Effective Split-Merge Monte Carlo Methods for Nonparametric Models of Sequential Data

Michael C. Hughes,+ Emily B. Fox,w and Erik B. Sudderth+

BROWN UNIVERSITY of WASHINGTON

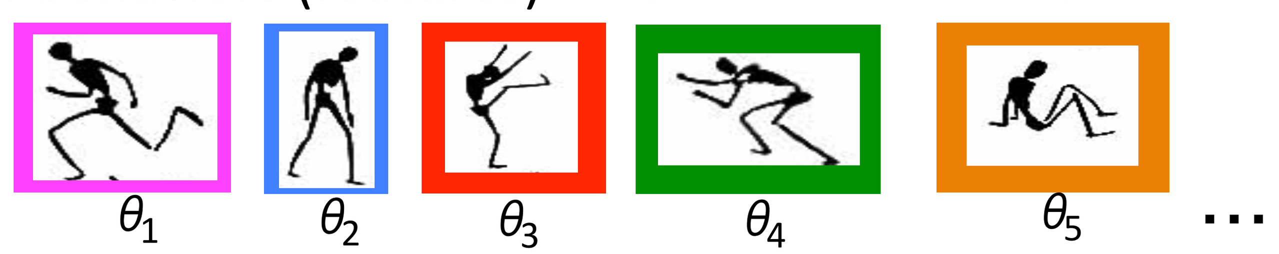
Model

Beta Process HMM [Fox et al. NIPS '09]

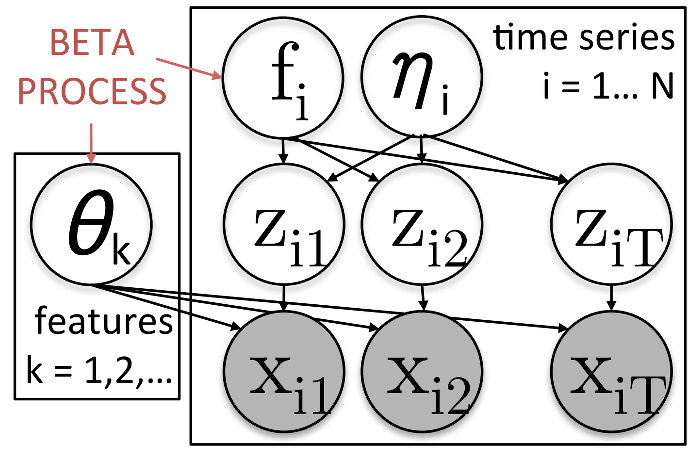
*Generates **collections** of sequential data*

Global set of behaviors (features) *# of behaviors learned from data*

θ emission parameters



Each sequence uses a sparse subset of behaviors



η_i seq. *i*'s HMM transition weights

f_i seq. *i*'s binary feature assignment

z_{it} active feature at time *t* in seq. *i*

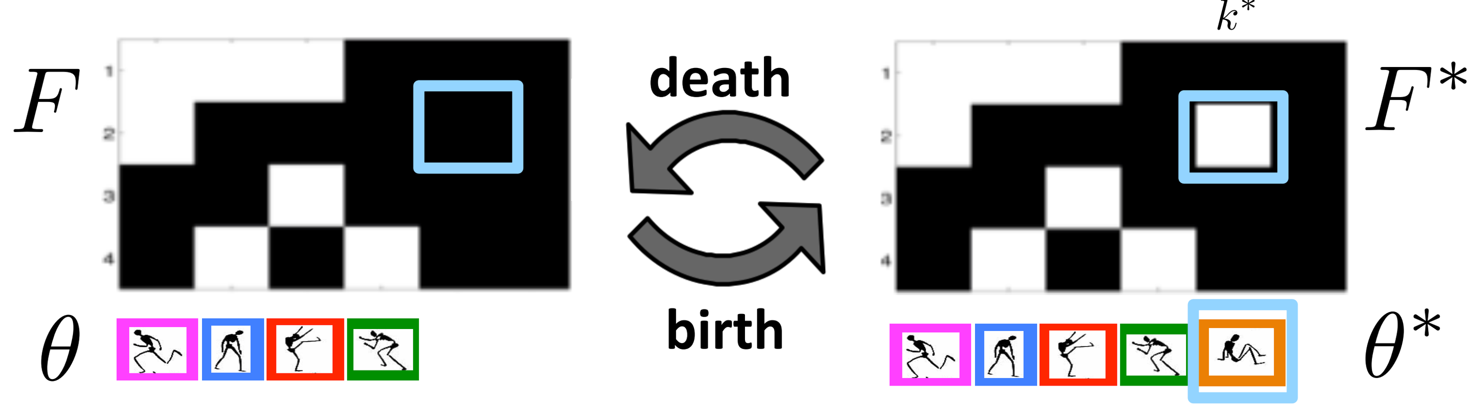
x_{it} observed data at time *t* in seq. *i*

