Exploiting compositionality to explore a large space of model structures

Roger Grosse

Dept. of Computer Science,
University of Toronto

Introduction

How has the life of a machine learning engineer changed in the past decade?

Many tasks that previously required human experts are starting to be automated

feature engineering

Hyperopt

AutoWeka

WHETLAB

algorithm configuration

Stan

infer.net

probabilistic programming

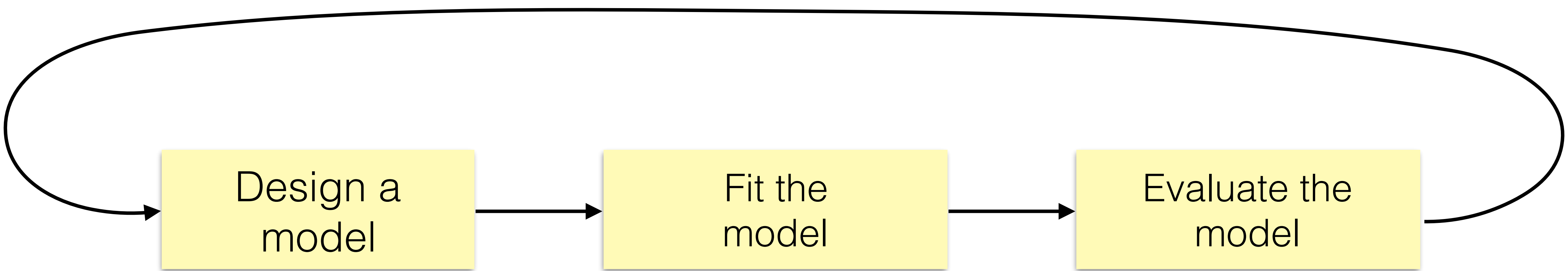
probabilistic inference

?

model selection

The probabilistic modeling pipeline

Can we identify good models automatically?

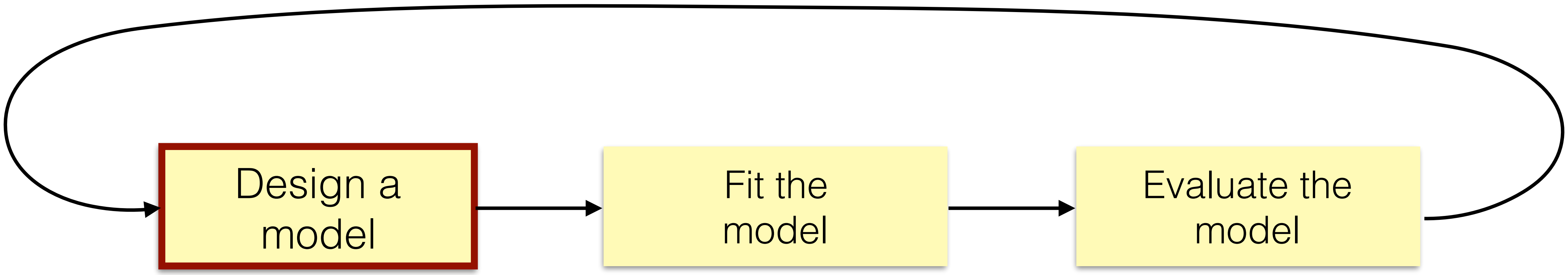


Two challenges:

Automating each stage of this pipeline

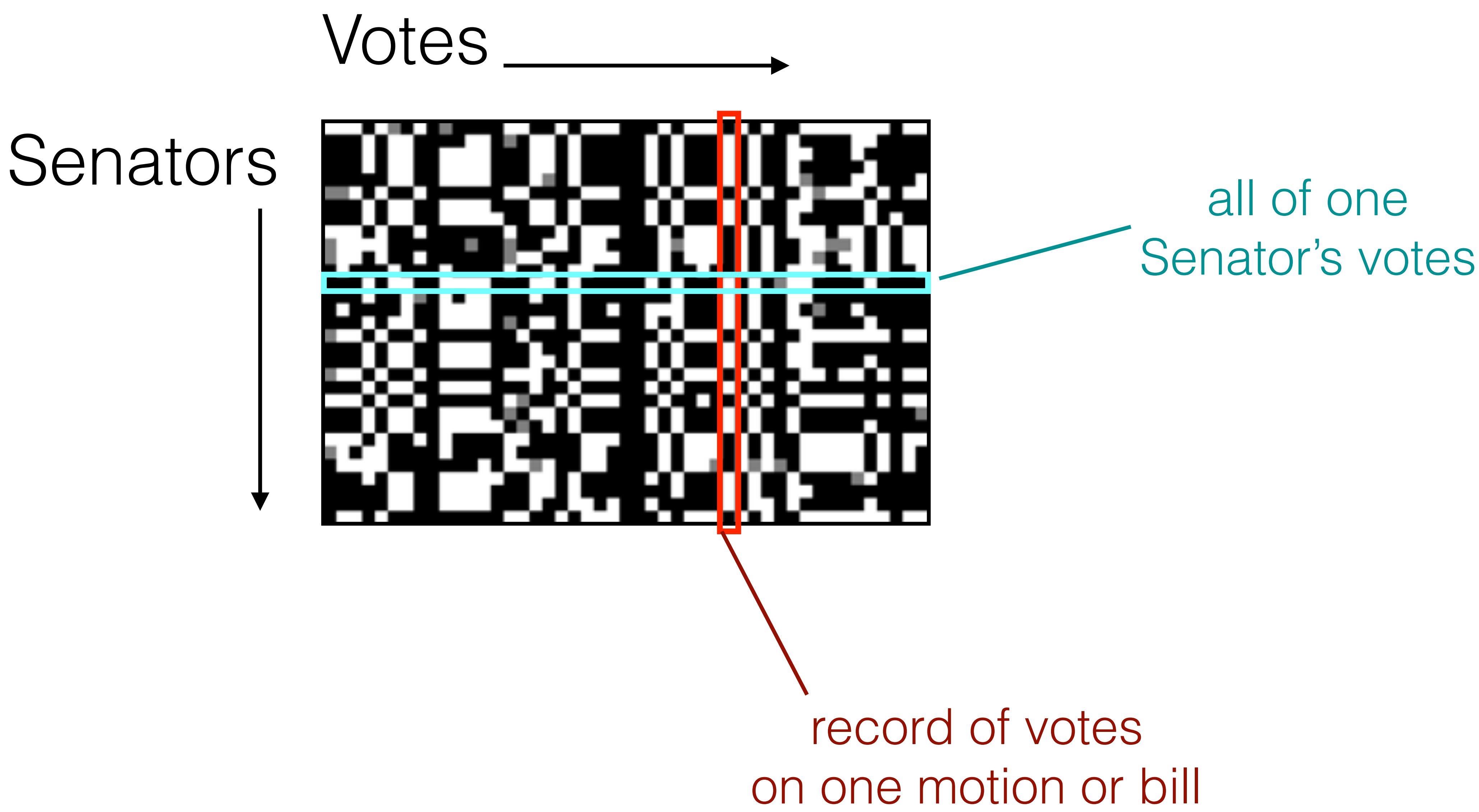
Identifying a promising set of candidate models

