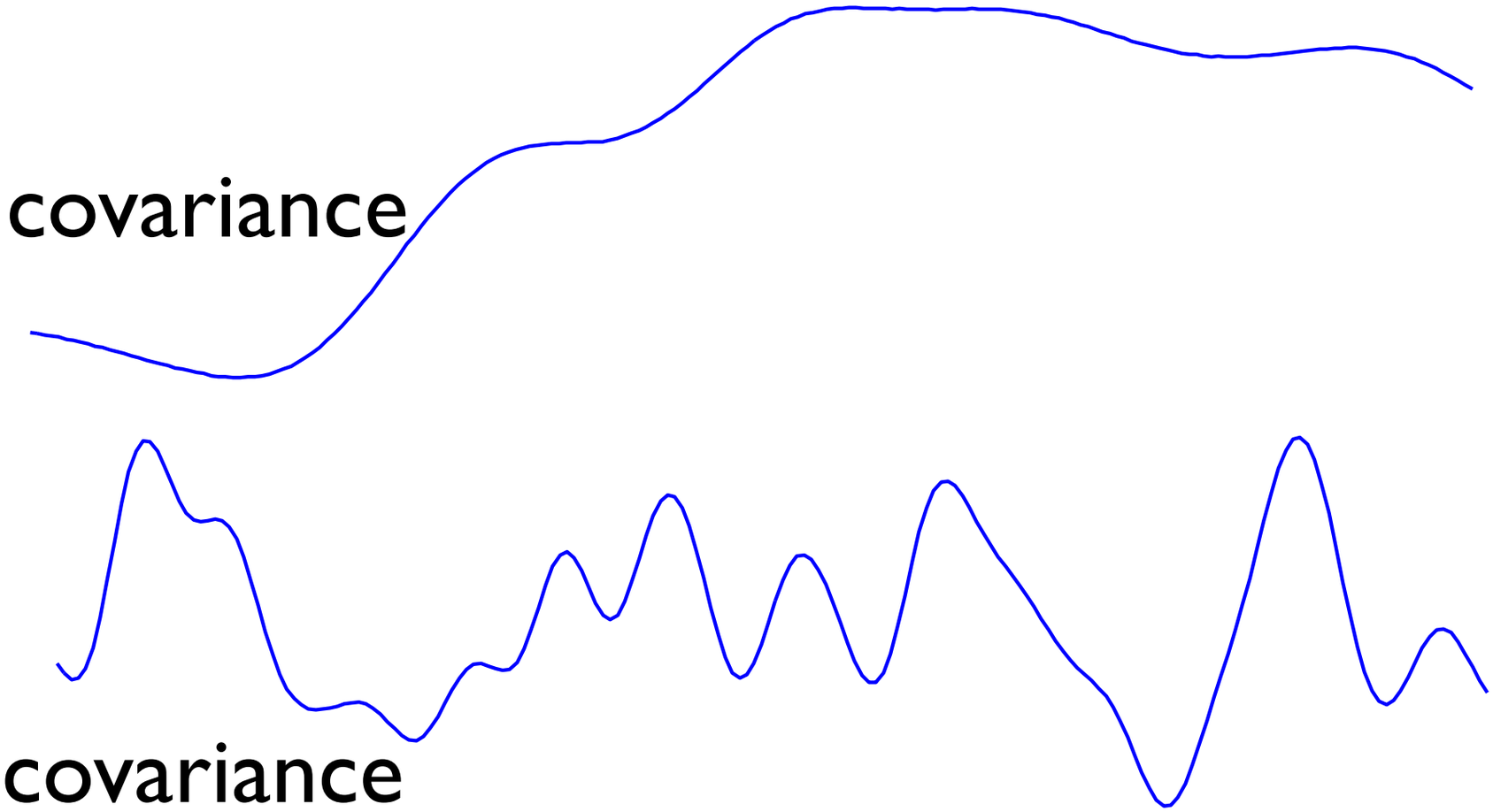
Underlying signal is smooth

- $p(f(t_1), \ldots, f(t_n)) \sim N_n(\mu, K)$
  - $\mu = [m(t_1),\ldots,m(t_n)]$
  - $K_{ij} = c(t_i,t_j)$

Gaussian Processes

c: kernel function (covariance)