$$x_{it} \sim p(\cdot|\theta_{z_{it}})$$

Data-Driven Birth/Death

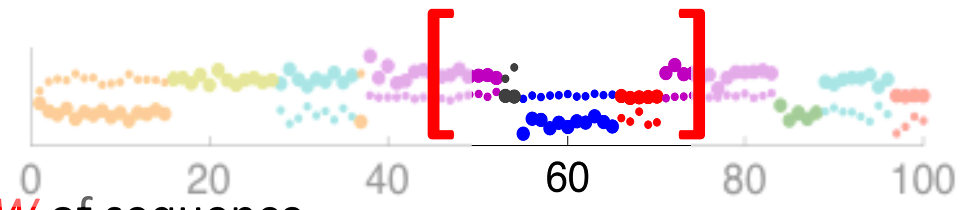
Add or delete unique feature to **one sequence**

Reversible jump proposal for F, θ (z marginalized out)



Propose from prior [Fox et al. NIPS 2009]

$$\theta^*_{k^*} \sim p(\theta)$$



Data-driven proposal

- select random window *W* of sequence
- proposal: mixture of prior and posterior over *W*

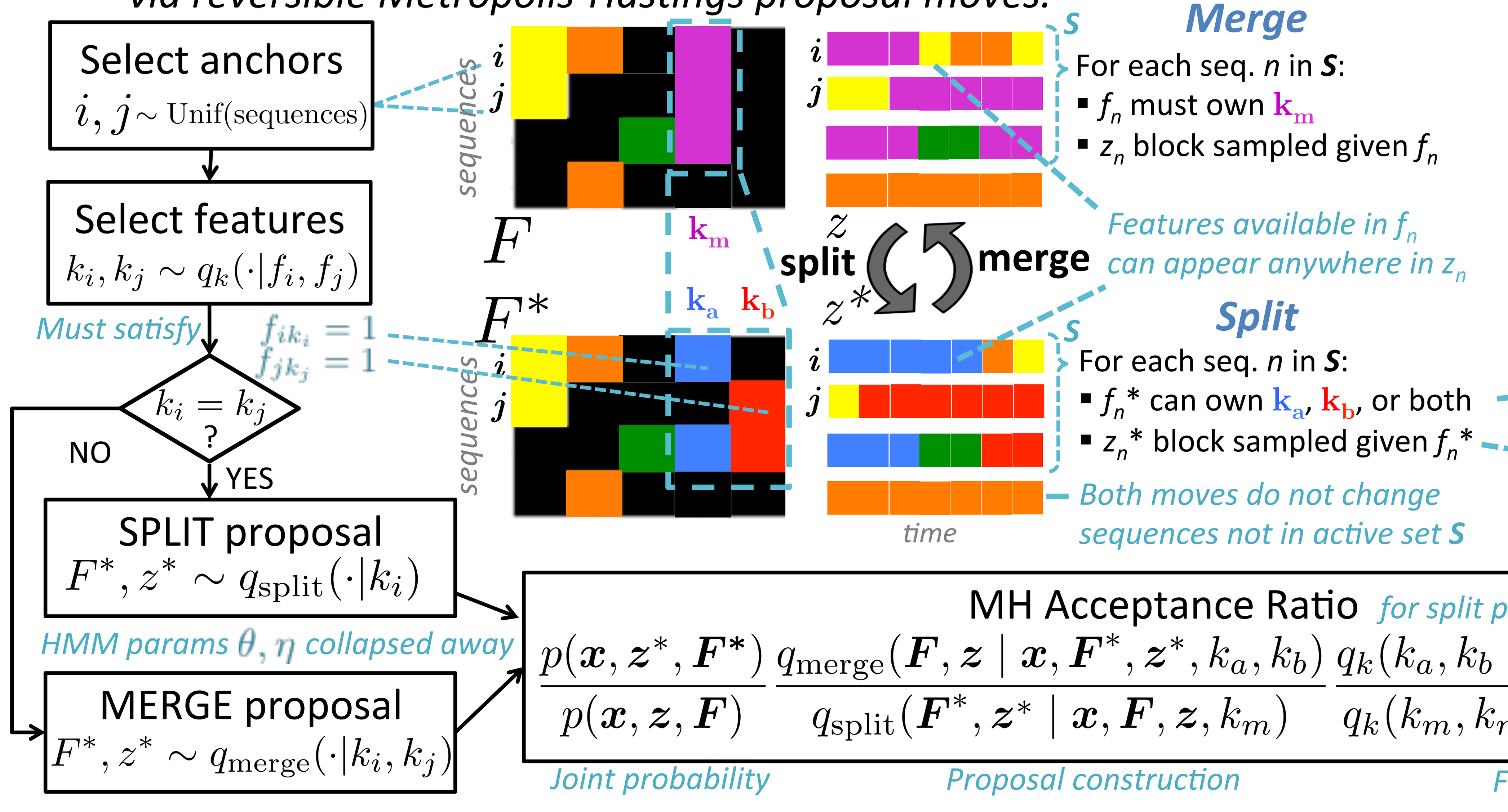
$$\theta^*_{k^*} \sim \tfrac{1}{2}p(\theta) + \tfrac{1}{2}p(\theta|x_{it} : t \in W)$$