The probabilistic modeling pipeline



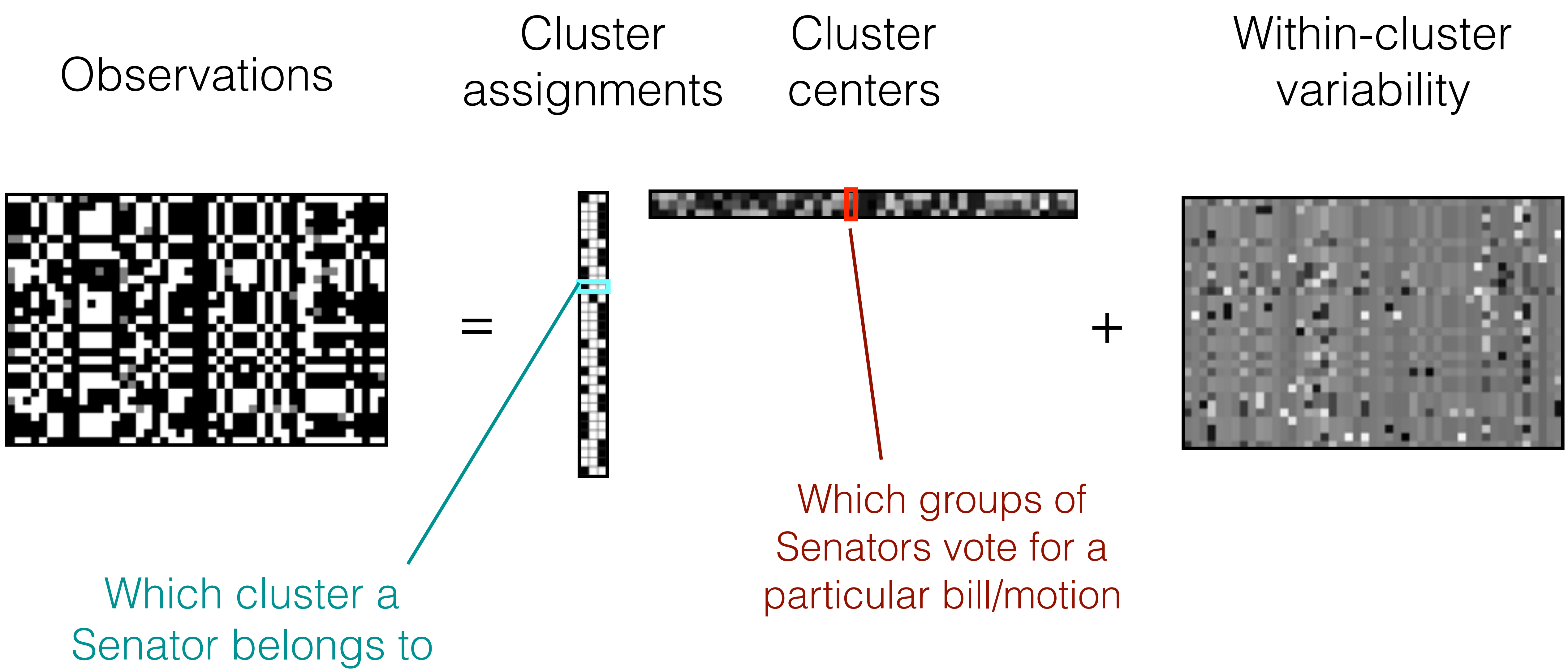
Matrix decompositions

Example: Senate votes, 2009-2010



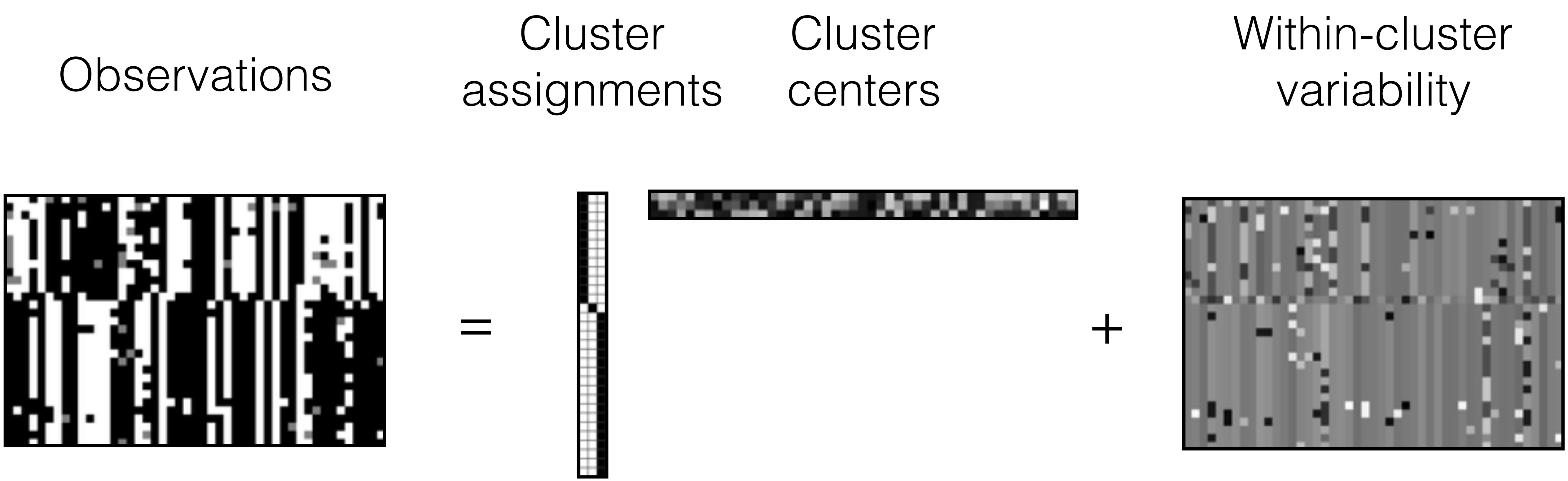
Matrix decompositions

Clustering the Senators



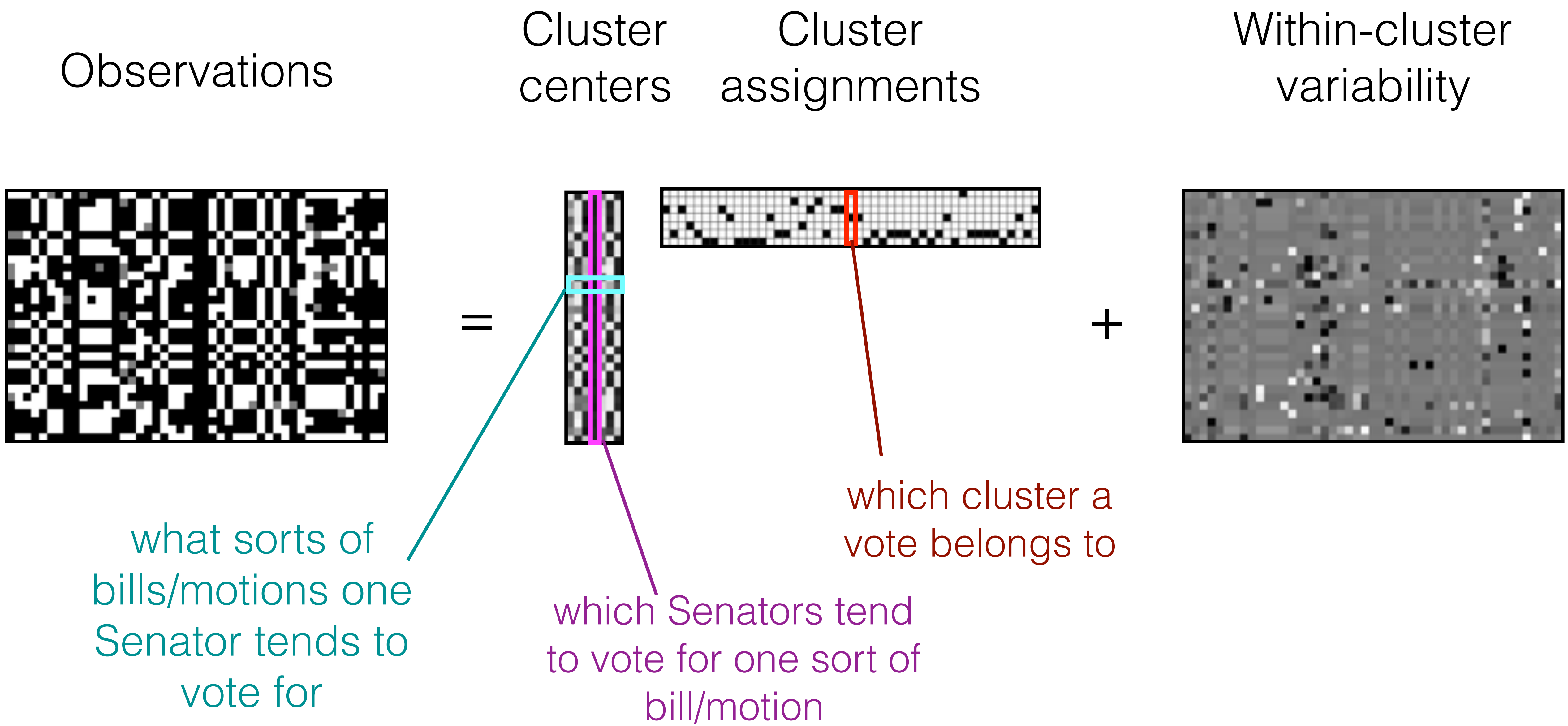
Matrix decompositions

Clustering the Senators



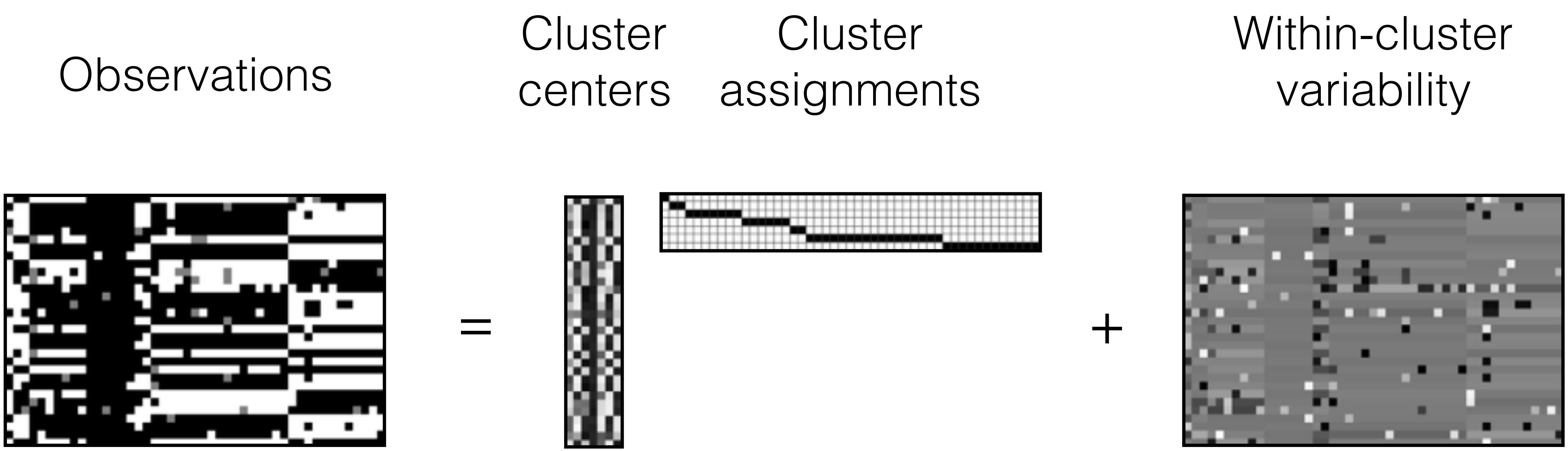
Matrix decompositions

Clustering the votes



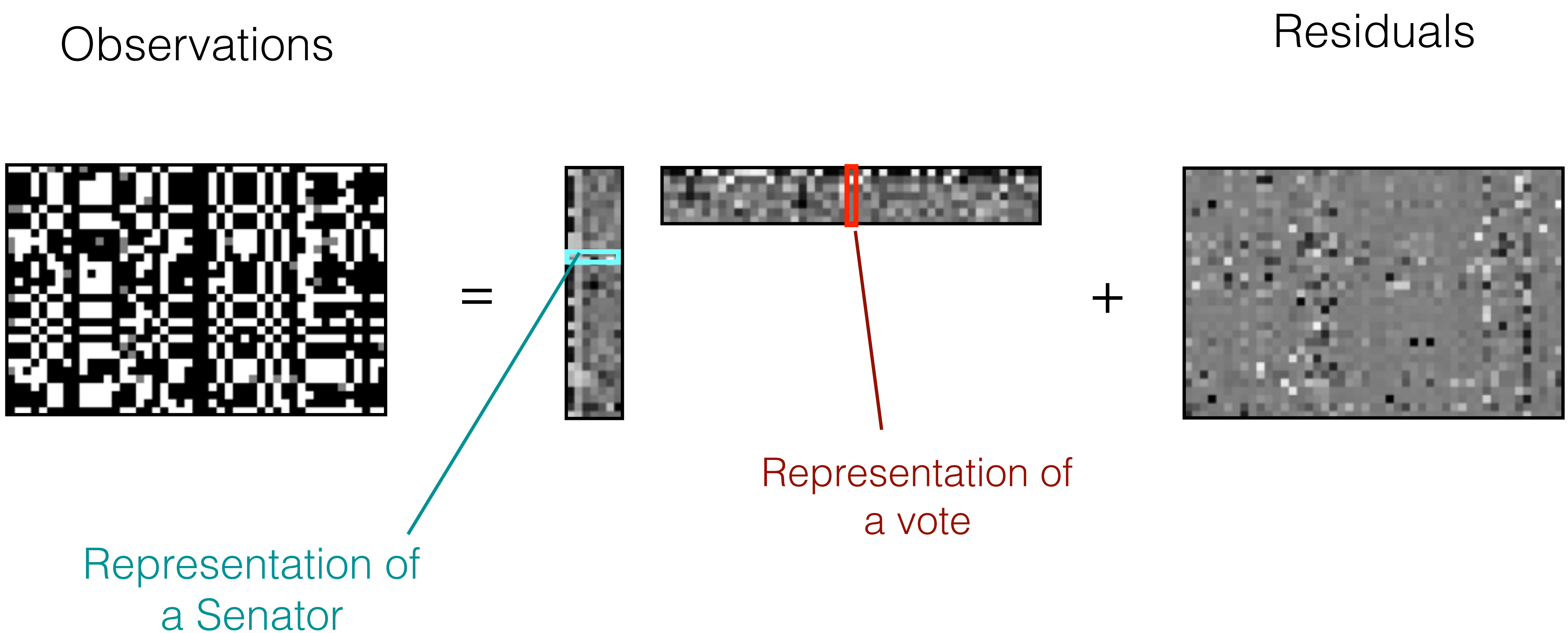
Matrix decompositions

Clustering the votes



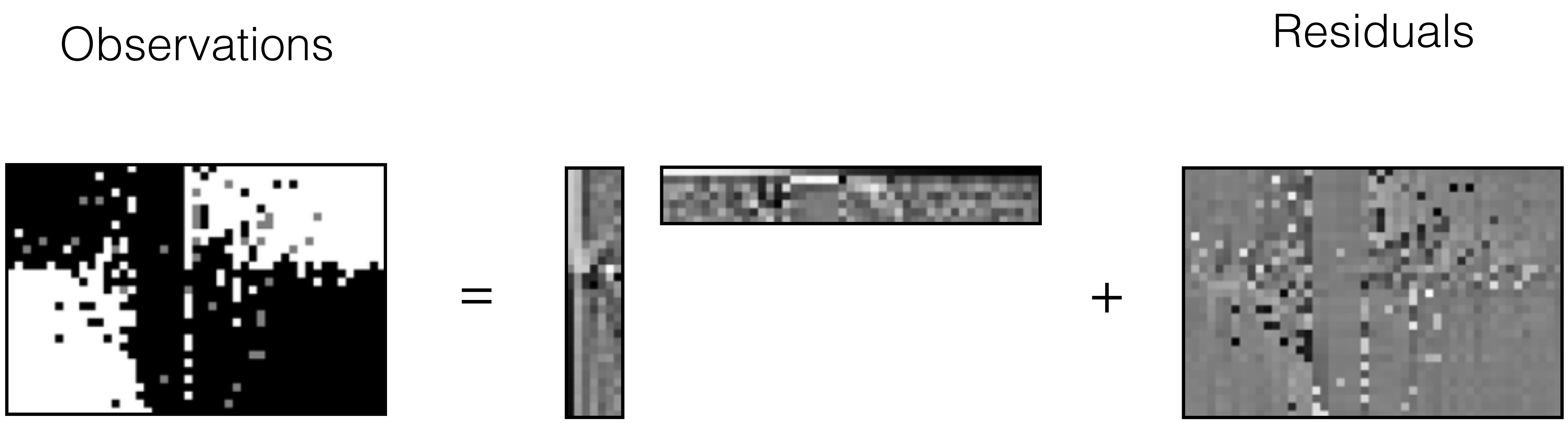
Matrix decompositions

Dimensionality reduction



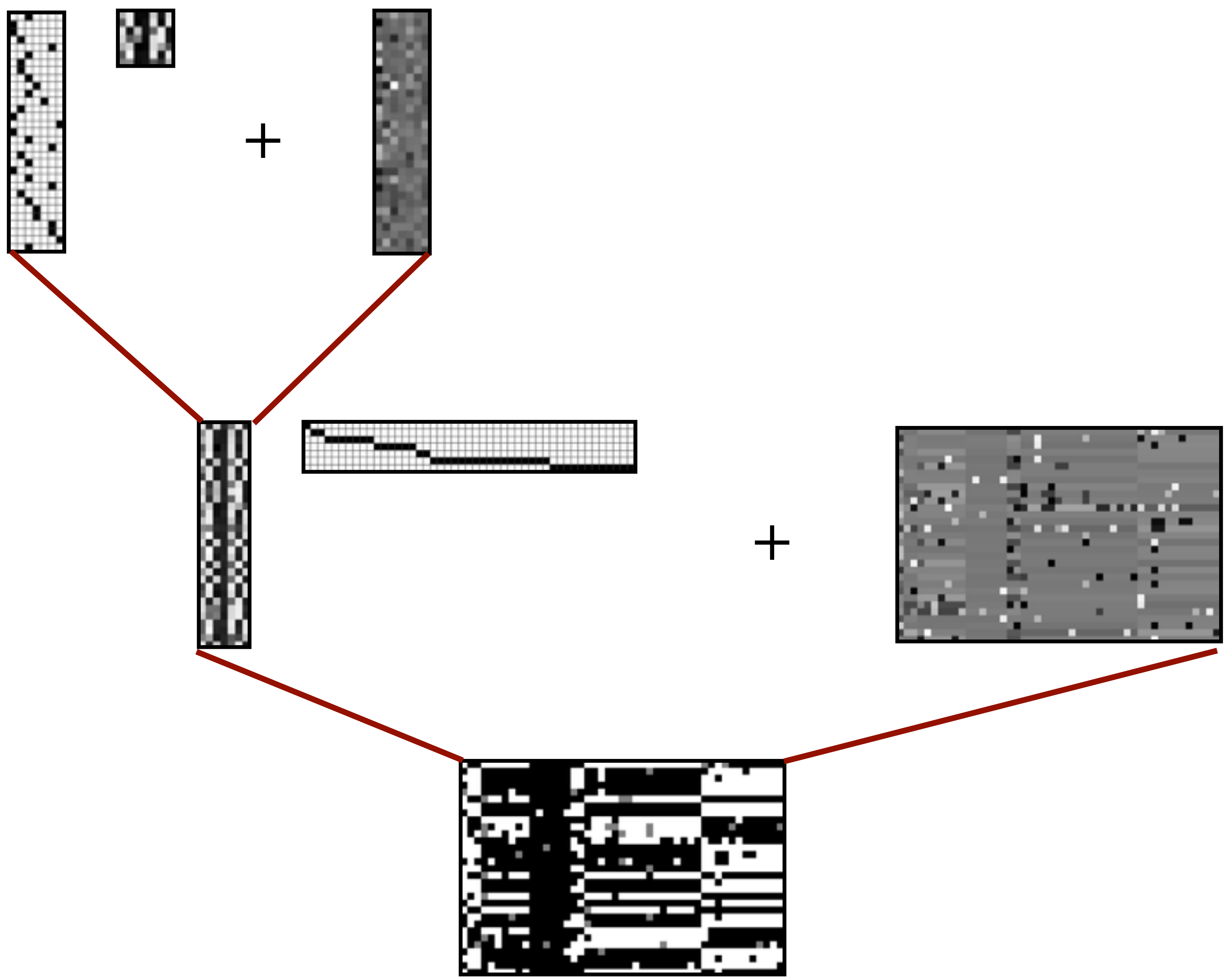
Matrix decompositions