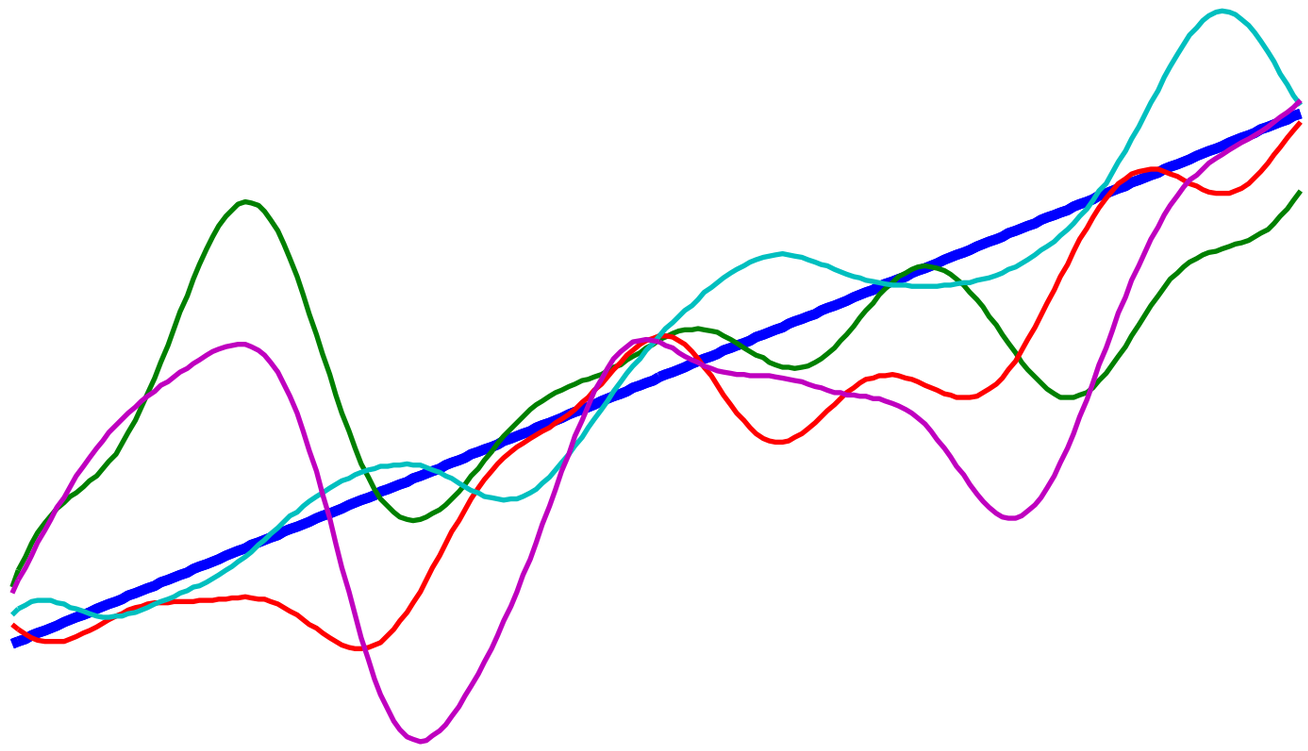
High covariance



Low covariance

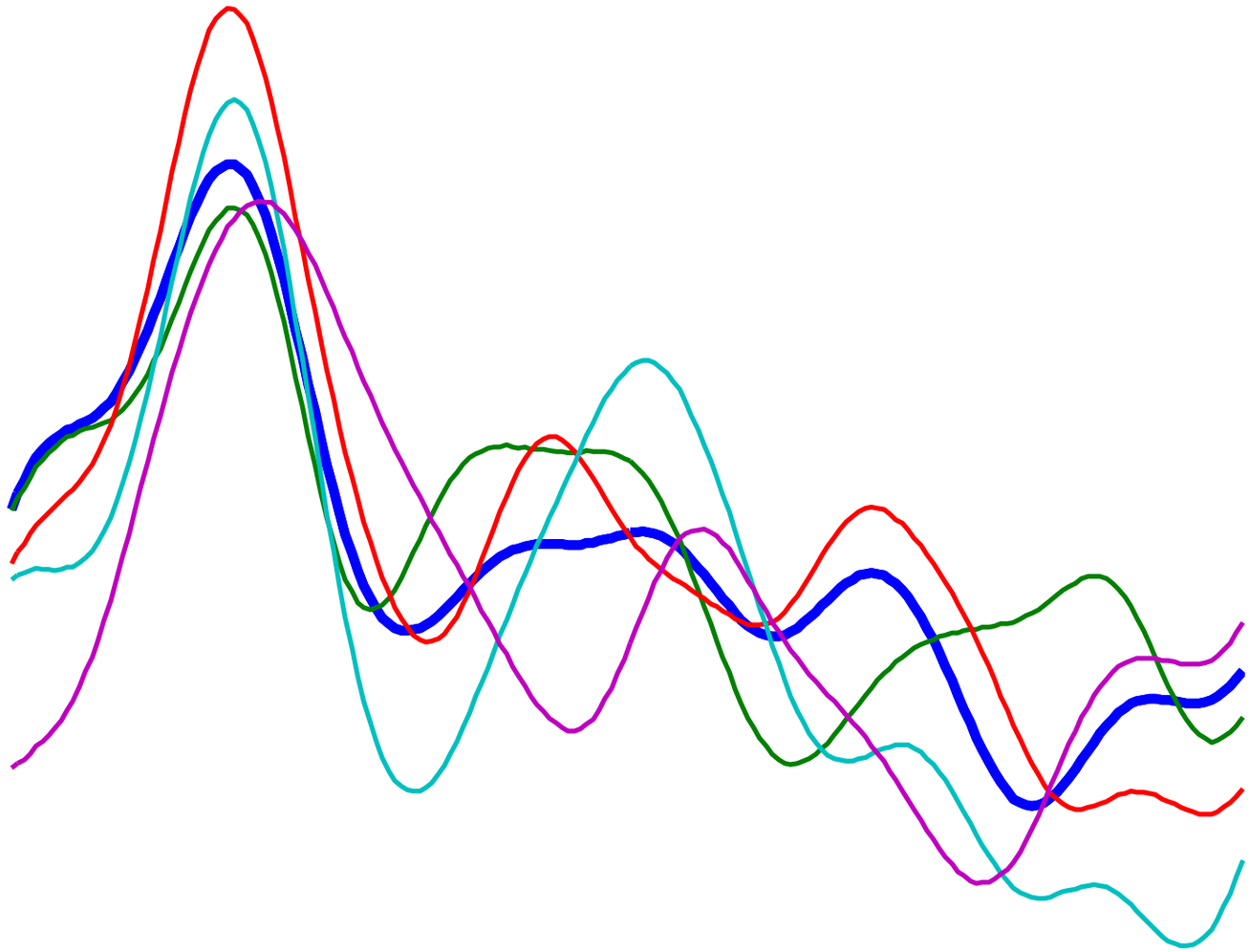
Gaussian Processes

m: mean function

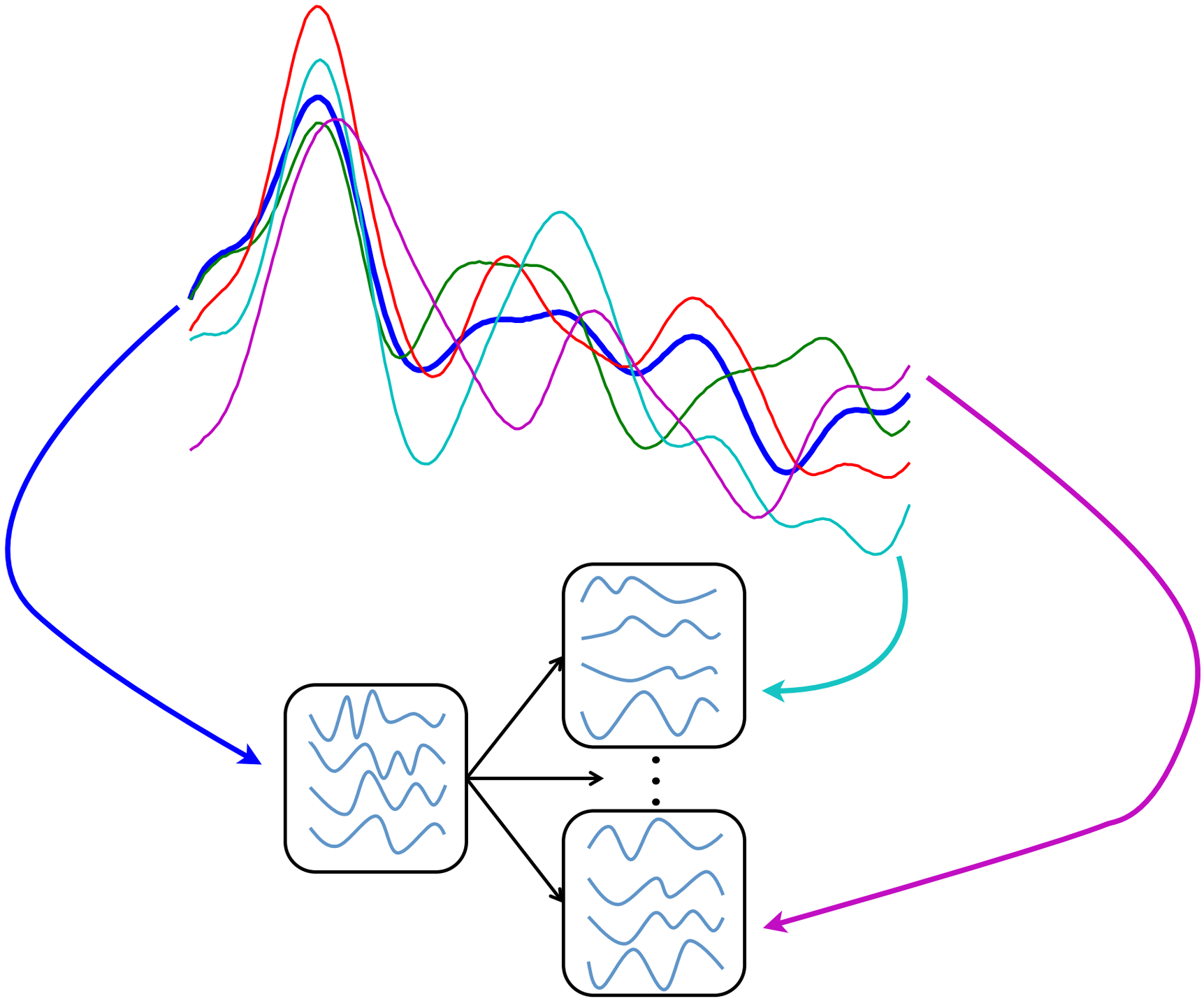


Gaussian Processes

m: mean function



Gaussian Processes



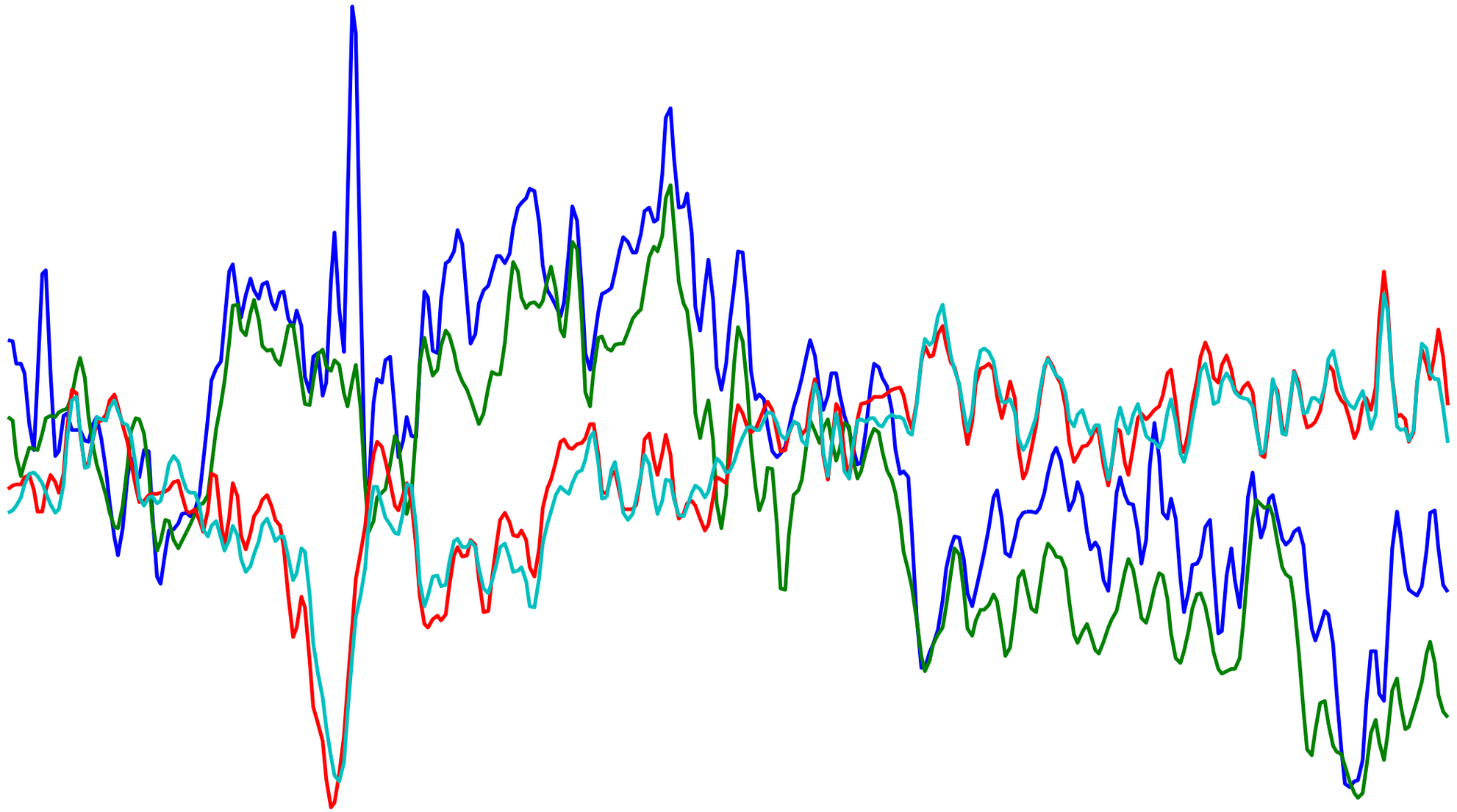
Challenge 2: Data is of high dimension

Data is of high dimension

Observation: Sensors are redundant

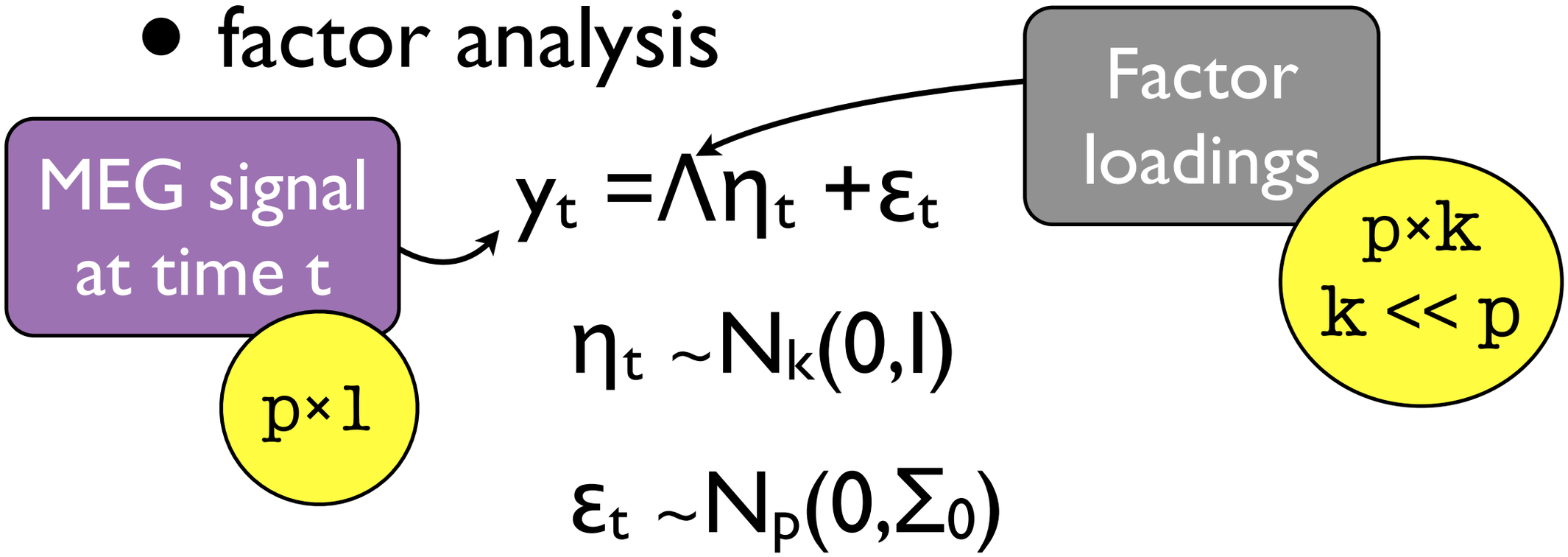
- spatially localized correlation
- long-range correlation