Using mixture ensures good death move acceptance rate

Efficiently adds new behaviors informed by the data

Split-Merge

Yields exact posterior samples with big changes to feature assignments via reversible Metropolis-Hastings proposal moves.



MH Acceptance Ratio *for split proposal*

$$\frac{p(\boldsymbol{x}, \boldsymbol{z}^*, \boldsymbol{F}^*)}{p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{F})} \frac{q_{\text{merge}}(\boldsymbol{F}, \boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{F}^*, \boldsymbol{z}^*, k_a, k_b)}{q_{\text{split}}(\boldsymbol{F}^*, \boldsymbol{z}^* \mid \boldsymbol{x}, \boldsymbol{F}, \boldsymbol{z}, k_m)} \frac{q_k(k_a, k_b \mid \boldsymbol{x}, \boldsymbol{F}^*, \boldsymbol{z}^*, i, j)}{q_k(k_m, k_m \mid \boldsymbol{x}, \boldsymbol{F}, \boldsymbol{z}, i, j)}$$

Joint probability *Proposal construction* *Feature selection*

Split Proposal Details

Inspired by sequential allocation [Dahl '05]

builds good proposals in one scan of data

Proposed F,z* collapses away HMM params.*

non-random auxiliary $\hat{\theta}, \hat{\eta}$ allow sampling via dyn. prog.

1) Initialize

Transitions $\hat{\eta}$ set to prior mean

Emissions $\hat{\theta}$ set to posterior mean

$$\hat{\theta} \leftarrow \mathbb{E}[\theta \mid \mathbf{x}, \mathbf{z}, \lambda]$$

2) Sequentially allocate non-anchors

IBP prior given previous allocations

$$f_n \sim p(f_{n,[k_a\ k_b]}|F_{S_{\text{prev}},[k_a\ k_b]}) \times p(x_n|f_n, \hat{\theta}, \hat{\eta}_n)$$

HMM marg. likelihood

$$z_n \sim p(z_n|x_n, f_n, \hat{\theta}, \hat{\eta}_n)$$

Backward Filter, Forward Sample via dynamic programming

3) Allocate anchors

To ensure reversibility via a merge, *i* must own k_a *j* must own k_b

This does *not* require that k_a appears in z_i or k_b appears in z_j

Problem

Existing MCMC inference slow to mix

- Requires long time to make big changes
  - Each update touches small subset of variables
- Rarely creates new features
  - Proposals from vague prior poorly-matched to data