Dimensionality reduction



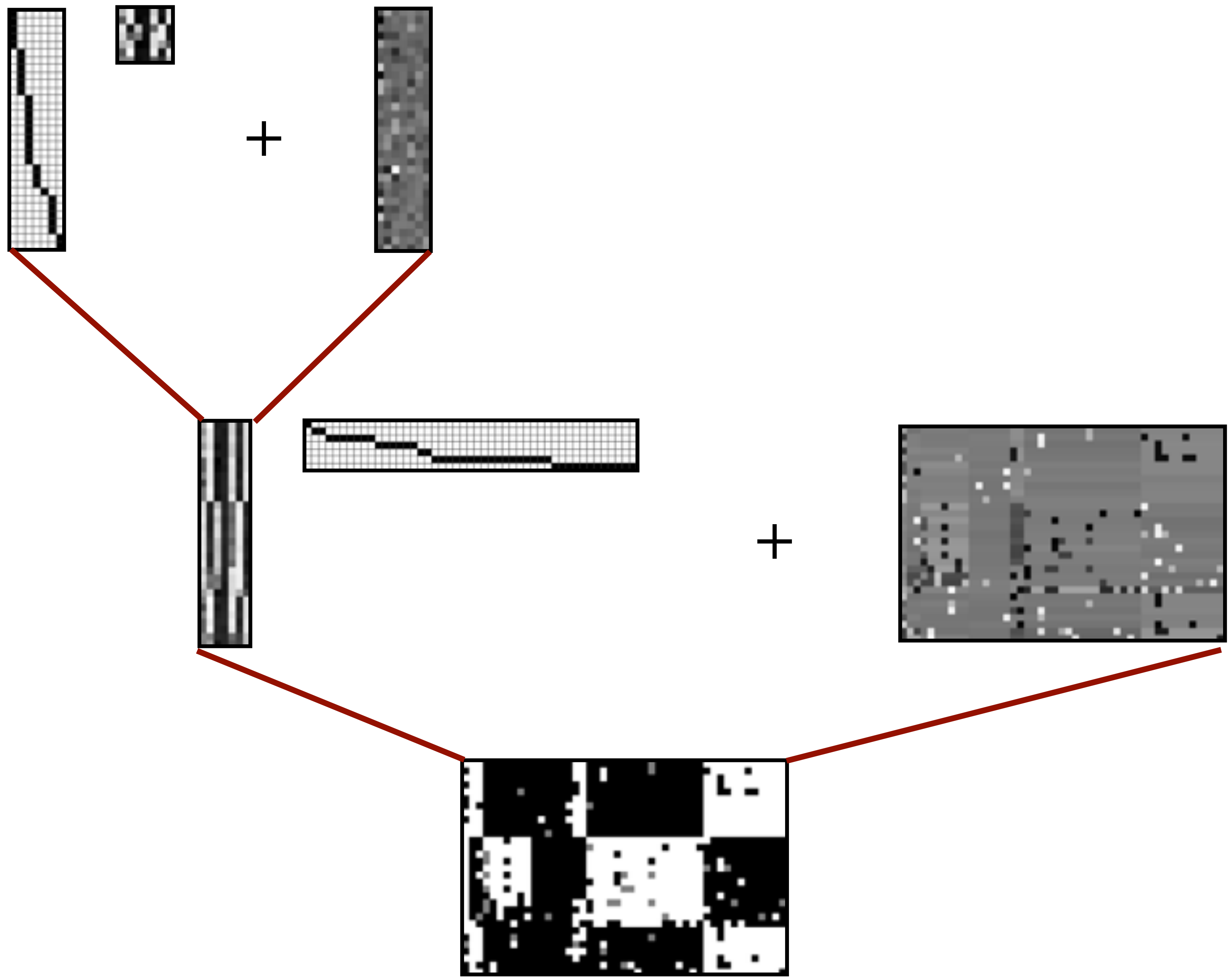
Matrix decompositions

Co-clustering Senators and Votes

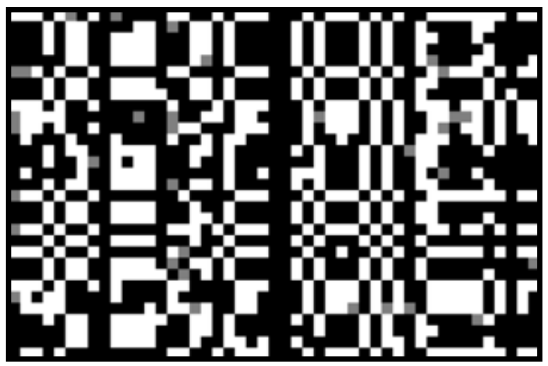


Matrix decompositions

Co-clustering Senators and Votes



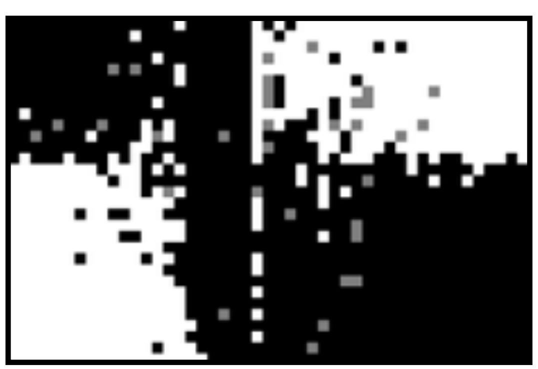
Matrix decompositions



No structure

Cluster columns

Cluster rows

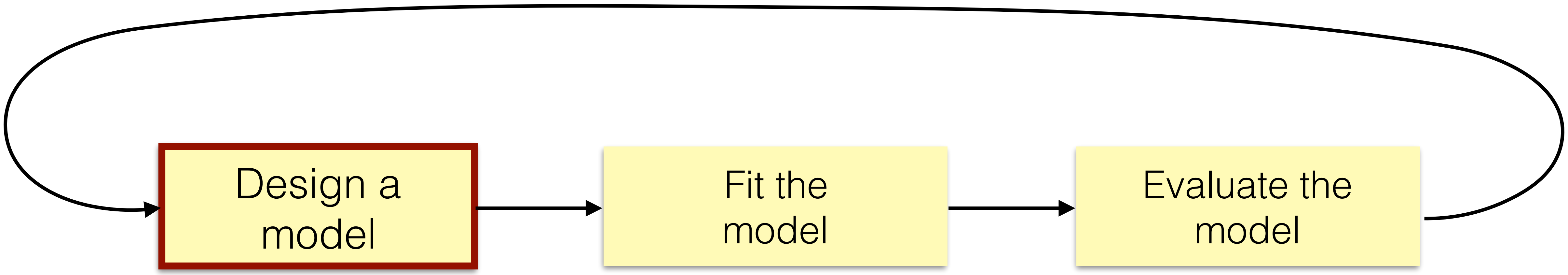


Dimensionality reduction

Co-clustering

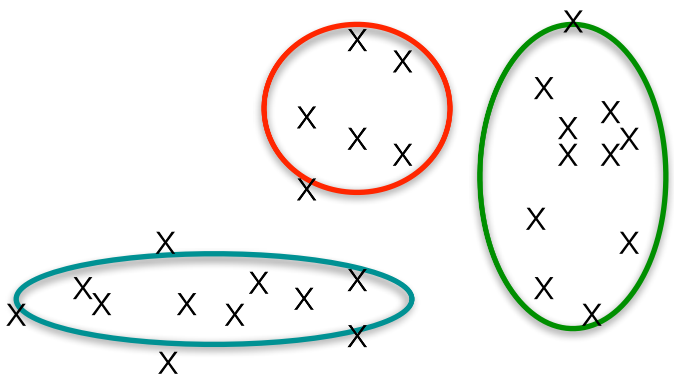
. . .

The probabilistic modeling pipeline

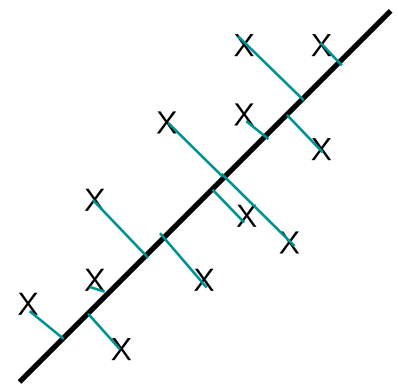


Building models compositionally

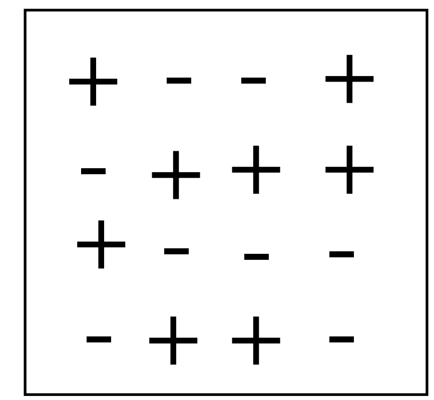
We build models by **composing simpler motifs**