Redundant Sensors



Redundant Sensors

- Model in a lower dimensional space
  - factor analysis



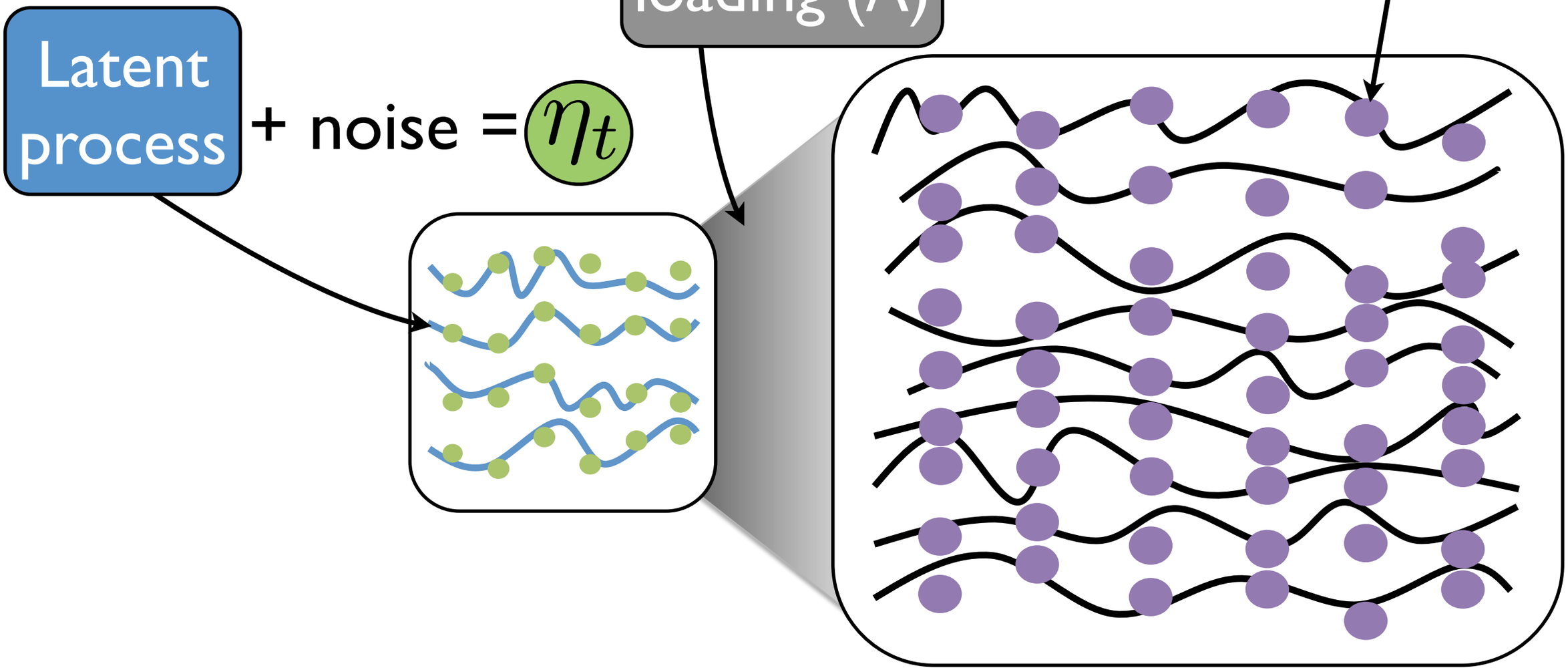
Redundant Sensors

$$y_t = \Lambda \eta_t + \epsilon_t$$

Observed MEG signal

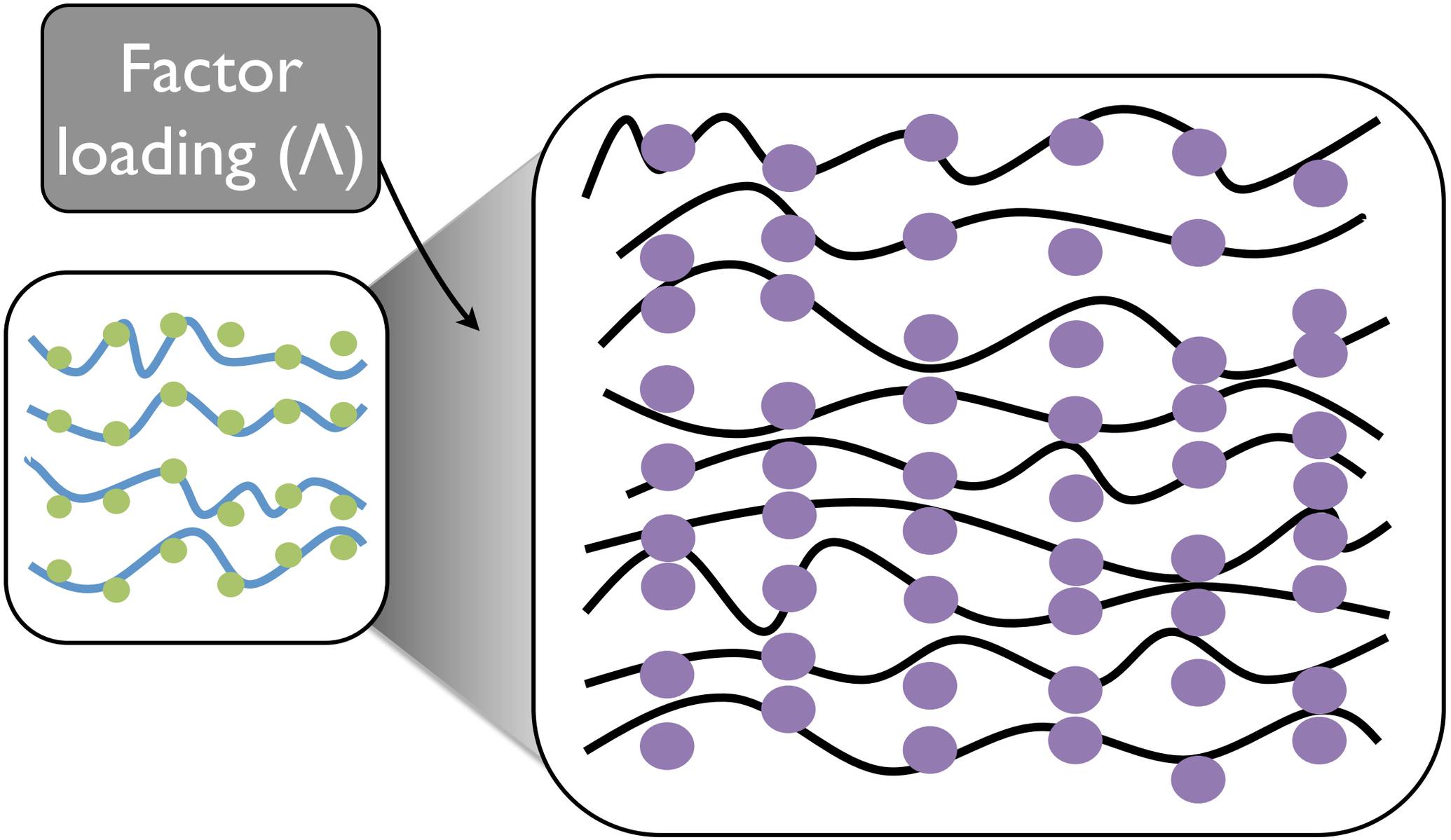
Factor loading (Λ)

Latent process + noise = η_t

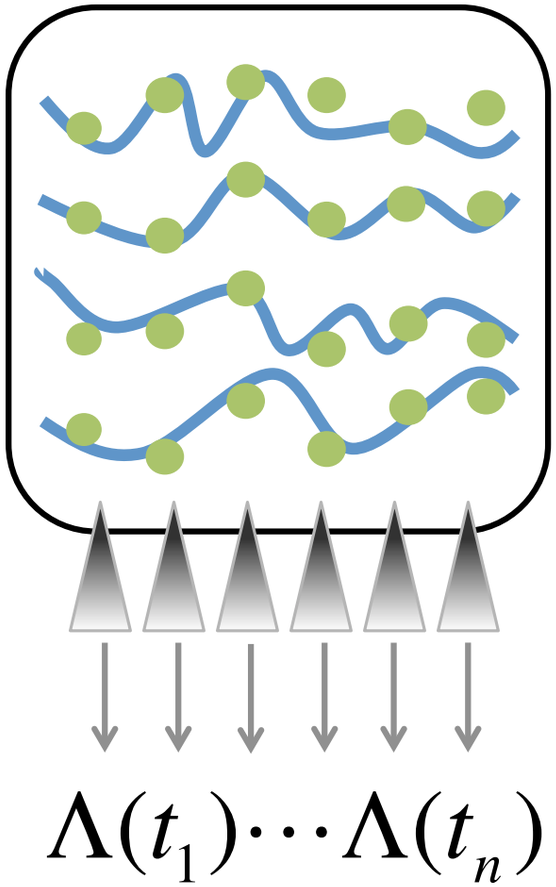


Challenge 3: Coordination changes with time (Heteroskedasticity)