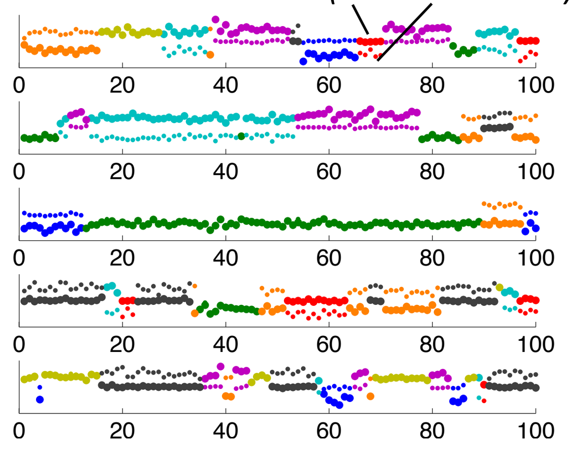
Contributions

- **Split-Merge (SM)** move
  - Change many feature assignments at once
- **Data-Driven (DD)** birth/death move
  - Propose new features consistent with observed data

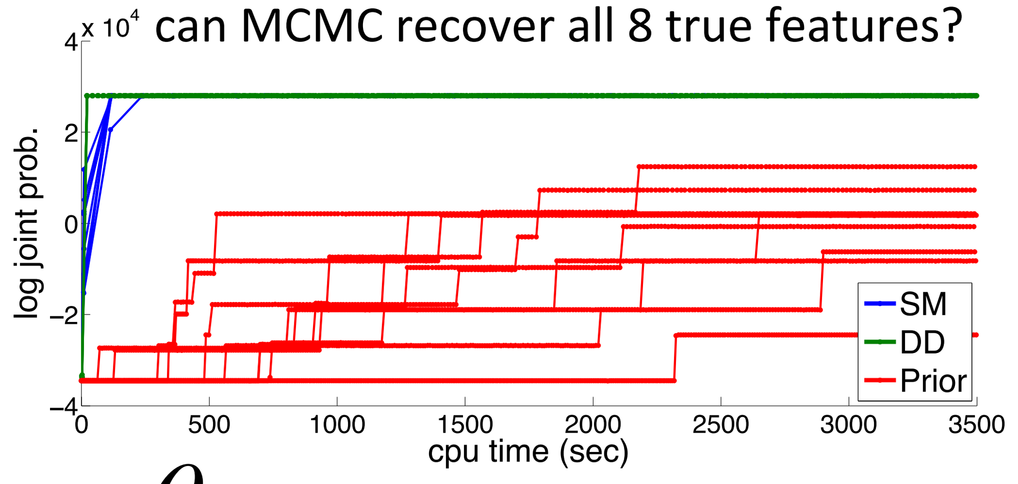
Significantly improves mixing.
Scales to 100+ sequences.

Toy Experiments

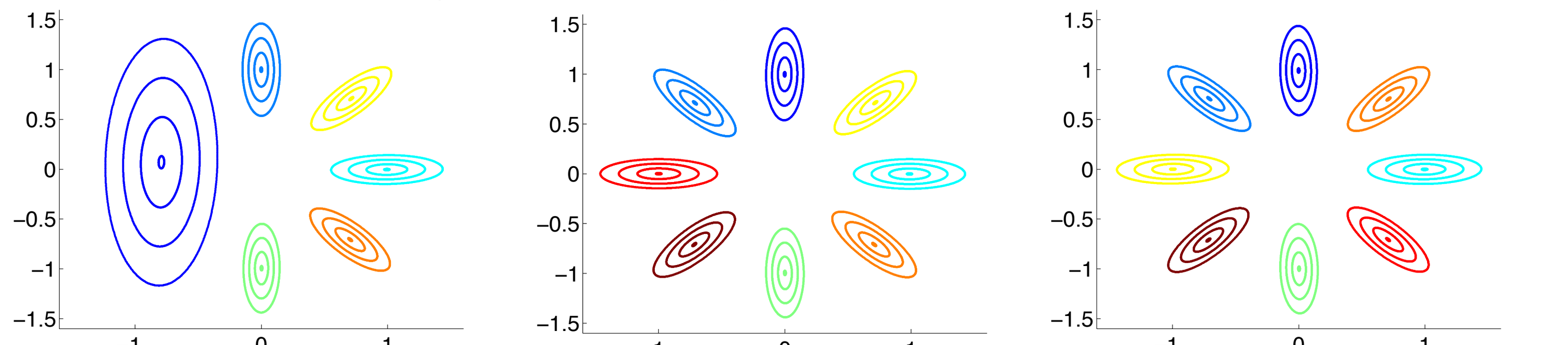
Example toy sequences
5 dim. Gaussian obs. (1st & 2nd shown)



Starting at poor initialization of one feature, can MCMC recover all 8 true features?