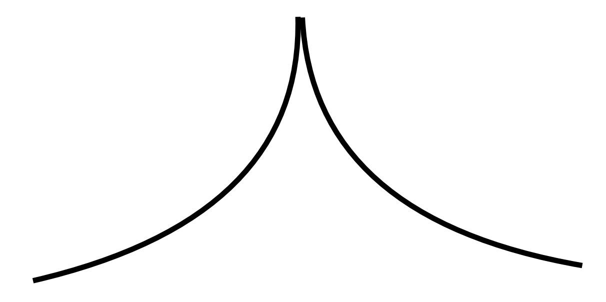
Clustering



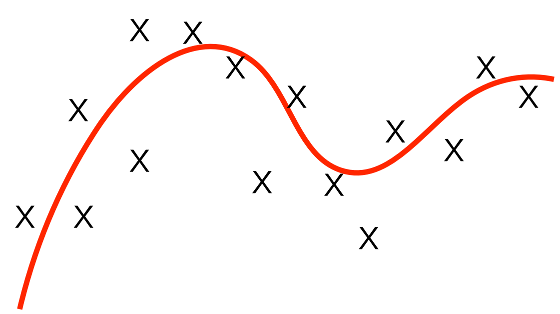
Dimensionality reduction



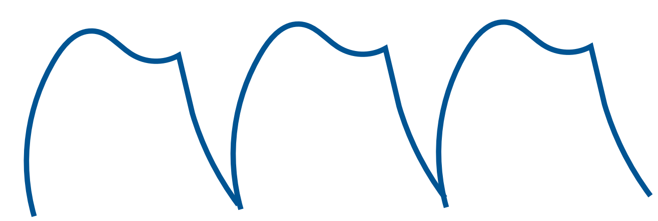
Binary attributes



Heavy-tailed distributions

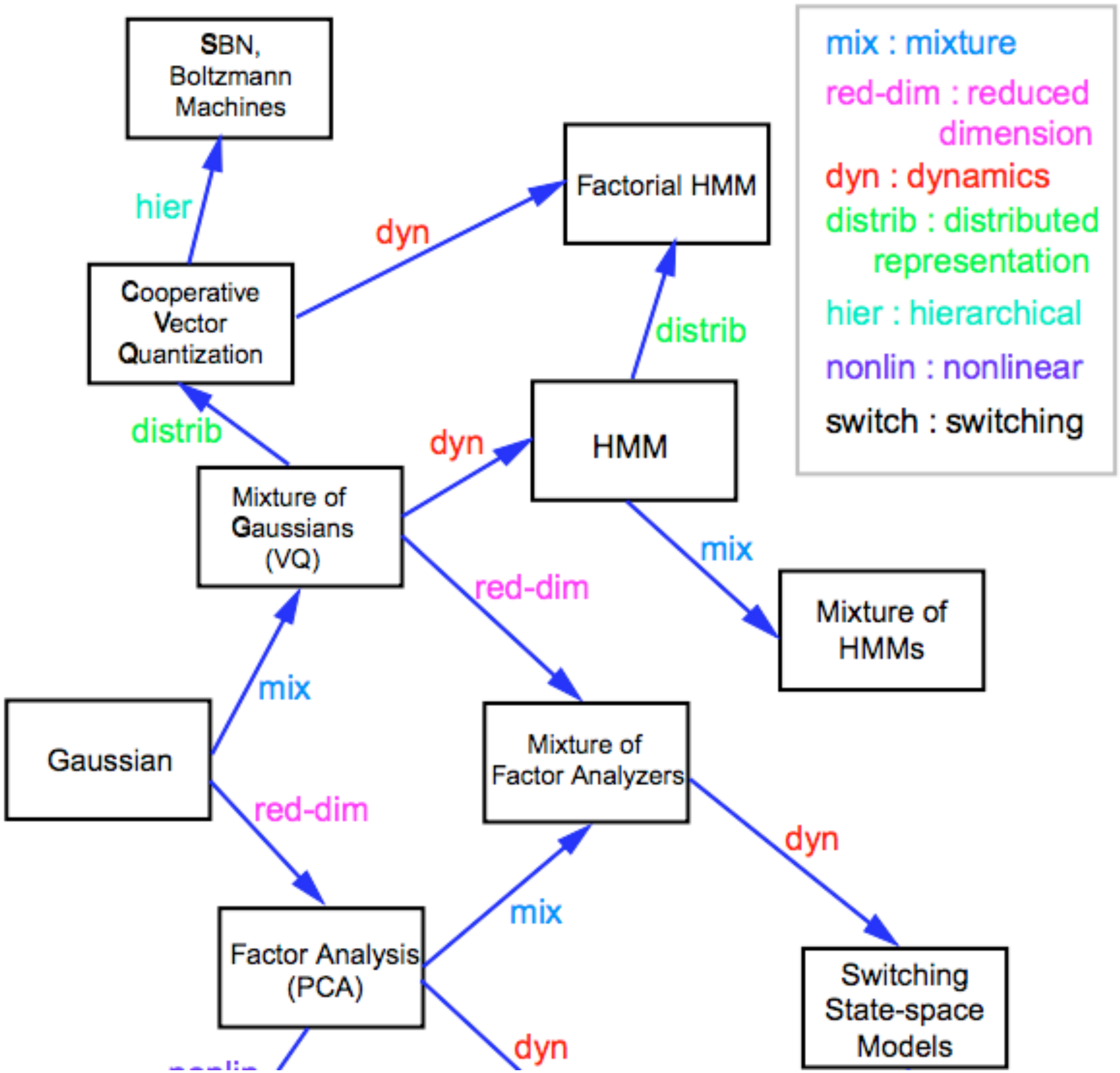


Smoothness



Periodicity

Building models compositionally



(Ghahramani, 1999 NIPS tutorial)

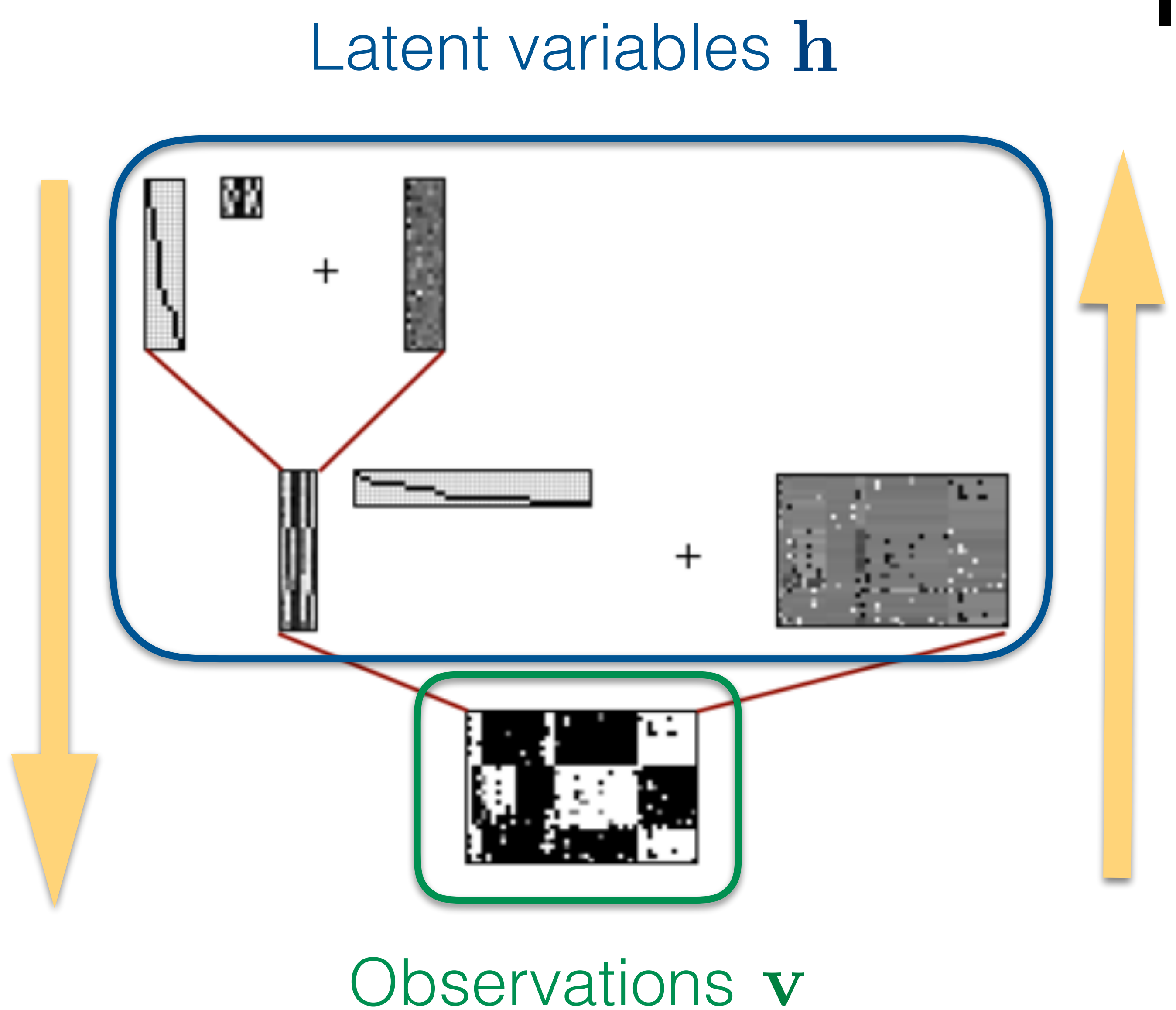
Generative models

Generation

Tell a story of how datasets get generated

This gives a joint probability distribution over observations and latent variables

$$p(\mathbf{h}, \mathbf{v}) = p(\mathbf{h})p(\mathbf{v}|\mathbf{h})$$



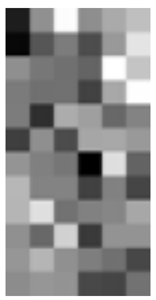
Posterior Inference

Infer a good explanation of how a particular dataset was generated

Find likely values of the latent variables conditioned on the observations

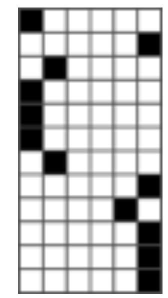
$$p(\mathbf{h}|\mathbf{v})$$

Space of models: building blocks



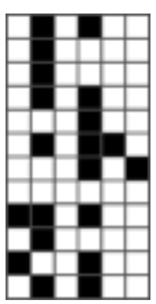
$$\lambda_i \sim \mathrm{Gamma}(a, b)$$
$$\nu_j \sim \mathrm{Gamma}(a, b)$$
$$u_{ij} \sim \mathrm{Normal}(0, \lambda_i^{-1}\nu_j^{-1})$$

Gaussian
(G)



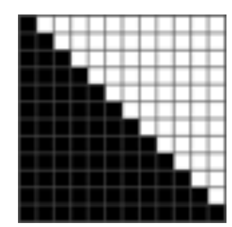
$$\pi \sim \mathrm{Dirichlet}(\alpha)$$
$$u_i \sim \mathrm{Multinomial}(\pi)$$

Multinomial
(M)



$$p_j \sim \mathrm{Beta}(\alpha, \beta)$$
$$u_{ij} \sim \mathrm{Bernoulli}(p_j)$$

Bernoulli
(B)



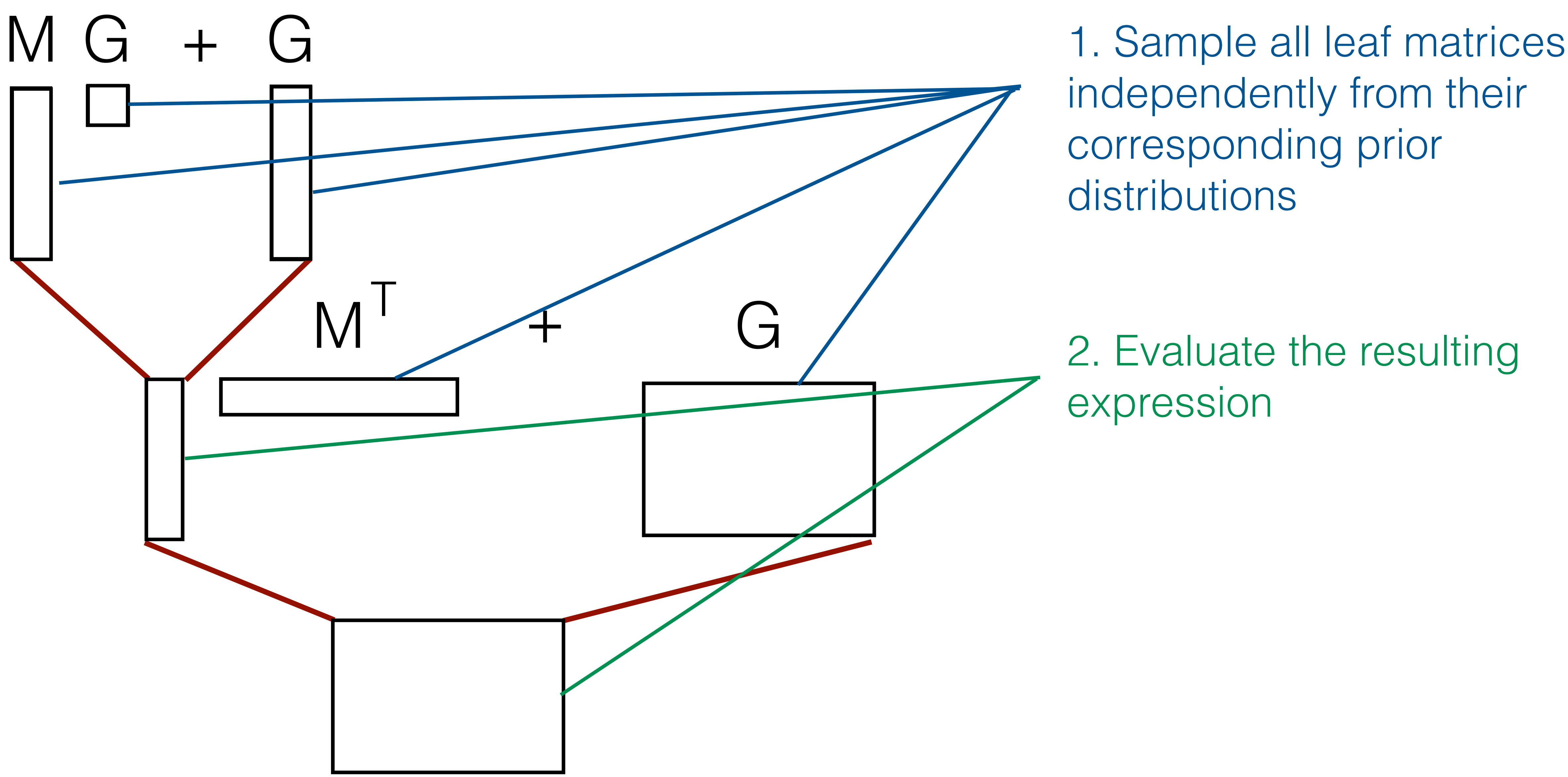
$$u_{ij} = \begin{cases} 1 & \text{if } i \geq j \\ 0 & \text{otherwise} \end{cases}$$

Integration
(C)

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

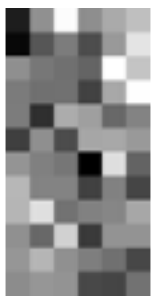
Space of models: generative process

We represent models as algebraic expressions.

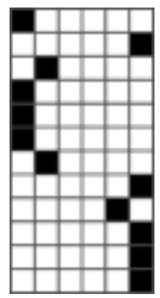


Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

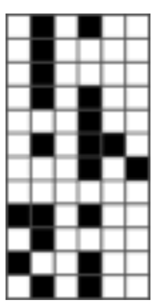
Space of models: grammar



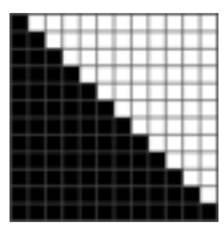
Gaussian
(G)



Multinomial
(M)



Bernoulli
(B)



Integration
(C)

Starting symbol: G

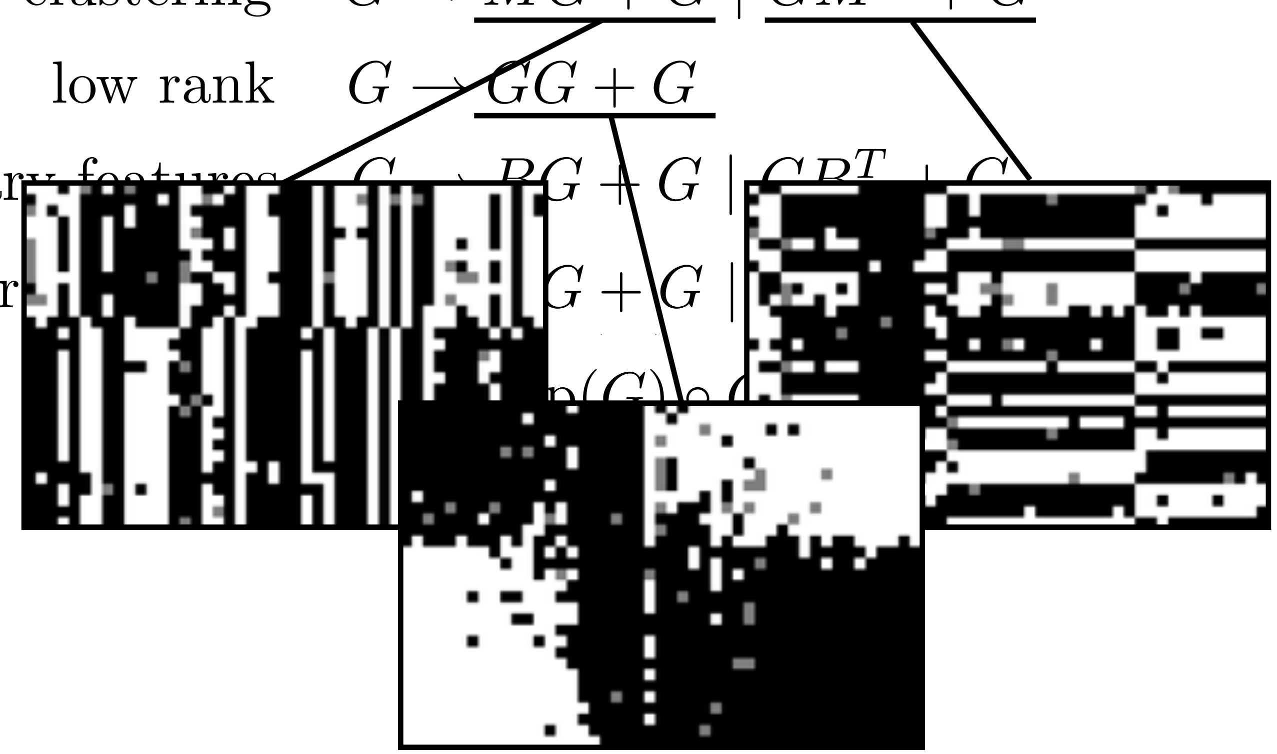
Production rules:

clustering $G \rightarrow \underline{MG + G} \mid \underline{GM^T + G}$

low rank $G \rightarrow \underline{GG + G}$

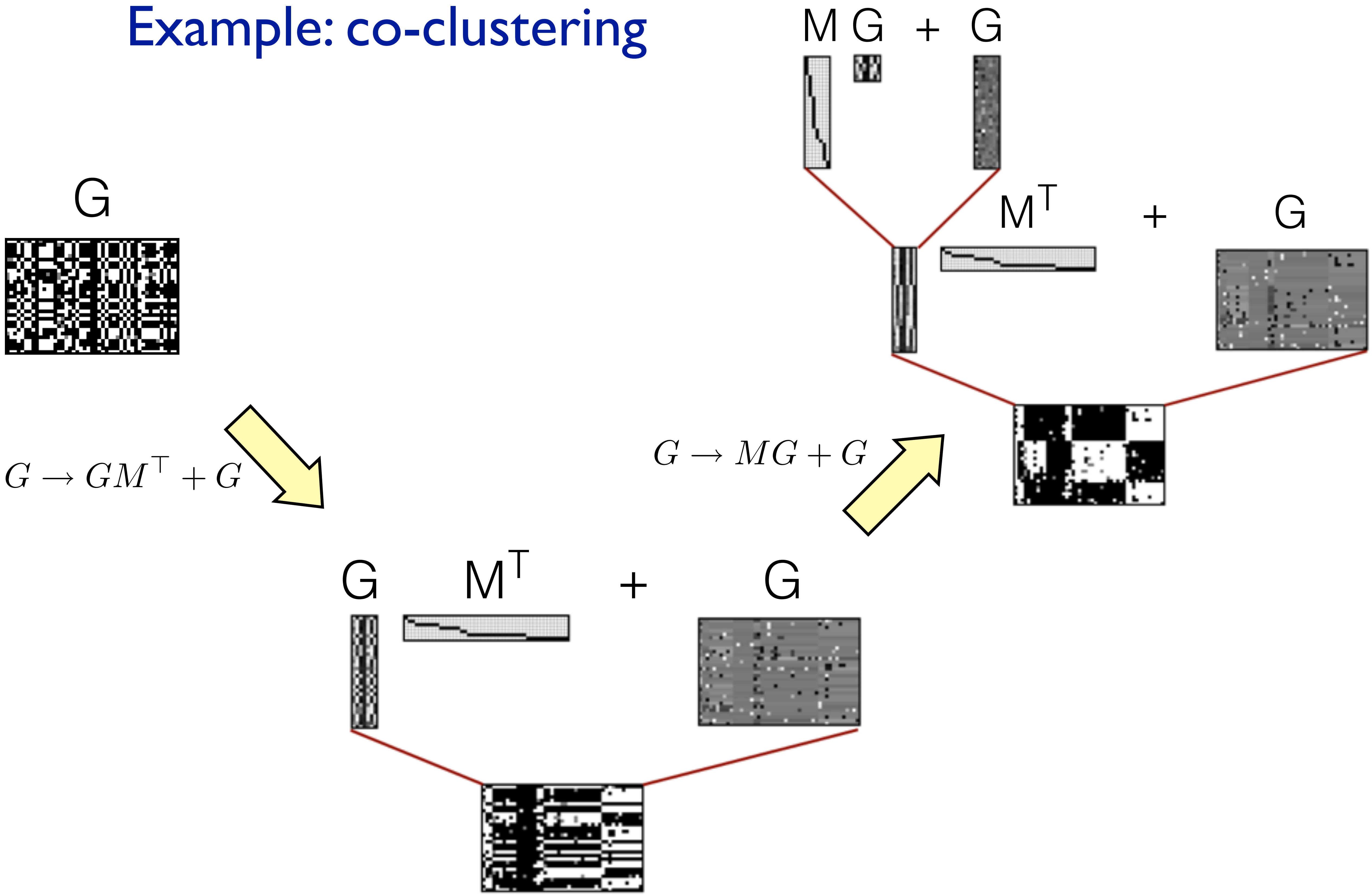
binary features $G \rightarrow BG + G \mid GB^T + G$

linear ... $G + G \mid$ ...



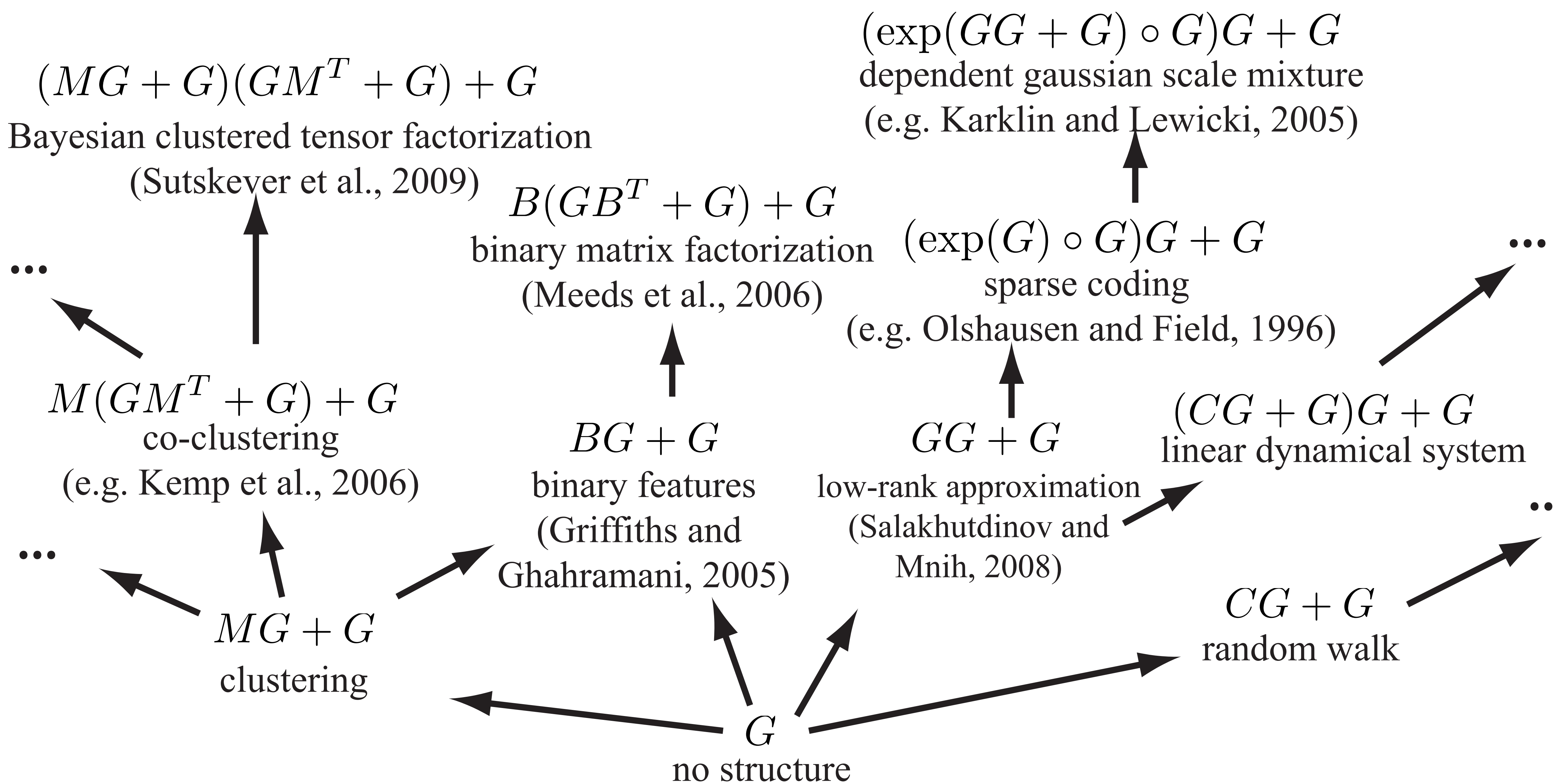
Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Example: co-clustering



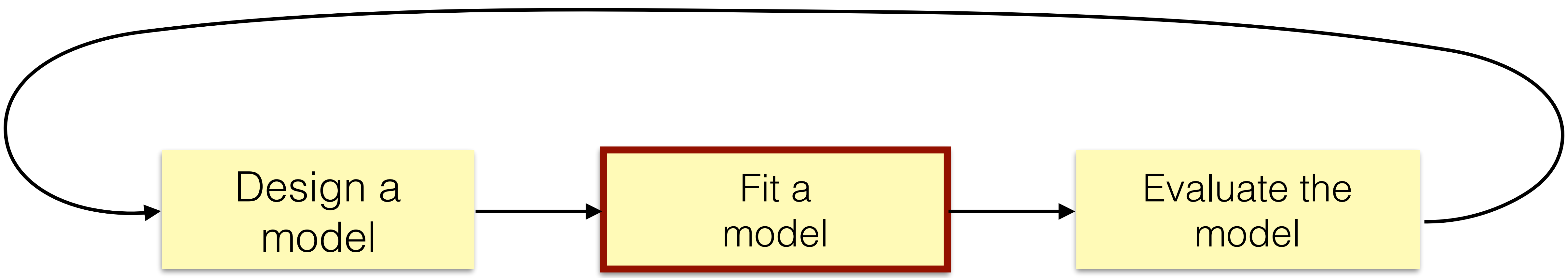
Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

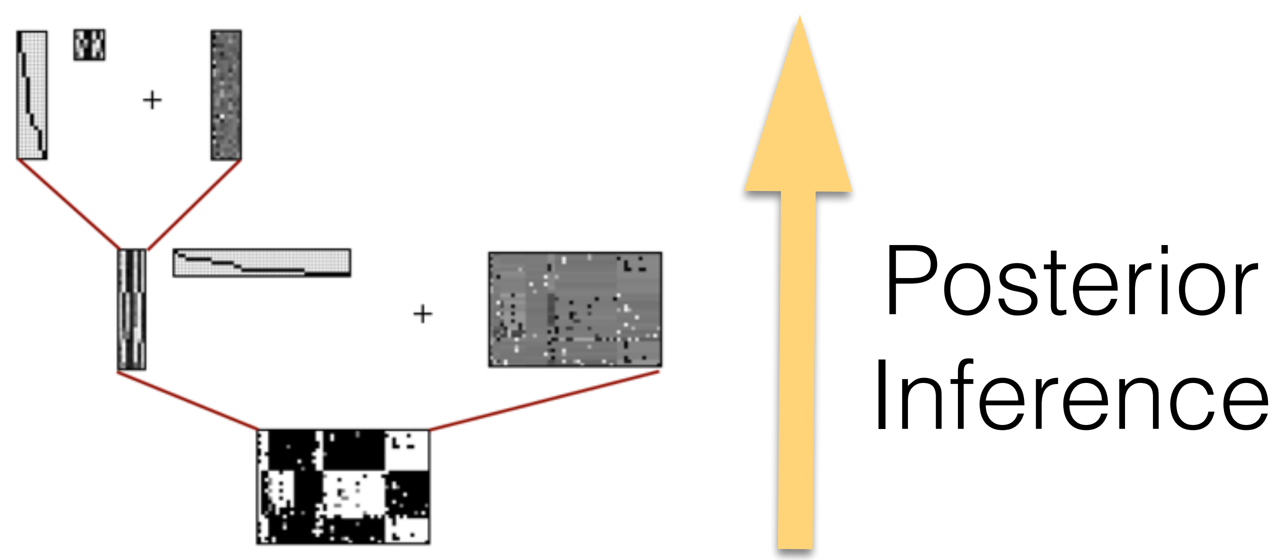
Examples from the literature



Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

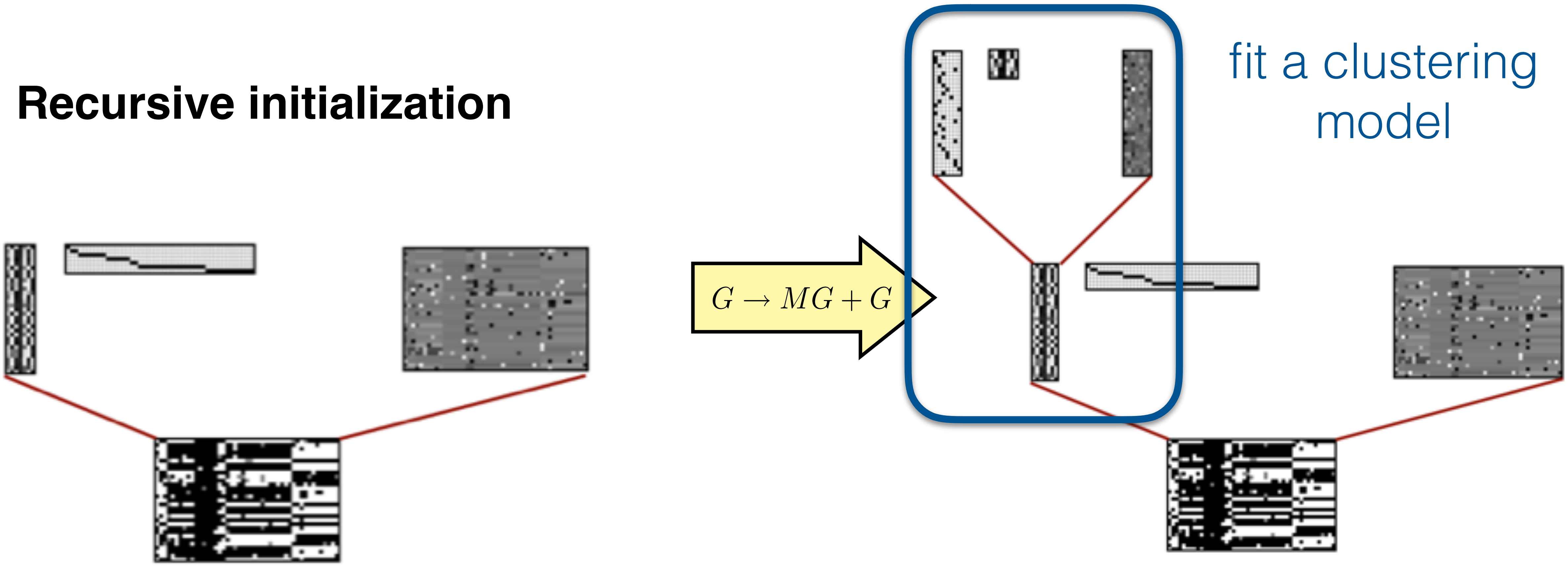
The probabilistic modeling pipeline





Algorithms: posterior inference

Recursive initialization



implement one algorithm per production rule

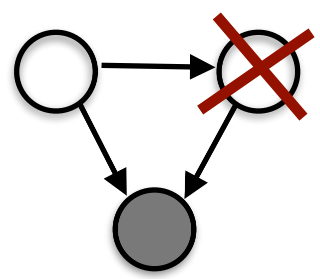
share computation between models

Choose the model dimension using Bayesian nonparametrics

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Posterior inference algorithms

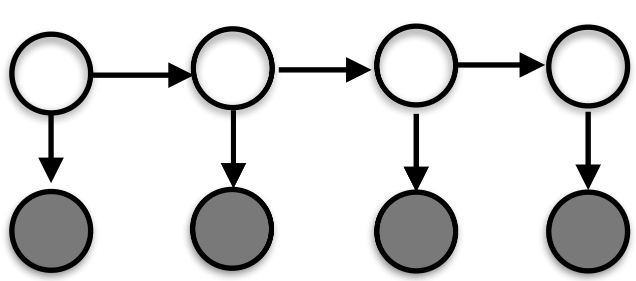
Can make use of **model-specific algorithmic tricks** carefully designed for **individual production rules**:



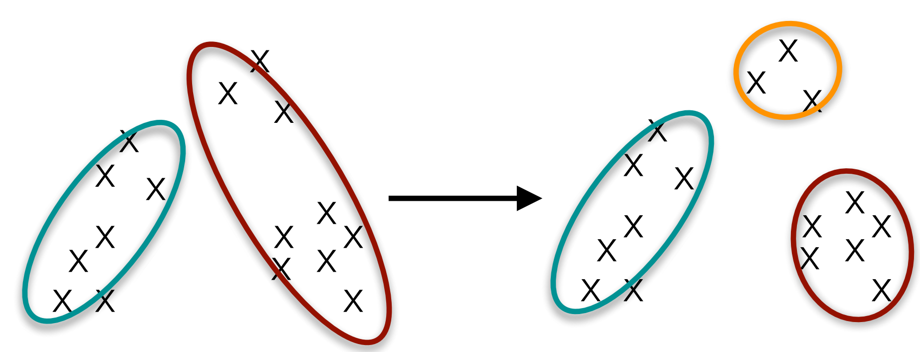
Eliminating variables analytically

$$(A + UCV)^{-1} =$$
$$A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Linear algebra identities

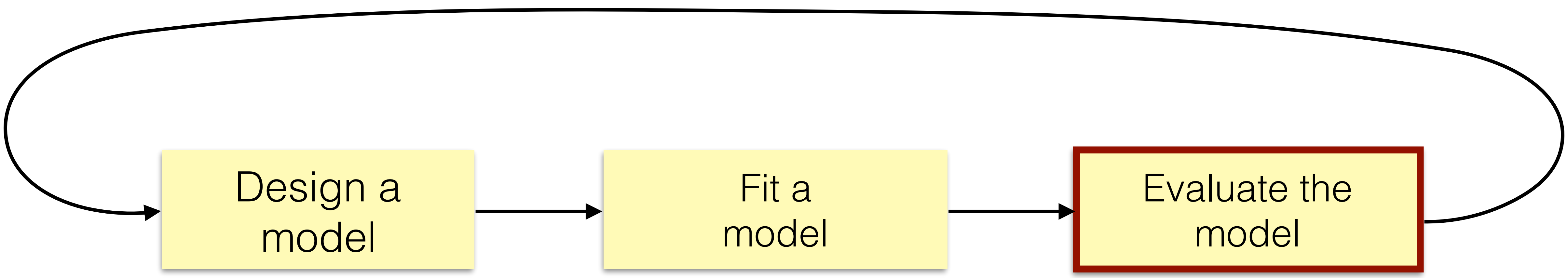


tractable substructures



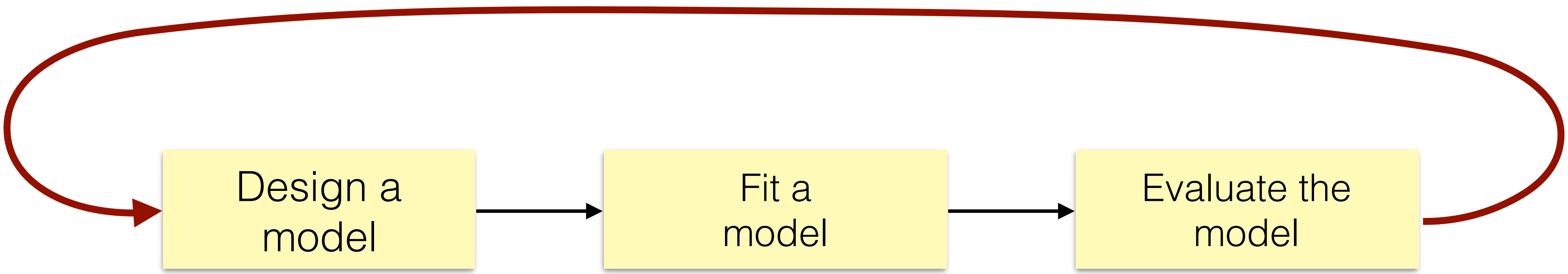
High-level transition operators

The probabilistic modeling pipeline



We evaluate models on the probability they assign to held-out subsets of the observation matrix.

The probabilistic modeling pipeline



Want to search over the large, open-ended space of models

Key problem: the search space is very large!

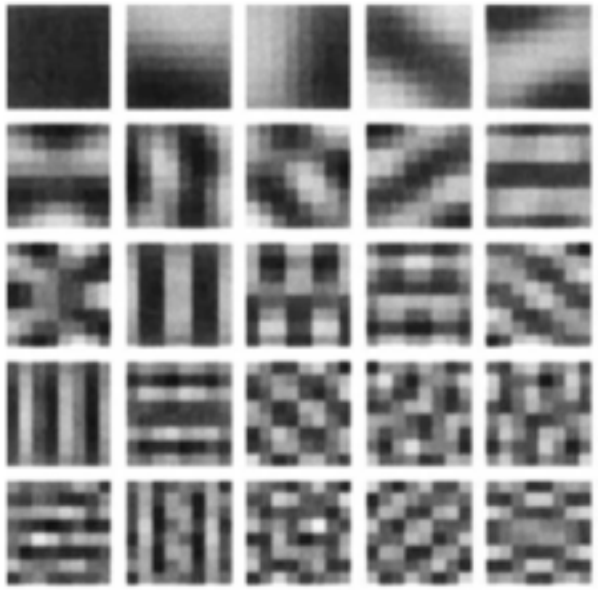
over 1000 models reachable in 3 productions

how to choose a promising set of models to evaluate?

Algorithms: structure search

A brief history of models of natural images…

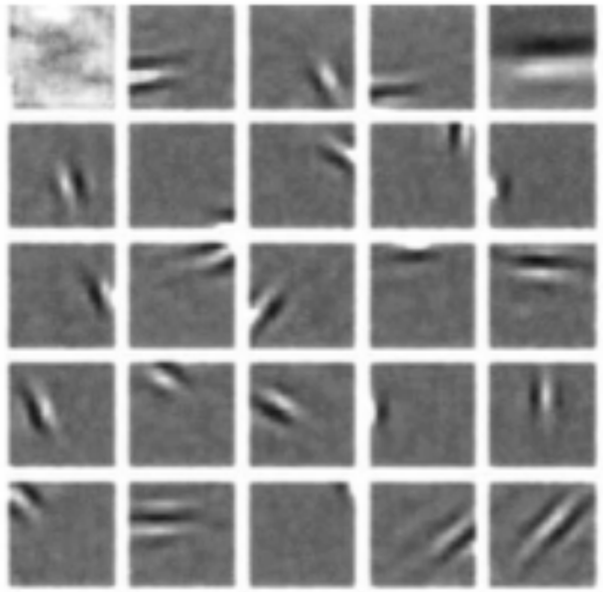
Sanger, 1988



Model patches as linear combinations of uncorrelated basis functions

Fourier representation

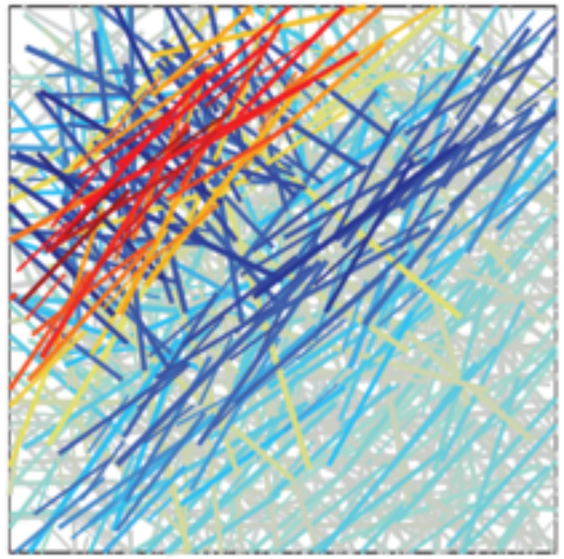
Olshausen and Field, 1994



Model the heavy-tailed distributions of coefficients

oriented edges similar to simple cells

Karklin and Lewicki, 2005, 2008



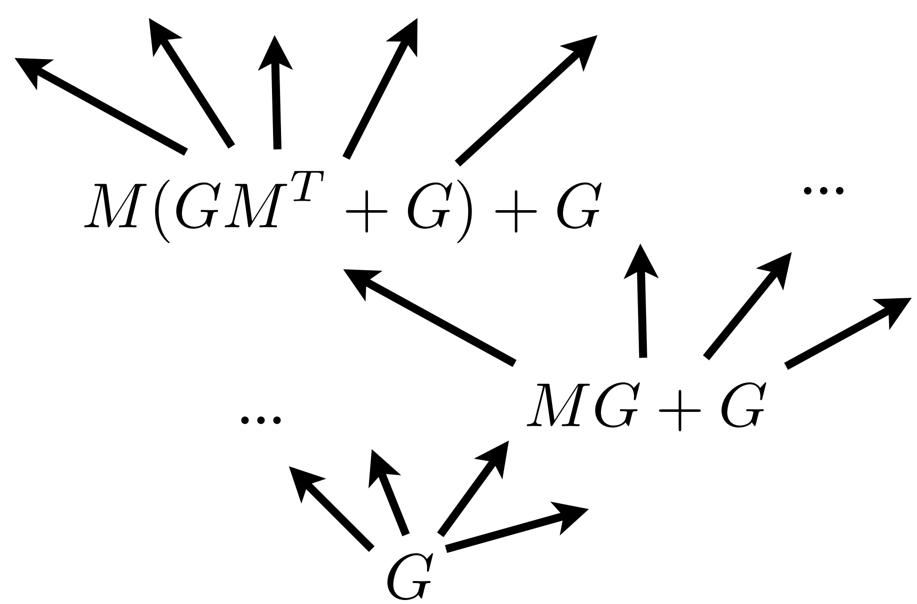
Model the dependencies between scales of coefficients

high-level texture representation similar to complex cells

Algorithms: structure search

Refining models = applying productions

Based on this intuition, we apply a greedy search procedure



Experiments: simulated data

Tested on simulated data where we know the correct structure

— Increasing noise $\longrightarrow$

	$\sigma^2 = 1$	$\sigma^2 = 3$	$\sigma^2 = 10$
low-rank			
clustering			
binary latent features			
co-clustering			
binary matrix factorization			
BCTF			
sparse coding			
dependent GSM			
random walk			
linear dynamical system			

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Experiments: simulated data

Tested on simulated data where we know the correct structure



— Increasing noise ⟶

	$\sigma^2 = 1$	$\sigma^2 = 3$	$\sigma^2 = 10$
low-rank	$GG + G$		
clustering	$MG + G$		
binary latent features	$BG + G$		
co-clustering	$M(GM^T + G) + G$		
binary matrix factorization	$(BG + G)B^T + G$		
BCTF	$(MG + G)(GM^T + G) + G$		
sparse coding	$(\exp(G) \circ G)G + G$		
dependent GSM	$(\exp(G) \circ G)G + G$		
random walk	$CG + G$		
linear dynamical system	$(CG + G)G + G$		

Usually chooses the correct structure in low-noise conditions

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Experiments: simulated data

Tested on simulated data where we know the correct structure

— Increasing noise ⟶

	$\sigma^2 = 1$	$\sigma^2 = 3$	$\sigma^2 = 10$
low-rank	$GG + G$	$GG + G$	① G
clustering	$MG + G$	$MG + G$	$MG + G$
binary latent features	$BG + G$	$BG + G$	$BG + G$
co-clustering	$M(GM^T + G) + G$	$M(GM^T + G) + G$	① $GM^T + G$
binary matrix factorization	$(BG + G)B^T + G$	② $GG + G$	② $GG + G$
BCTF	$(MG + G)(GM^T + G) + G$	② $GM^T + G$	③ G
sparse coding	$(\exp(G) \circ G)G + G$	$(\exp(G) \circ G)G + G$	② G
dependent GSM	① $(\exp(G) \circ G)G + G$	① $(\exp(G) \circ G)G + G$	③ $BG + G$
random walk	$CG + G$	$CG + G$	① G
linear dynamical system	$(CG + G)G + G$	$(CG + G)G + G$	② $BG + G$

Usually chooses the correct structure in low-noise conditions

Gracefully falls back to simpler models under heavy noise

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Experiments: real-world data

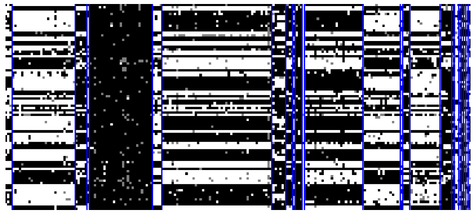
Senate votes 09-10 $GM^T + G$ $(MG + G)M^T + G$ —

Cluster votes.

22 clusters

largest: party line Democrat, party line Republican, all yea

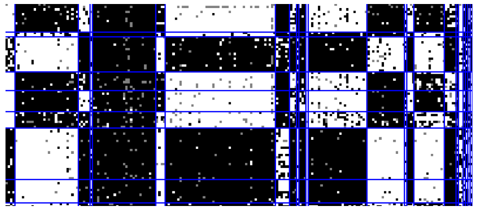
others are series of votes on single issues



Cluster Senators.

11 clusters

no cross-party clusters



No third level model improves by more than 1 nat

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Experiments: real-world data

Senate votes 09-10	$GM^T + G$	$(MG + G)M^T + G$	—
Motion capture	$CG + G$	$C(GG + G) + G$	—

Model 1: Independent Markov chains

Model 2: Correlations in joint angles

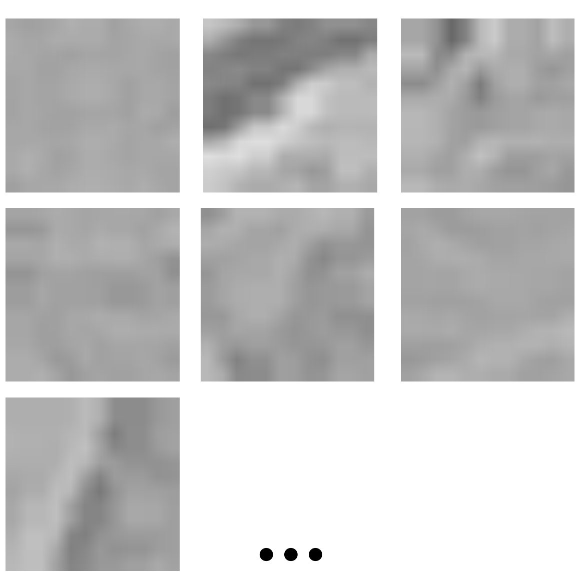
Data: motion capture of a person walking. Each row gives a person's displacement and joint angles in one frame.

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Experiments: real-world data

Senate votes 09-10	$GM^T + G$	$(MG + G)M^T + G$	—
Motion capture	$CG + G$	$C(GG + G) + G$	—
Image patches	$GG + G$	$(\exp(G) \circ G)G + G$	$(\exp(GG + G) \circ G)G + G$

Data: 1,000 12x12 patches from 10 blurred and whitened images.



...

Model 1: Low-rank approximation (PCA).

Model 2: Sparsify coefficients to get sparse coding

Model 3: Model dependencies between scale variables

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Experiments: real-world data

Senate votes 09-10	$GM^T + G$	$(MG + G)M^T + G$	—
Motion capture	$CG + G$	$C(GG + G) + G$	—
Image patches	$GG + G$	$(\exp(G) \circ G)G + G$	$(\exp(GG + G) \circ G)G + G$
Concepts	$MG + G$	$M(GG + G) + G$	—

Data: Mechanical Turk users' judgments to 218 questions about 1000 entities

Model 1:
Cluster entities.

39 clusters

Model 2:
Low-rank representation of cluster centers.

8 dimensions

Dimension 1: living vs. nonliving

Dimension 2: large vs. small

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

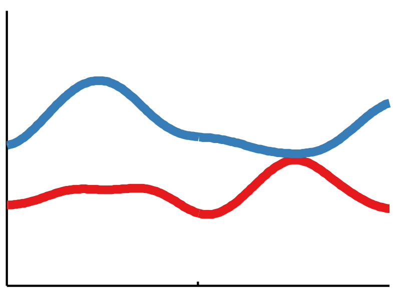
“Structure discovery in nonparametric regression through compositional kernel search,” ICML 2013.

David Duvenaud, James Lloyd, Roger Grosse,
Josh Tenenbaum, and Zoubin Ghahramani,

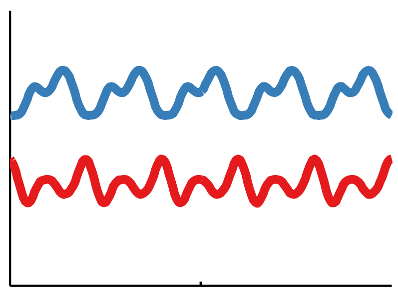
Compositional structure search for time series

Gaussian processes are distributions over functions, specified by kernels.

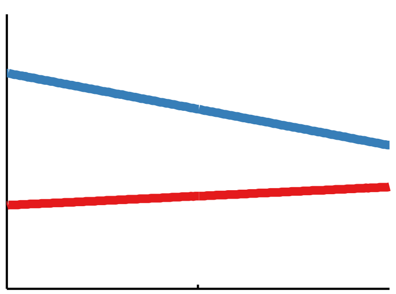
Primitive kernels:



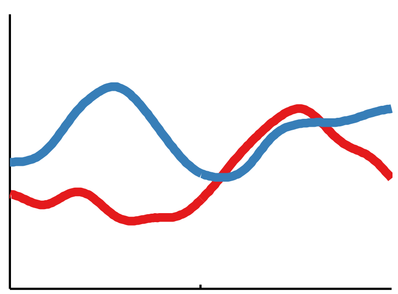
SE



PER

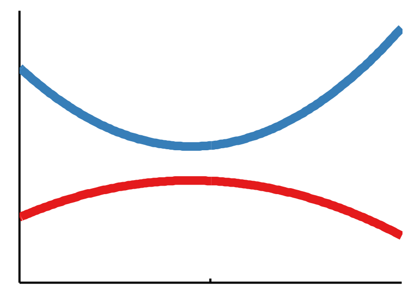


LIN

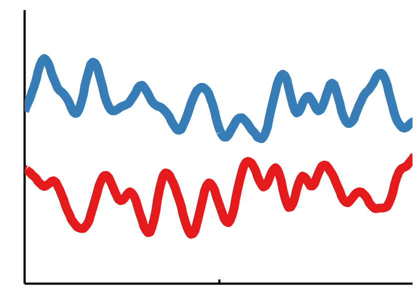


RQ

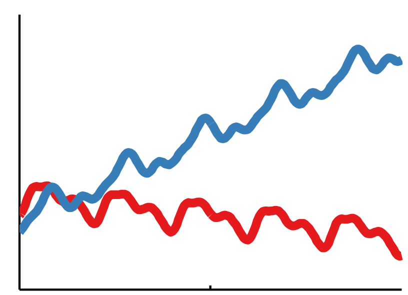
Composite kernels:



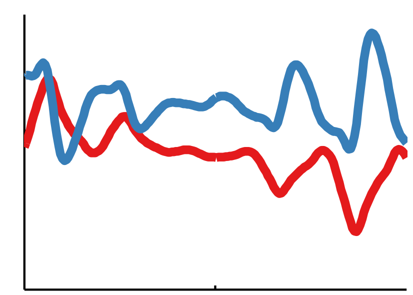
LIN $\times$ LIN



SE $\times$ PER

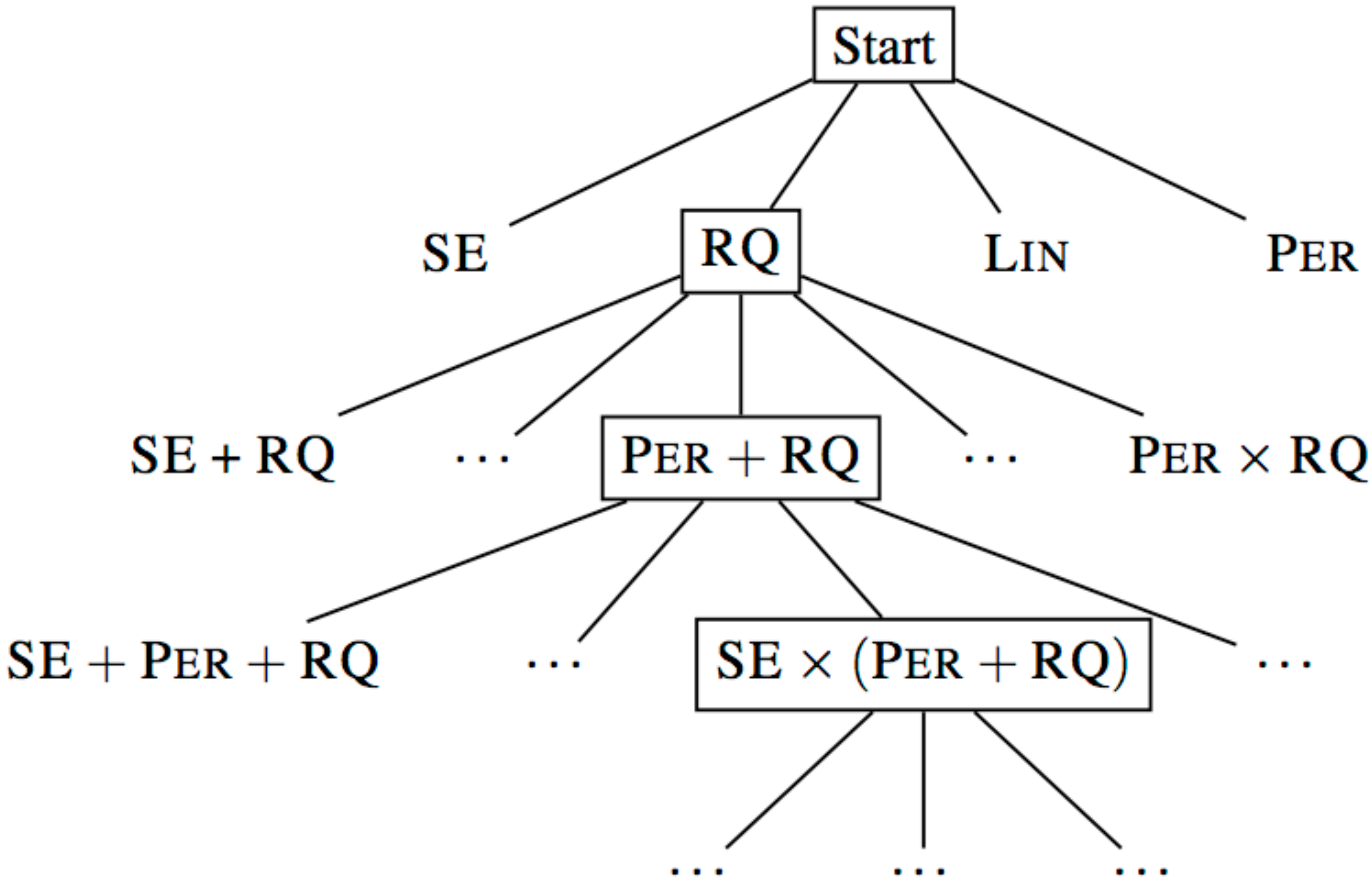


LIN $+$ PER



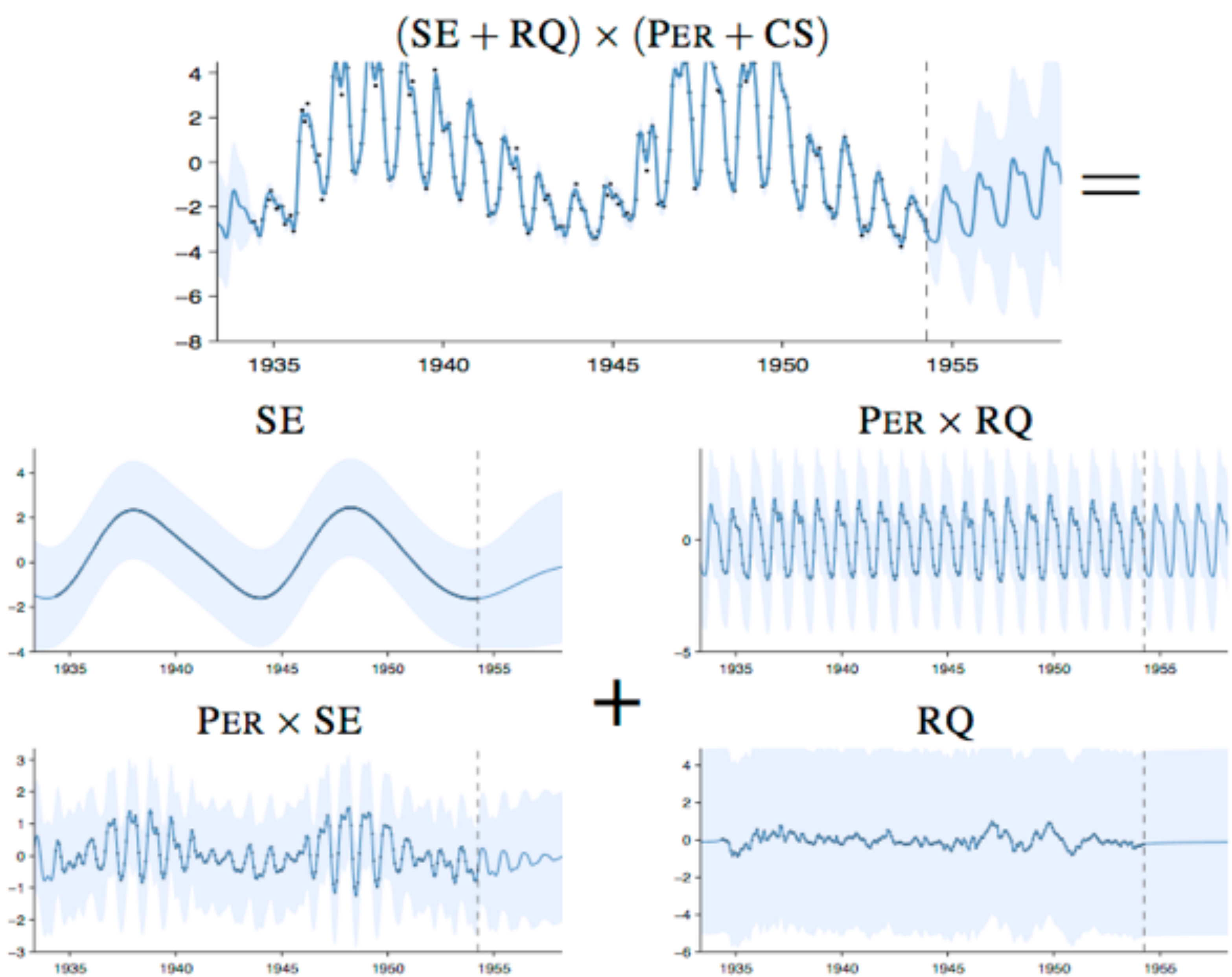
LIN $\times$ PER

Compositional structure search for time series



Compositional structure search for time series

radio critical frequency



An automatic report for the dataset : 01-airline

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

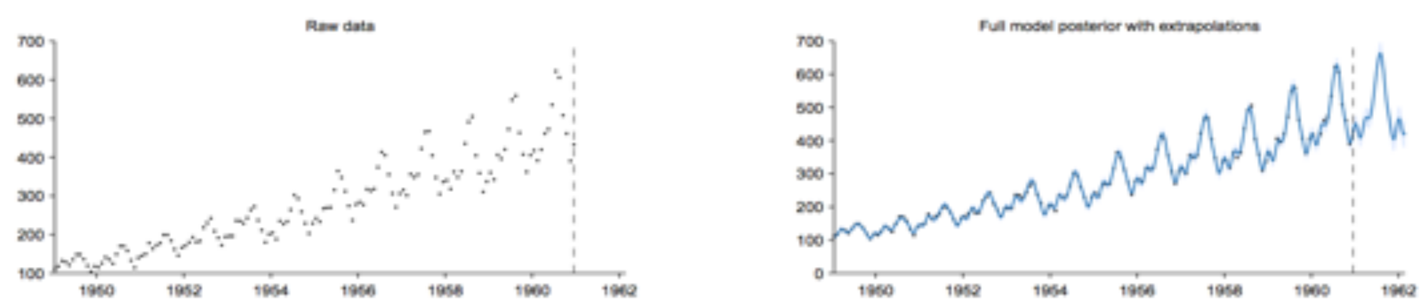
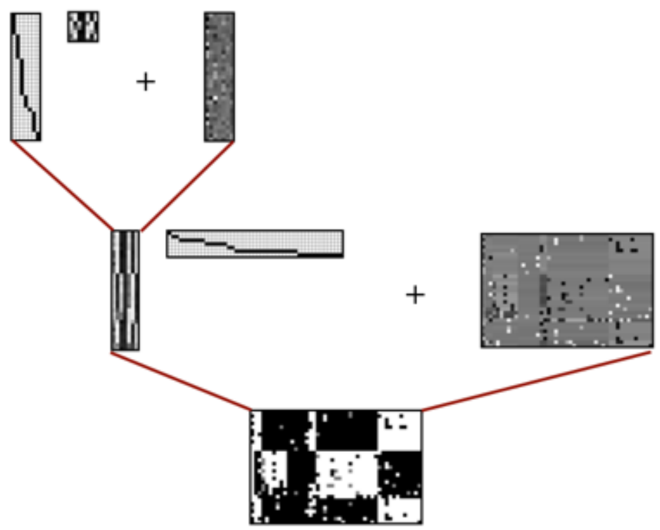
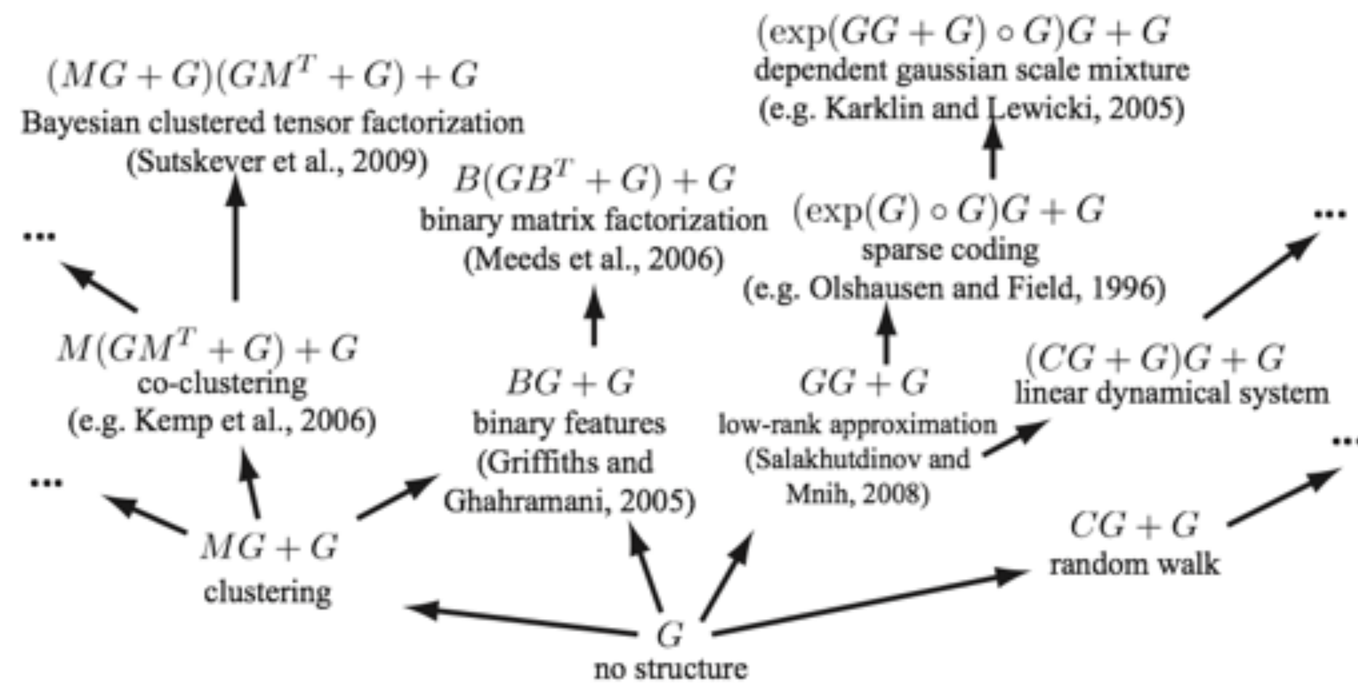


Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.8% of the variation in the data. After the first 3 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

. . .





10 minute break