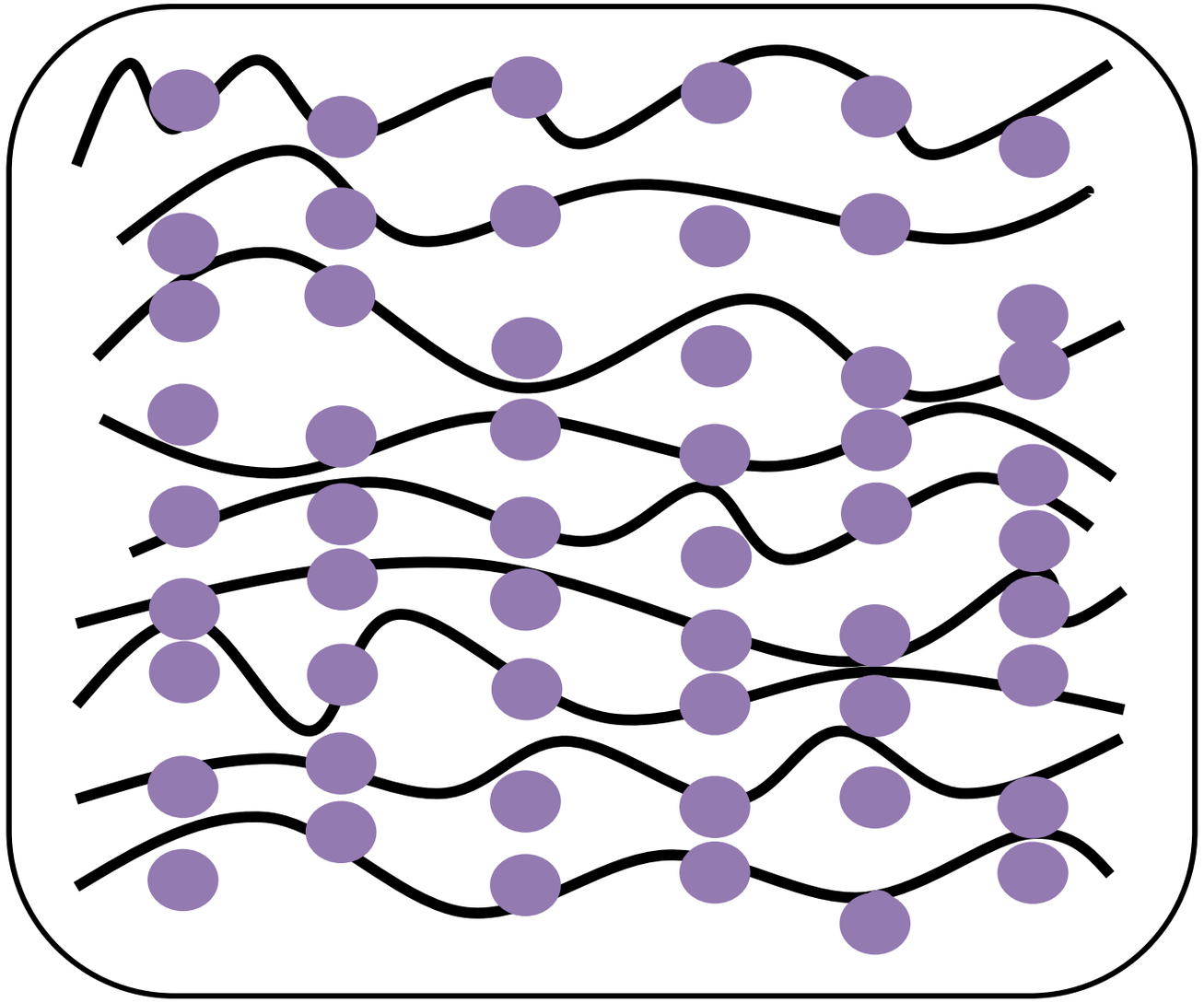
Coordination changes with time



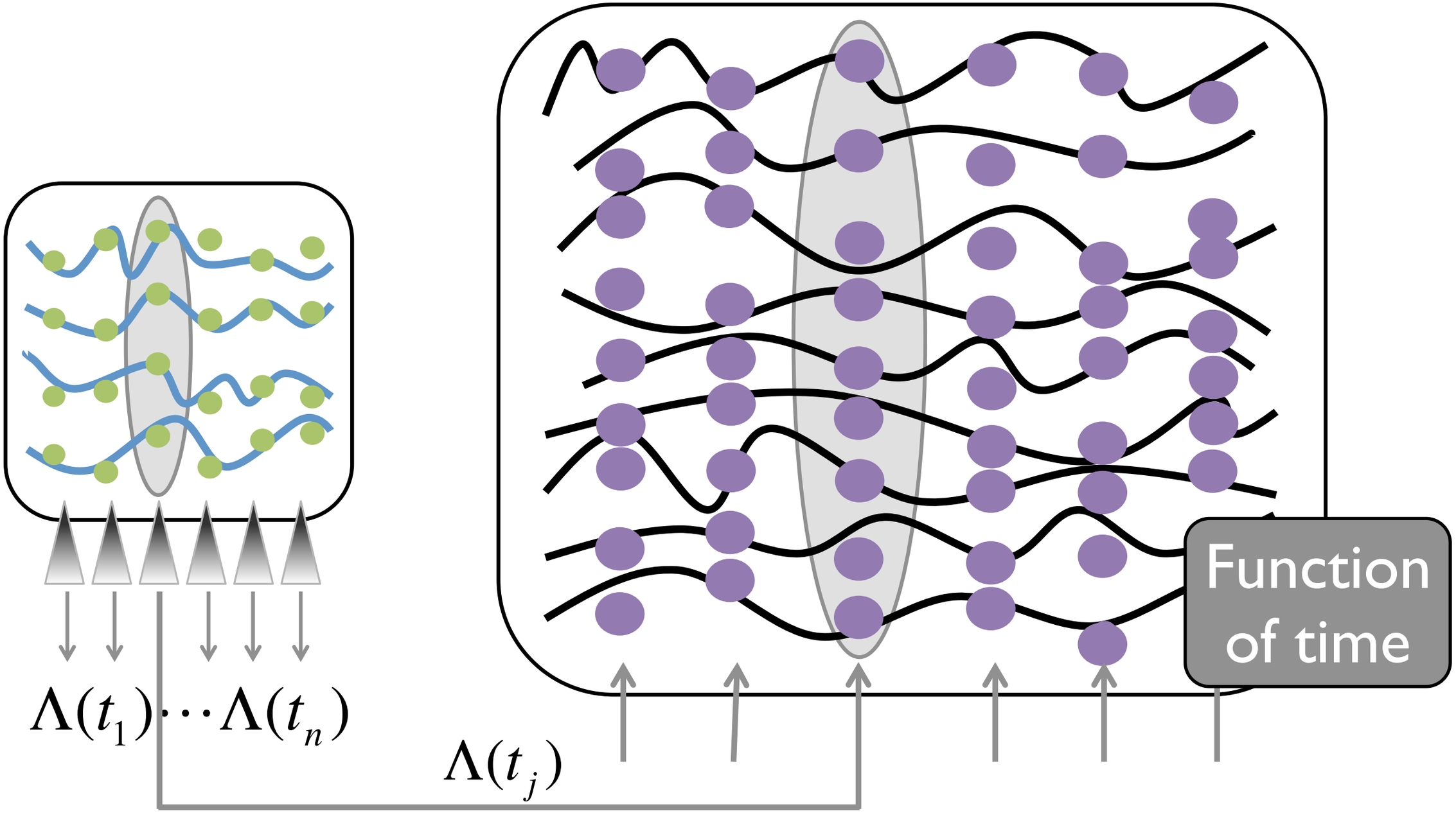
Coordination changes with time



$$\Lambda(t_1)\cdots\Lambda(t_n)$$



Coordination changes with time



Coordination changes with time

MEG signal

$$y_t^{(i,w)} = \Lambda^{(w)}(t)\eta_t^{(i,w)} + \epsilon_t^{(i,w)}, \quad \epsilon_t^{(i,w)} \sim N_p(0, \Sigma_0^{(w)})$$

$$\eta_t^{(i,w)} = \psi^{(i,w)}(t) + \nu_t^{(i,w)}, \quad \nu_t^{(i,w)} \sim N_k(0, I_k)$$

$$\Lambda^{(w)}(t) = \Theta^{(w)}\xi^{(w)}(t)$$

Latent mean

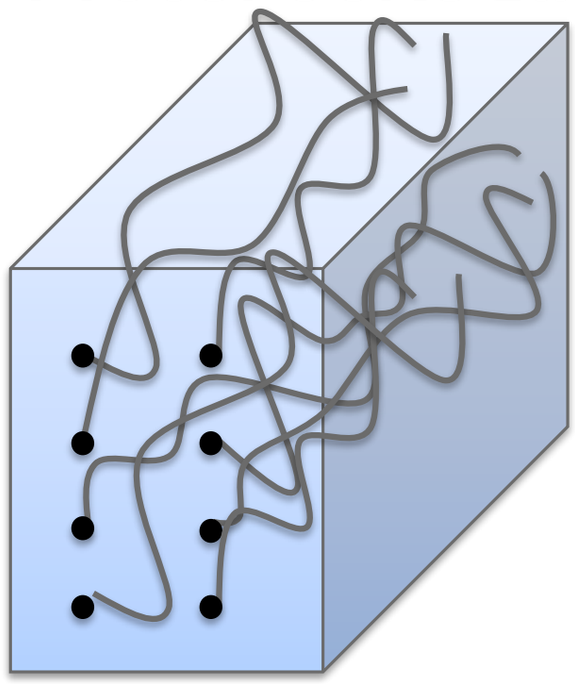
Factor loadings

Coordination changes with time

$$\Sigma^{(w)}(t) = \Lambda^{(w)}(t)\Lambda^{(w)\prime}(t)$$

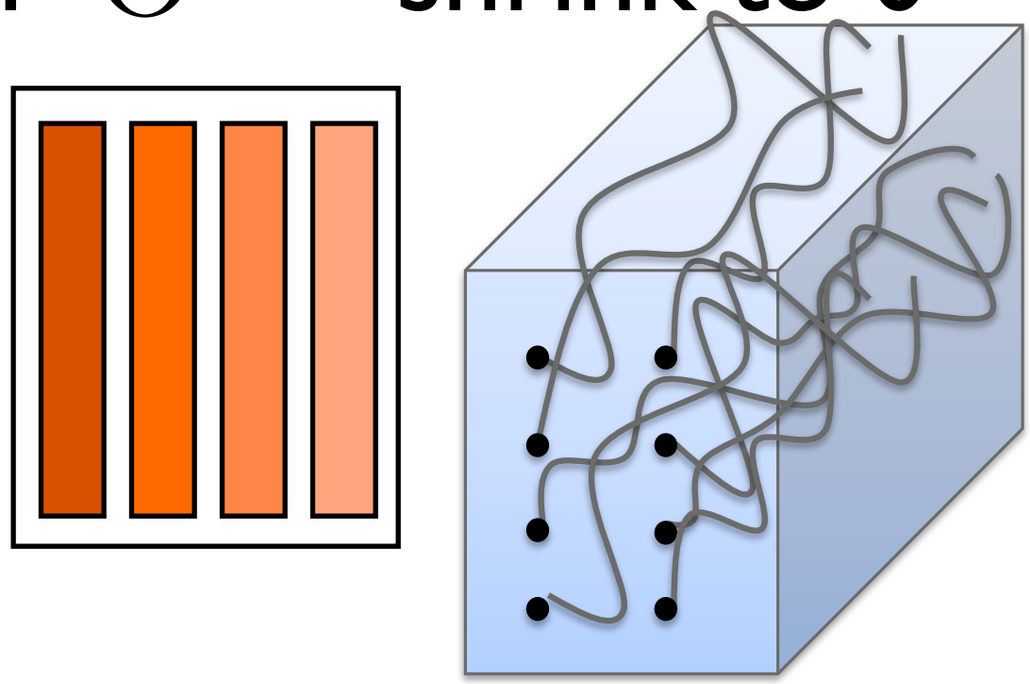
$$= \Theta^{(w)}\xi^{(w)}(t)\xi^{(w)}(t)'\Theta^{(w)\prime}$$

Heteroskedastic



Shrinkage Prior

- $\Lambda^{(w)}(t) = \Theta^{(w)}\xi^{(w)}(t)$
- $\xi^{(w)}(t)$ is a matrix L × K
- Columns of $\Theta^{(w)}$ shrink to 0