Discovered emission parameters θ



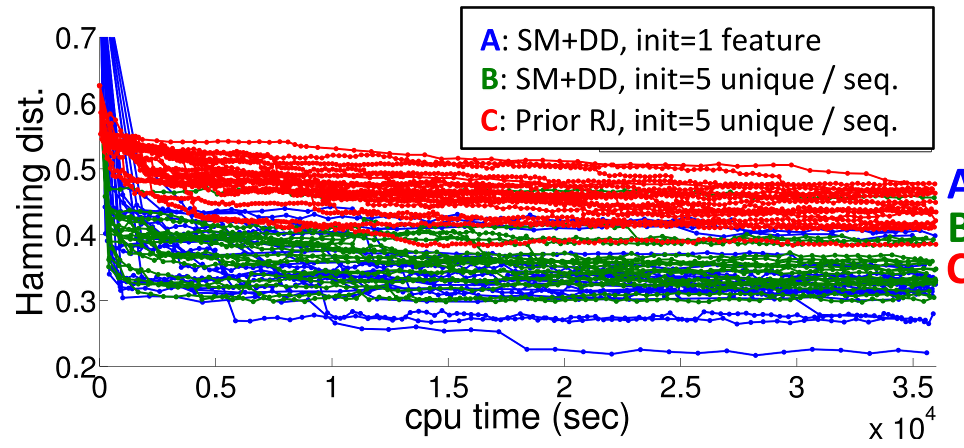
SM and DD moves find essential features
Prior methods stuck in bad local optima

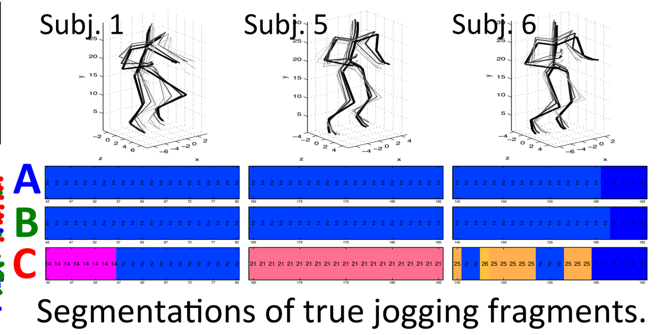
Motion Capture

6 subjects performing 12 exercises (7 shared, 5 unique).

Data: 12 joint angle sensors. **Model:** 1st-order auto-regressive.

Goal: Compare BP-HMM segmentation to human annotation.

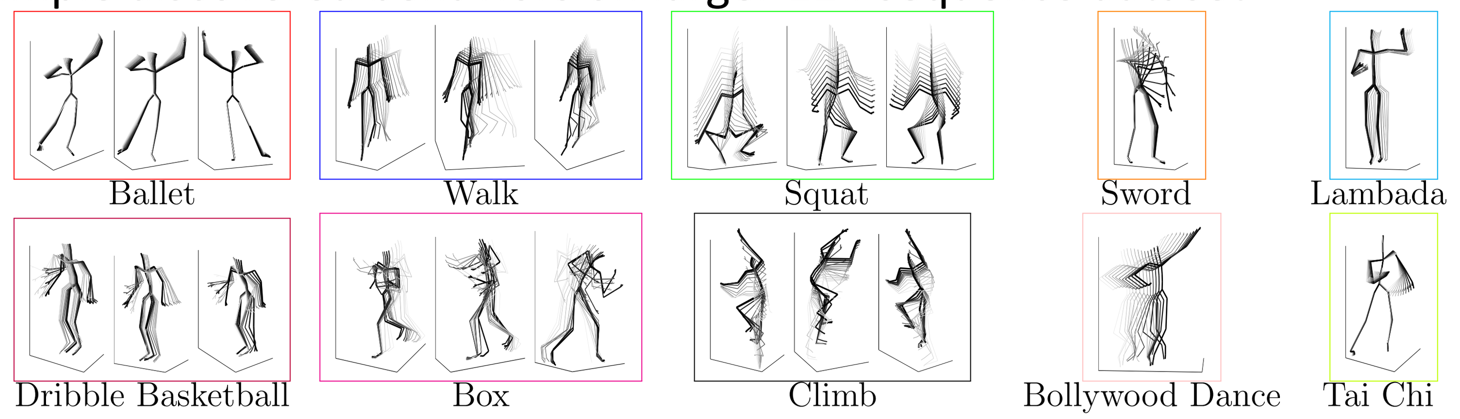




Segmentations of true jogging fragments.
SM+DD finds dominant single behavior.
Prior unable to merge due to local moves.