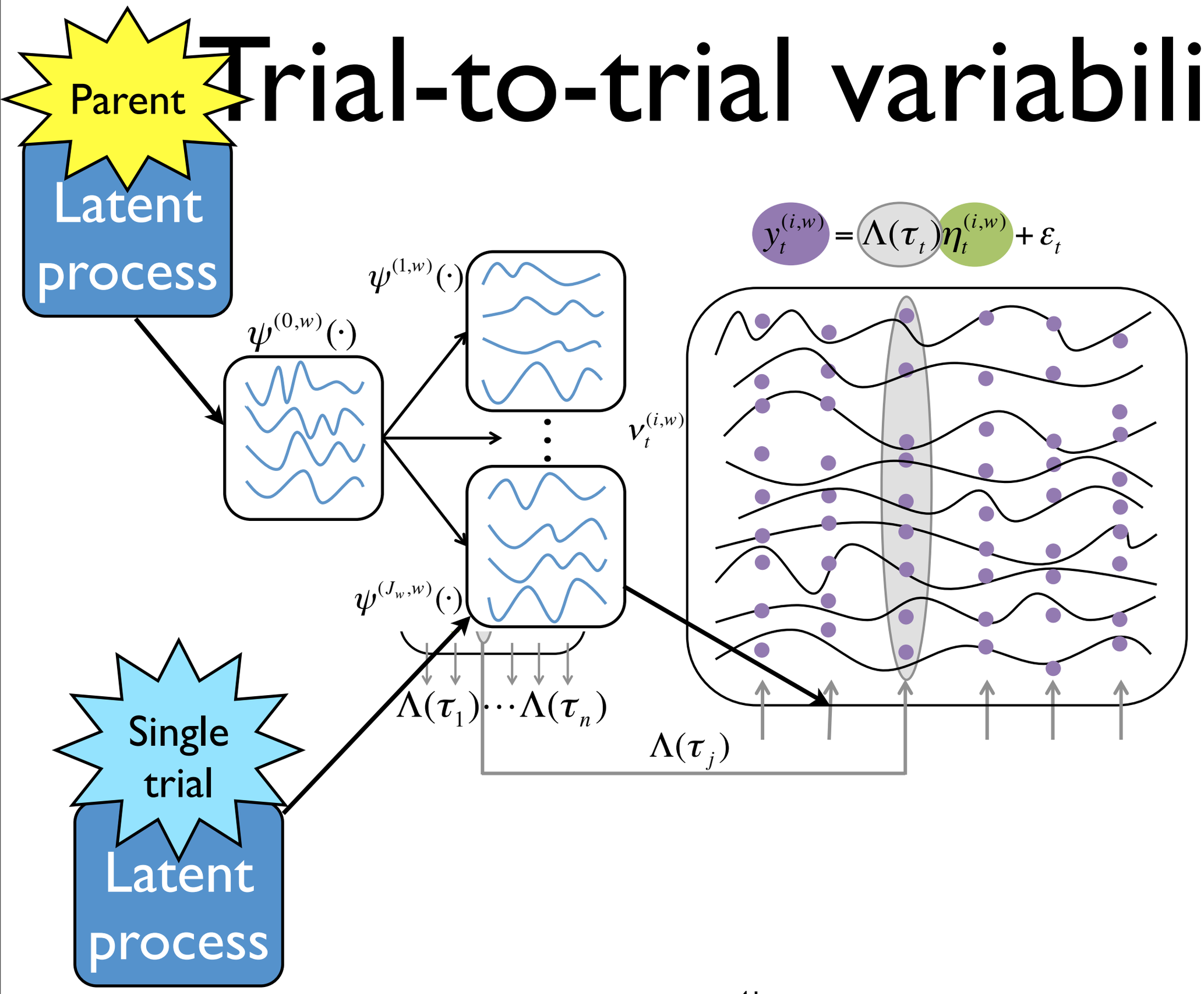
Challenge 4:
Trial-to-trial variability

Trial-to-trial variability

Observation: each trial is a (noisy) recording of the **same process**

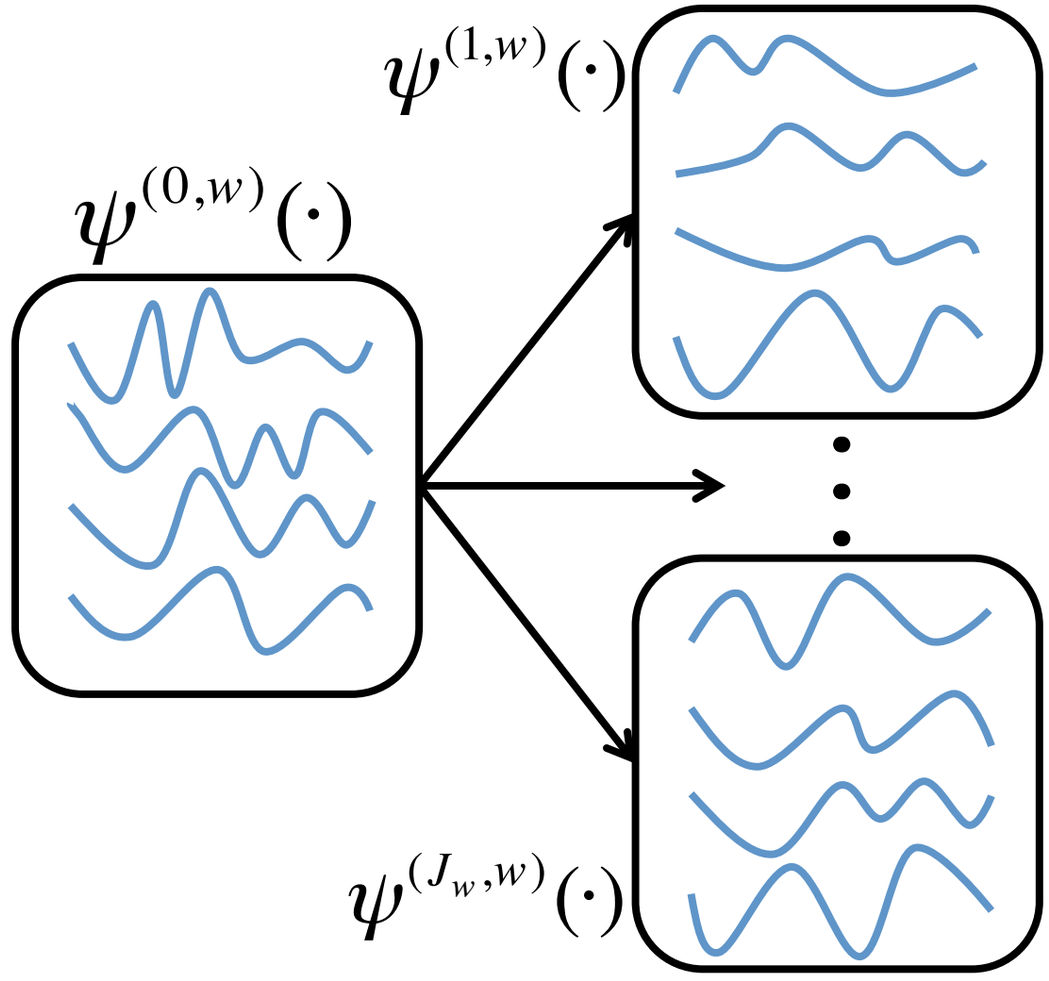
- Allow for multiple examples
- Each example **may deviate** somewhat

Trial-to-trial variability



Trial-to-trial variability

- Hierarchy **shares information** between trials



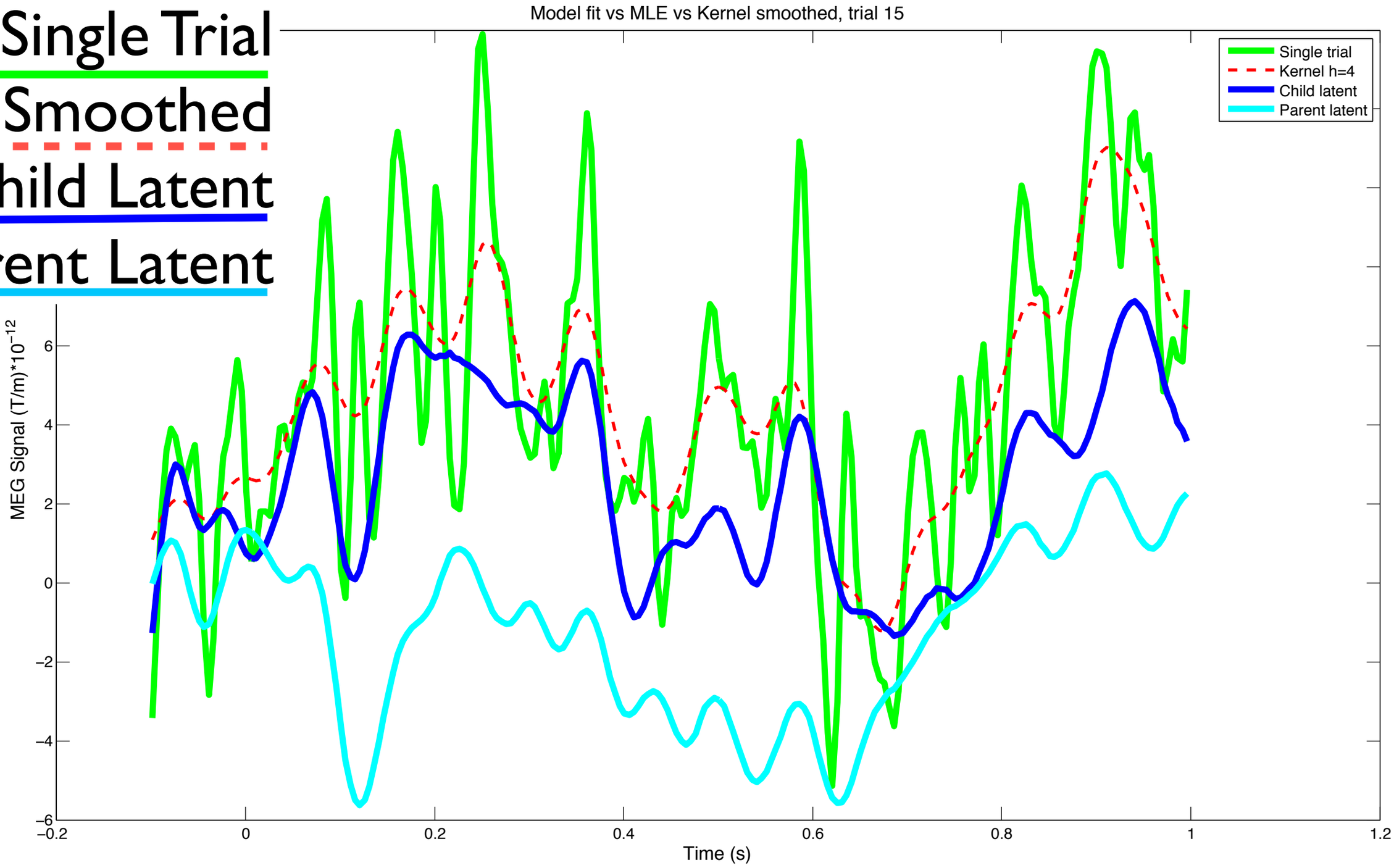
Trial-to-trial variability

Single Trial

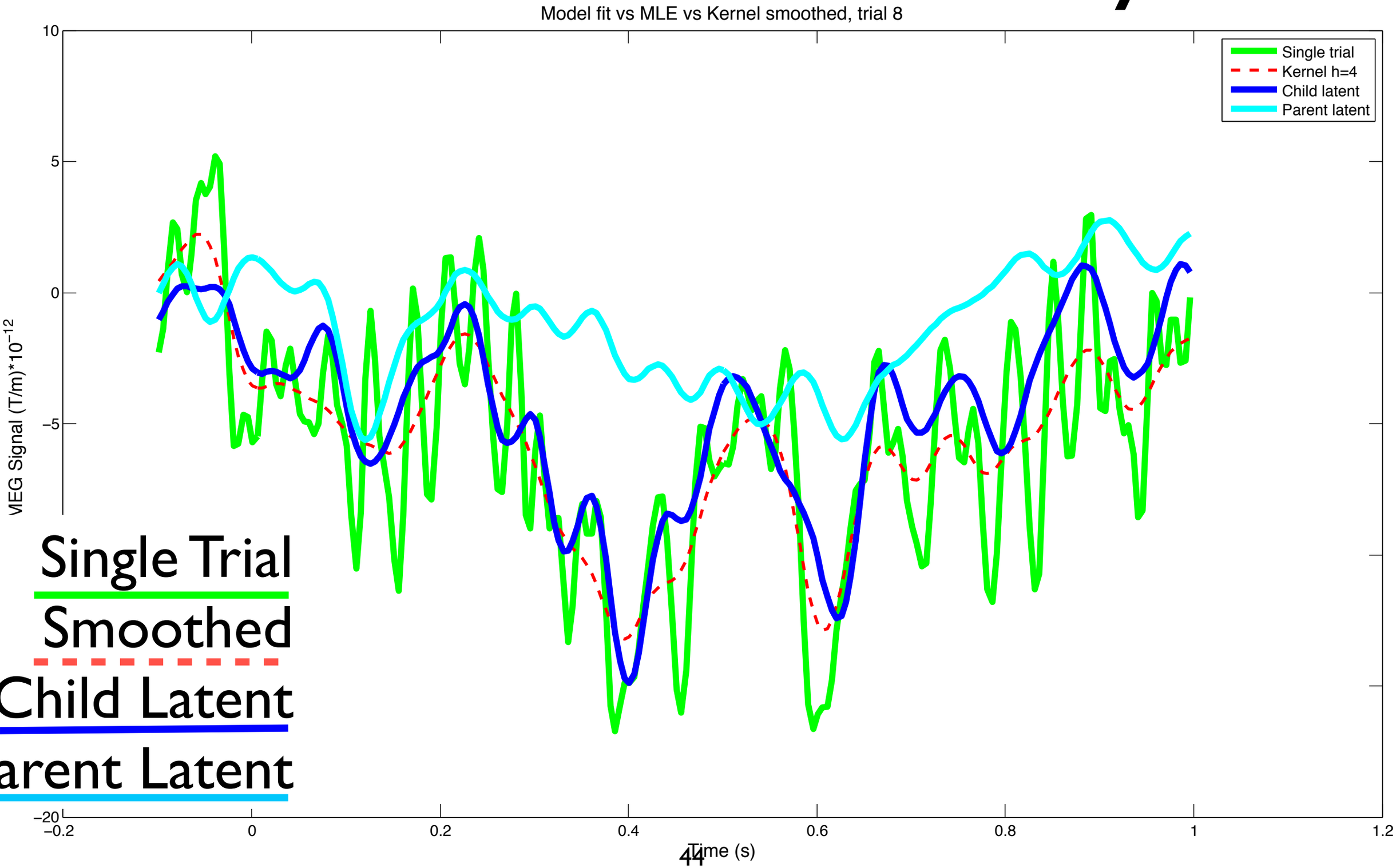
Smoothed

Child Latent

Parent Latent



Trial-to-trial variability



Single Trial

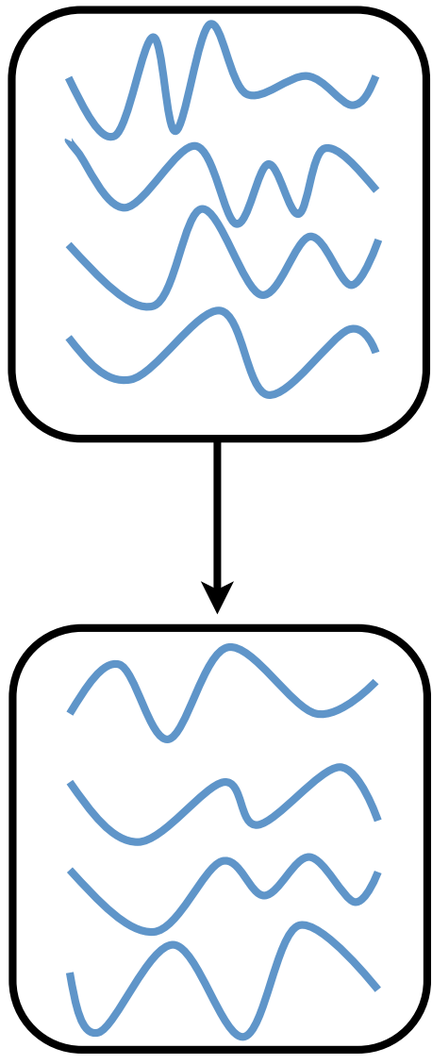
Smoothed

Child Latent

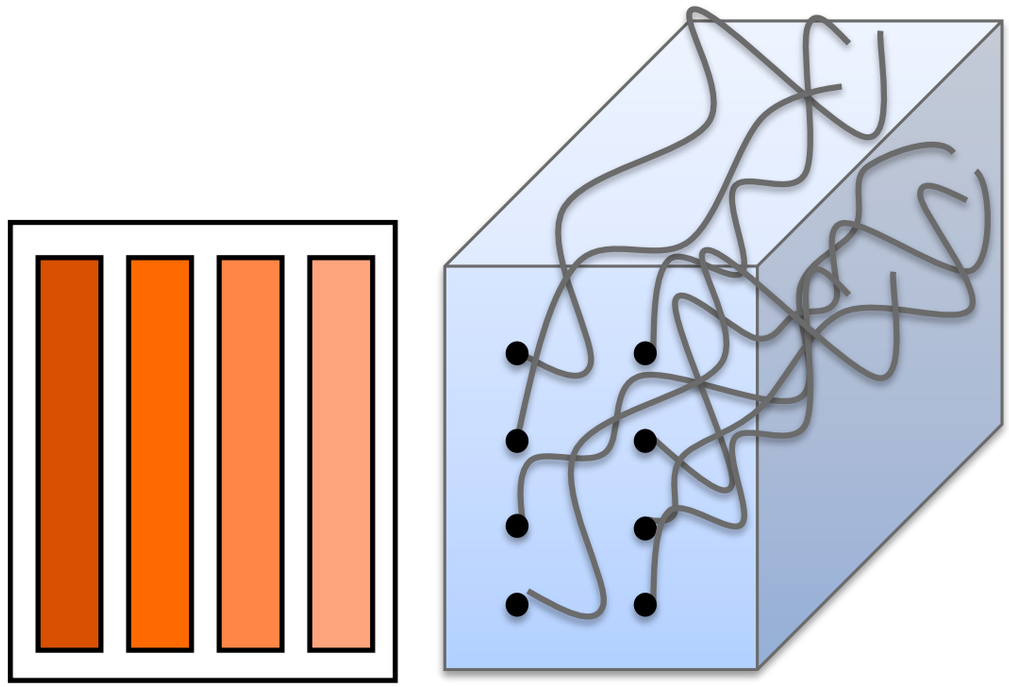
Parent Latent

Generative Story

lettuce

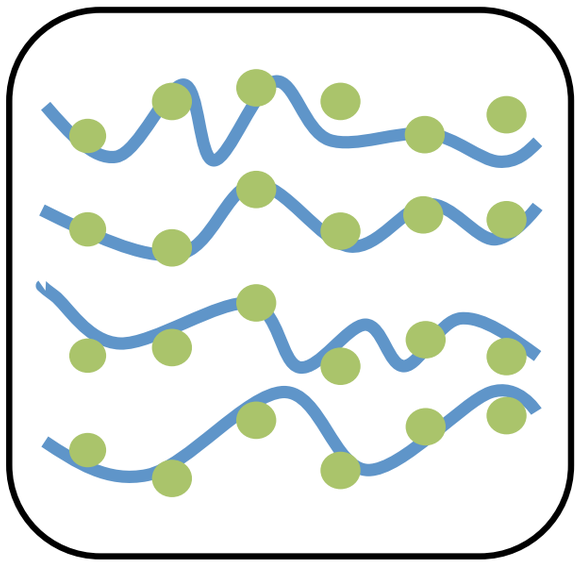


Generative Story

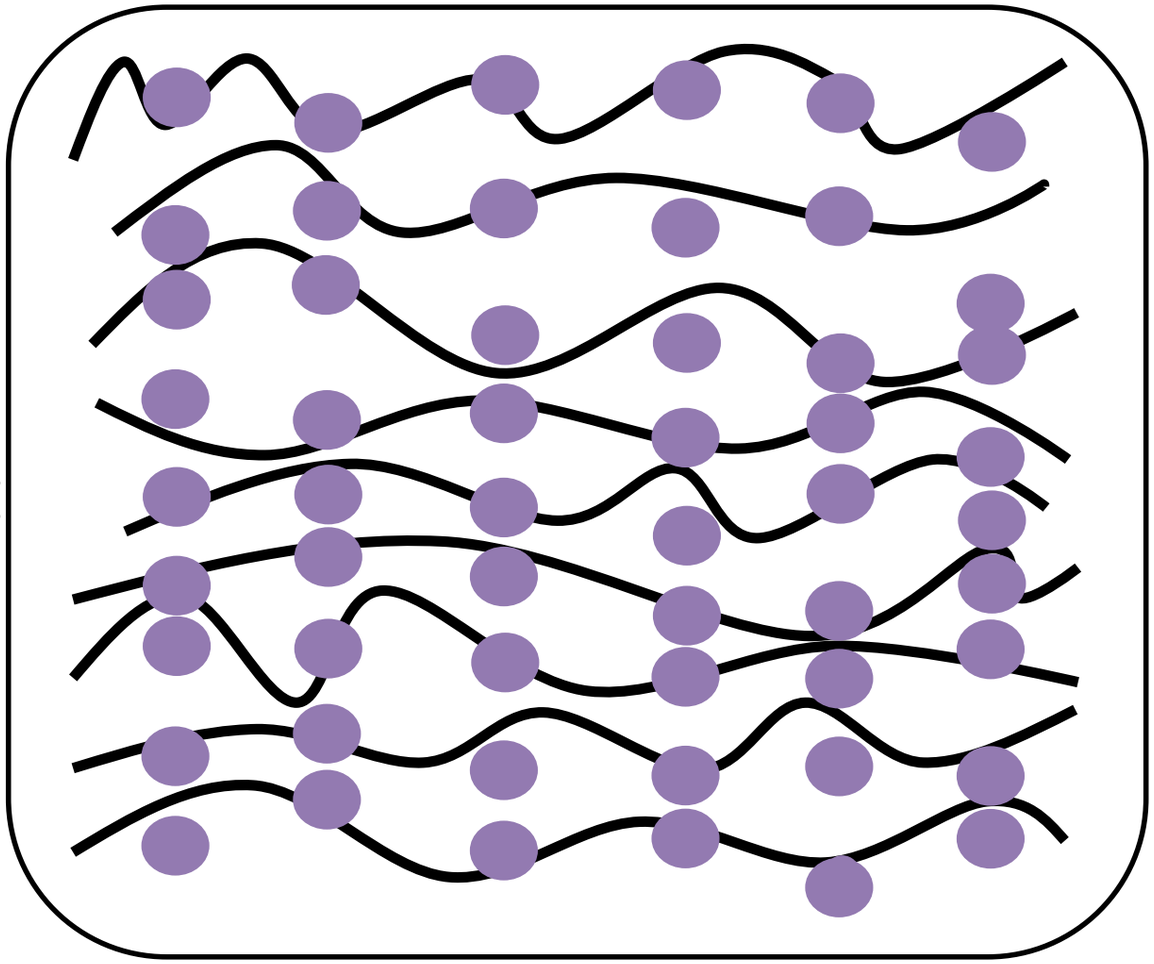


$$\Lambda^{(w)}(t) = \Theta^{(w)}\xi^{(w)}(t)$$

Generative Story



$$\Lambda^{(w)}(t) =$$

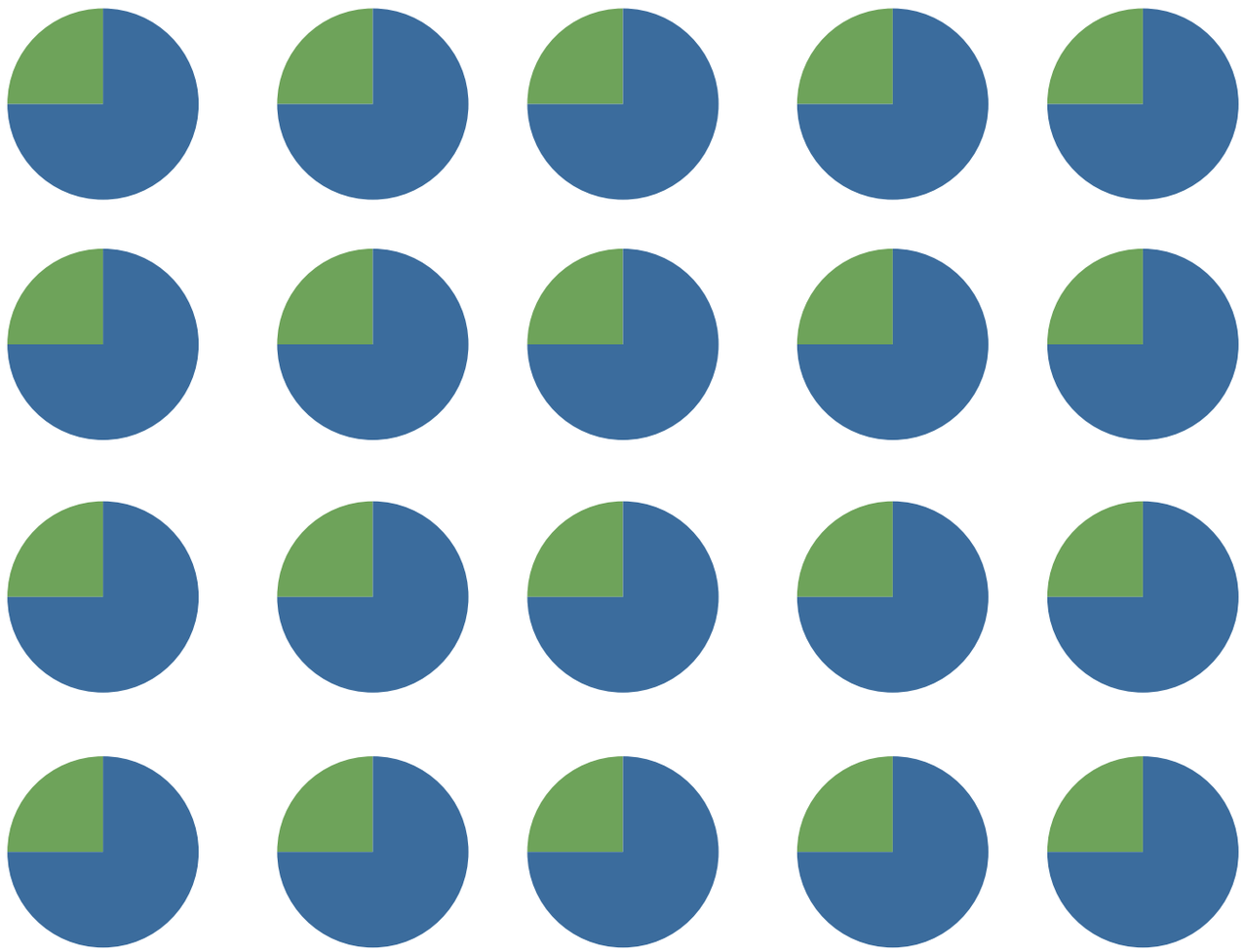


Evaluation

Recall the task

- Distinguish between 4 word categories
- Train on 15 single trials/word
  - 20 words
- Evaluate on 100 single trials
  - 5/word

Train & Test



Fit Model

---

for $i \leftarrow 1, J_w$ **do**

  for $\ell \leftarrow 1, L$ **do**

$$\tilde{\Sigma}_{\psi}^{(w)} = \tilde{K}^{-1} + \text{diag}\left([\Lambda^{(w)}(\tau_1)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_1)[\Lambda^{(w)}(\tau_1)]_{\cdot\ell}, \ldots [\Lambda^{(w)}(\tau_n)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_n)[\Lambda^{(w)}(\tau_n)]_{\cdot\ell}\right)$$

$$\tilde{y}_t^{(i,w)} = y_t^{(i,w)} - \textstyle\sum_{(r\neq\ell)}[\Lambda^{(w)}(\tau)]_{\cdot r}\psi_r^{(i,w)}(\tau)$$

$$\psi_\ell^{(i,w)}(\tau_{1:n}) \sim N_n^{-1}\left(\begin{bmatrix} [\Lambda^{(w)}(\tau_1)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_1)\tilde{y}_1^{(i,w)} + \psi_1^{(0,w)} \\ \vdots \\ [\Lambda^{(w)}(\tau_n)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_n)\tilde{y}_n^{(i,w)} + \psi_n^{(0,w)} \end{bmatrix}, \tilde{\Sigma}_{\psi}^{(w)}\right)$$

  for $t \leftarrow 1, n$ **do**

$$\tilde{y}_t^{(i,w)} = y_t^{(i,w)} - \Lambda^{(w)}(\tau)\psi^{(i,w)}(\tau)$$

$$\nu_t^{(i,w)} \sim N_k^{-1}\left(\Lambda^{(w)'}(\tau)\Sigma_0^{-(w)}\tilde{y}_t^{(i,w)}, I + \Lambda^{(w)'}(\tau)\Sigma_0^{-(w)}\Lambda^{(w)}(\tau)\right)$$

for $\ell \leftarrow 1, L$ **do**

$$\psi_\ell^{(0,w)}(\tau_{1:n}) \sim N_n^{-1}(K_1^{-1}\textstyle\sum_{i=1}^{J_w}\psi_\ell^{(i,w)}(\tau_{1:n}), K_0^{-1} + J_w K_1^{-1})$$

for $\ell \leftarrow 1, L$ **do**

  for $m \leftarrow 1, K$ **do**

$$\tilde{\Sigma}_\xi = K_1^{-1} + \textstyle\sum_{i=1}^{J_w}\text{diag}\left(\left(\eta_{1,m}^{(i,w)}\right)^2\sum_{j=1}^{p}\left(\Theta_{j\ell}^{(w)}\right)^2\sigma_{j,w}^{-2}, \ldots \left(\eta_{n,m}^{(i,w)}\right)^2\sum_{j=1}^{p}\left(\Theta_{j\ell}^{(w)}\right)^2\sigma_{j,w}^{-2}\right)$$

$$\tilde{y}_{t,j}^{(i,w)} = y_{t,j}^{(i,w)} - \textstyle\sum_{(r,s)\neq(\ell,m)}\Theta_{jr}^{(w)}\xi_{rs}^{(w)}(\tau).$$

$$\xi_{\ell m}^{(w)}(\tau_{1:n}) \sim N_n^{-1}\left(\sum_{i=1}^{J_w}\begin{bmatrix} \eta_{1,m}^{(i,w)}\sum_{j=1}^{p}\Theta_{j\ell}^{(w)}\sigma_{j,w}^{-2}\tilde{y}_{1,j}^{(i,w)} \\ \vdots \\ \eta_{n,m}^{(i,w)}\sum_{j=1}^{p}\Theta_{j\ell}^{(w)}\sigma_{j,w}^{-2}\tilde{y}_{n,j}^{(i,w)} \end{bmatrix}, \tilde{\Sigma}_\xi\right)$$

for $j \leftarrow 1, p$ **do**

$$\Theta_{j\cdot}^{(w)} = \begin{bmatrix}\Theta_{j1}^{(w)} & \ldots & \Theta_{jL}^{(w)}\end{bmatrix},\ \eta_t^{(i,w)} = \psi^{(i,w)}(\tau) + \nu_t^{(i,w)}$$

$$\sigma_{w,j}^{-2} \sim \text{Ga}\left(a_\sigma + \tfrac{nJ_w}{2}, b_\sigma + \tfrac{1}{2}\textstyle\sum_{i=1}^{J_w}\sum_{t=1}^{n}(y_{t,j}^{(i,w)} - \Theta_{j\cdot}^{(w)}\xi^{(w)}(\tau)\eta_t^{(i,w)})^2\right)$$

for $j \leftarrow 1, p$ **do**

$$y_{\cdot,j} = [y_{(1,j)}(t_{1:n}) \ldots y_{(J_w,j)}(t_{1:n})]'$$

$$\tilde{\eta}^{(w)'} = \begin{bmatrix}\xi^{(w)}(\tau_1)\eta_1^{(1,w)} & \ldots & \xi^{(w)}(\tau_n)\eta_n^{(1,w)} & \ldots & \xi^{(w)}(\tau_1)\eta_1^{(J_w,w)} & \ldots & \xi^{(w)}(\tau_n)\eta_n^{(J_w,w)}\end{bmatrix}$$

$$\tilde{\Sigma}_\Theta^{-(w)} = \sigma_{w,j}^{-2}\tilde{\eta}^{(w)'}\tilde{\eta}^{(w)} + \text{diag}(\phi_{j1}^{(w)}\zeta_1^{(w)}, \ldots, \phi_{jL}^{(w)}\zeta_L^{(w)})$$

$$\Theta_{j\cdot}^{(w)} \sim N_L\left(\sigma_{w,j}^{-2}\tilde{\Sigma}_\Theta^{(w)}\tilde{\eta}^{(w)'}y_{\cdot,j}, \tilde{\Sigma}_\Theta^{(w)}\right)$$

Evaluating Models

Single trial

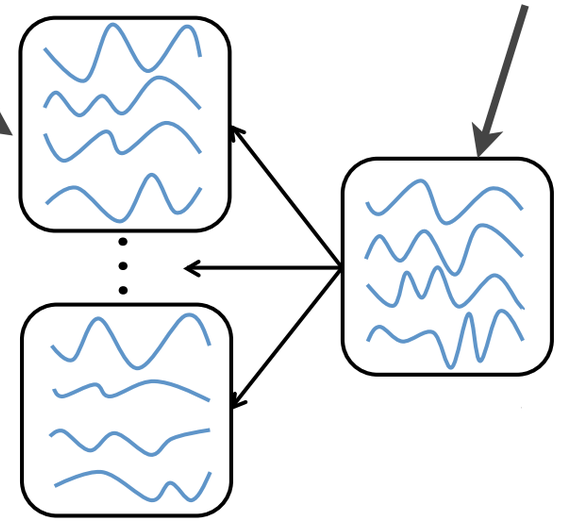
Training data

$$p(y^*_{1:n} \mid \{y^{i,w}_{1:n}\}_{i=1}^{J_w}) \approx \frac{1}{300 \cdot M} \sum_{m=1}^{M} \left(\sum_{n=1}^{300} p(y^*_{1:n} | \Omega_m^{(w)}, \psi^{*(i,m)}) \right)$$

Gibbs Samples

Draws of

$$\psi_m^{*(i,w)}(t_{1:n}) \sim N_n(\psi_m^{(0,w)}, K)$$



Testing

- MLE-based models (likelihood)
  - Homoskedastic $\mu_{MLE}^{(w)} \quad \hat{\Sigma}^{(w)}$
  - Heteroskedastic $\mu_{MLE}^{(w)} \quad \hat{\Sigma}_t^{(w)}$
- Our model $\psi^*(t) \sim \mathrm{GP}(\psi^{(0,w)}(t), K)$

$$\mu^*(t) = \Lambda^{(w)}(t)\psi^*(t)$$

$$\Sigma^{(w)}(t) = \Lambda^{(w)}(t)\Lambda^{(w)}(t)'$$

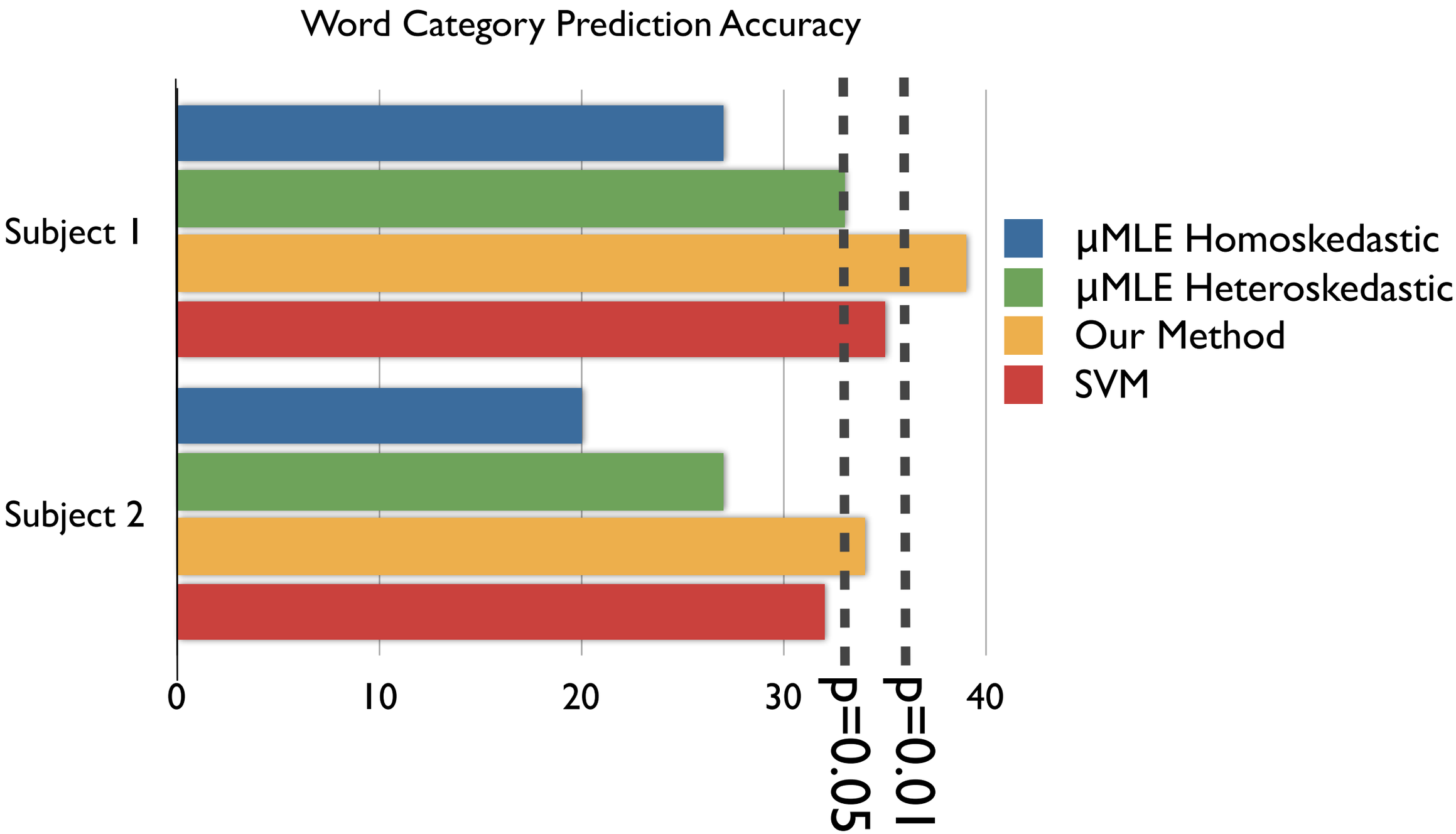
Testing

- Choose category of word that maximizes log likelihood

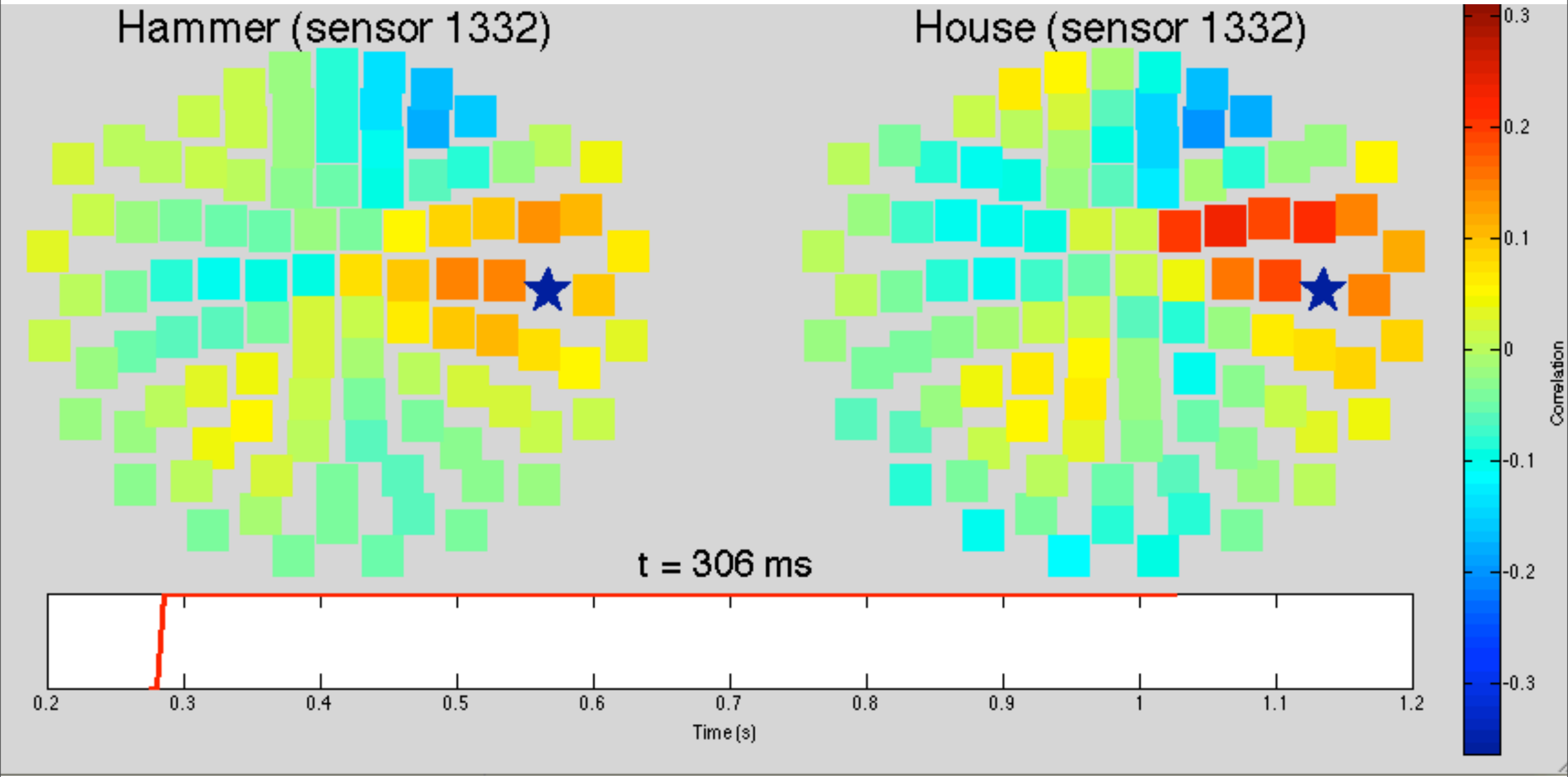
$$\underset{w}{\text{argmax}}\{ll(y^{(*)}_{1:n}|\mu,\Sigma)\}$$

- Evaluate Linear SVM for comparison
  - 4 one vs all SVMs
  - ties broken by distance to hyperplane

Results


Word Category Prediction Accuracy
Subject 1
Subject 2
0
10
20
30
40
p=0.05
p=0.01
μMLE Homoskedastic
μMLE Heteroskedastic
Our Method
SVM

Results


Hammer (sensor 1332)
House (sensor 1332)
t = 306 ms
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
Time (s)
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Correlation

Summary

- Heteroskedastic model
- Hierarchy of latent processes

1) Noise

2) High Dimension

3) Brain coordination changes over time

4) Trial-to-Trial Variability

Future Work

- Extend hierarchy
  - over words in categories
  - all words
  - subjects

Thanks!

Contact:

Alona Fyshe afyshe@cs.cmu.edu

Code (soon):

http://www.cs.cmu.edu/~afyshe/papers/aistats2012

Funding:

- Natural Sciences and Engineering Research Council of Canada (NSERC)
- W. M. Keck Foundation
- National Science Foundation (NSF)
- National Institute of Environmental Health Sciences