Previous work stuck with split behaviors even with clever initialization.
New MCMC finds better segmentation starting with just one feature.

Example discovered behaviors on larger 124 sequence dataset



Kitchen Video

126 subjects prepare 5 recipes in common kitchen. **Model**: Multinomial.

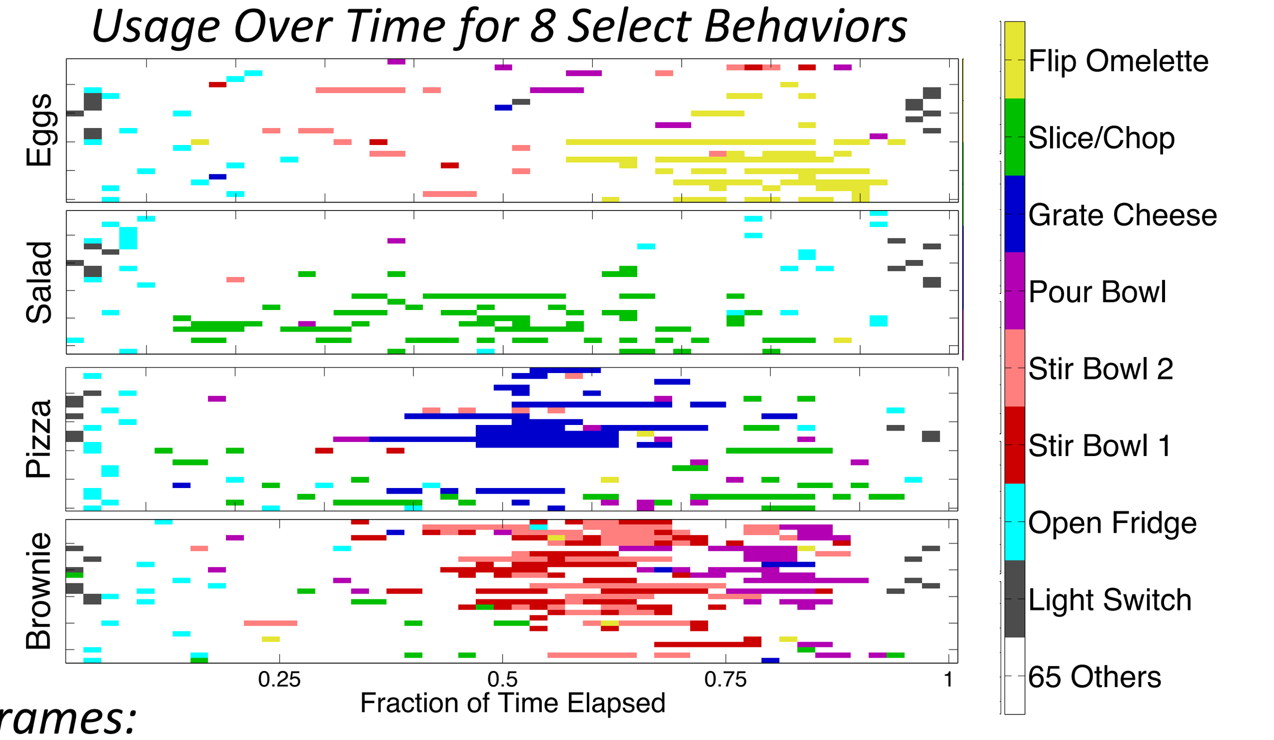
Data: Bag of visual words (motion + appearance), at 1 second intervals.

Usage Over Time for 8 Select Behaviors

Each row shows behaviors used throughout one video at final MCMC iteration.

Unsupervised. Recipe labels not given as input.

Not shown: Sandwich videos.



Example Key Frames: