WYSS INSTITUTE

Multimodal Clustering of Frogs

Using Gaussian Process and Dirichlet Process Priors

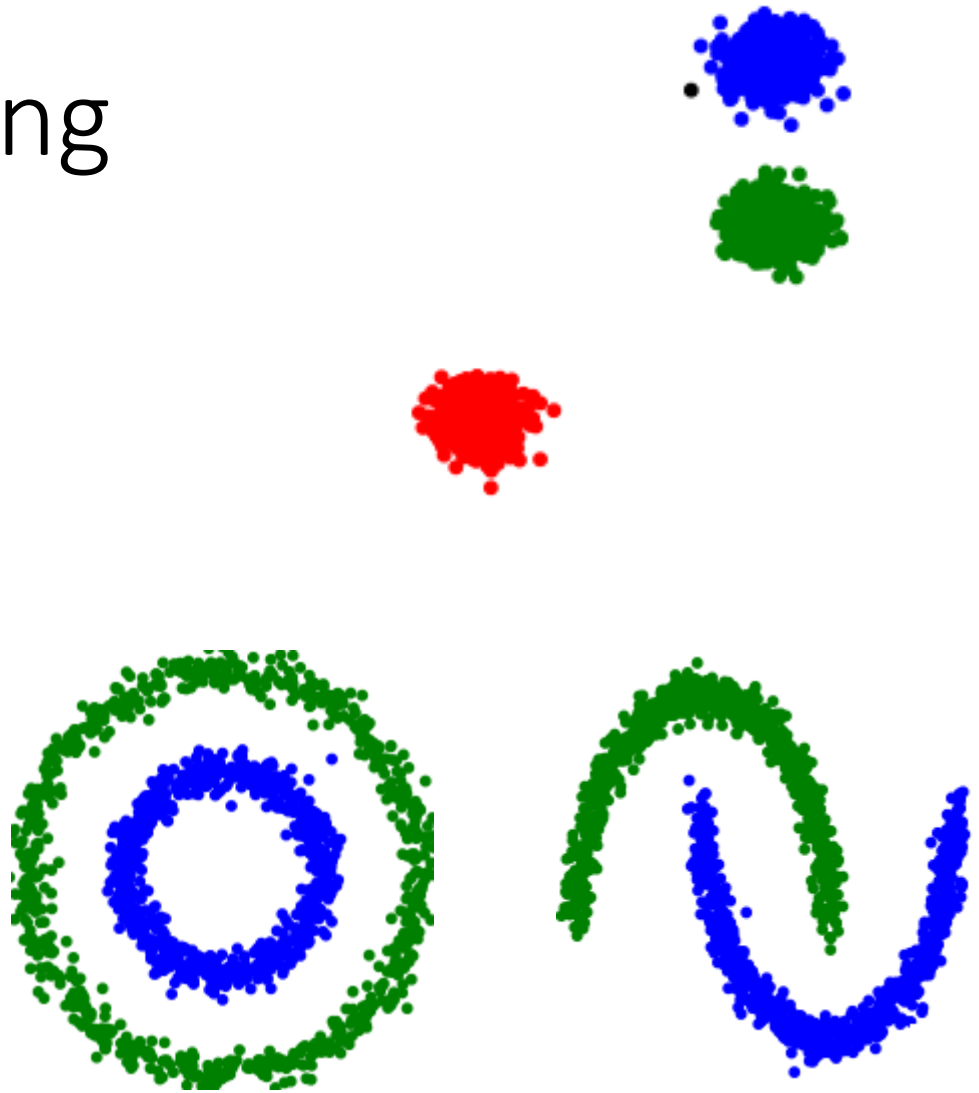
SAHIL LOOMBA

Clustering

To group “objects” by “similarity” in some “space”

Such that objects within one cluster (group) are “closer” to each other than to objects of another cluster

It’s a difficult problem! (even in 2D)



WYSS INSTITUTE

Feature Space

	gene_1	gene_2	…	gene_g
frog_1	6.321	4.287	…	1.432
frog_2	5.009	2.411	…	3.091
…	…	…	…	…
frog_n	4.487	3.932	…	1.254

A g-dimensional feature space for transcriptomics of n frogs

WYSS INSTITUTE

Multiple Feature Spaces in Omics

	gene_1	gene_2	…	gene_g
frog_1				
frog_2				
…				
frog_n				

g dimensional transcriptome

	prot_1	prot_2	…	prot_p
frog_1				
frog_2				
…				
frog_n				

p dimensional proteome

	metb_1	metb_2	…	metb_m
frog_1				
frog_2				
…				
frog_n				

m dimensional metabolome

	micb_1	micb_2	…	micb_b
frog_1				
frog_2				
…				
frog_n				

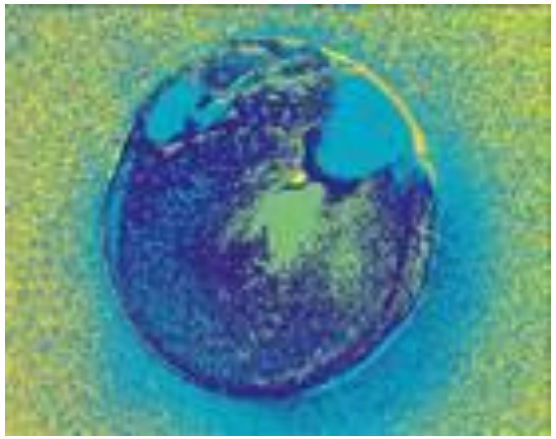
b dimensional microbiome

WYSS INSTITUTE

Multiple Feature Spaces in HSI

	pixl_1	pixl_2	…	pixl_l
frog_1				
frog_2				
…				
frog_n				

`l` dimensional 2D HSI images



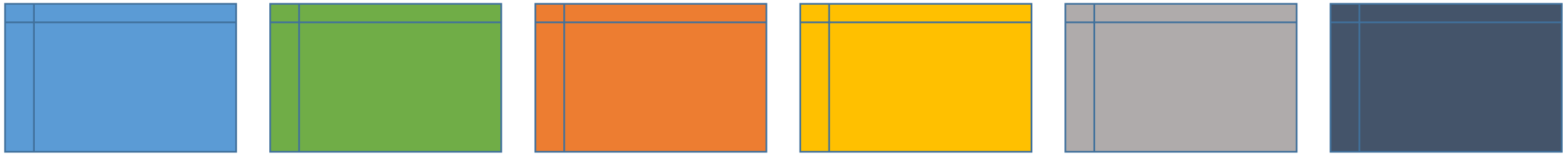
	feat_1	feat_2	…	feat_h
frog_1				
frog_2				
…				
frog_n				

`h` dimensional processed HSI features

```
eccentricity,
convex area,
orientation,
…
```

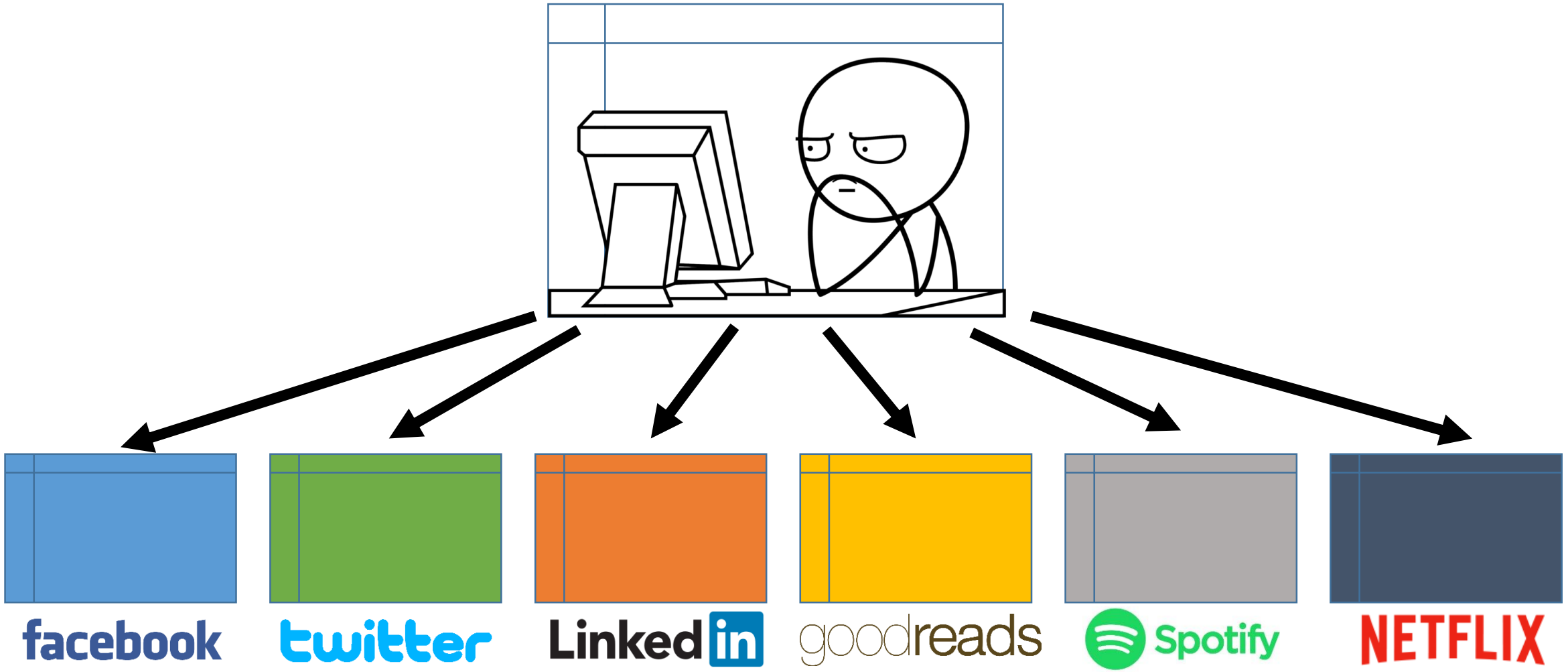
WYSS INSTITUTE

multiple “observed” modalities



WYSS INSTITUTE

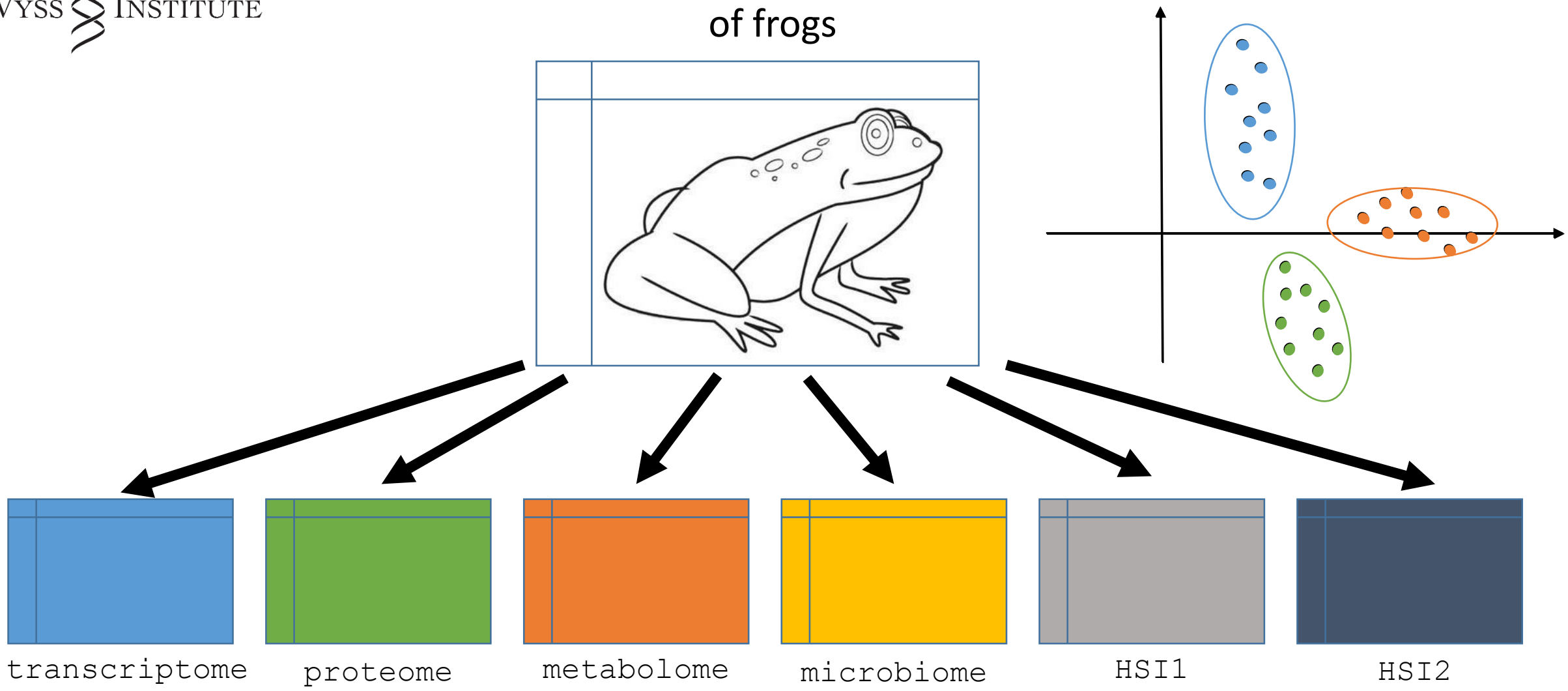
d dimensional latent ("hidden") space of people on the internet



multiple "observed" modalities

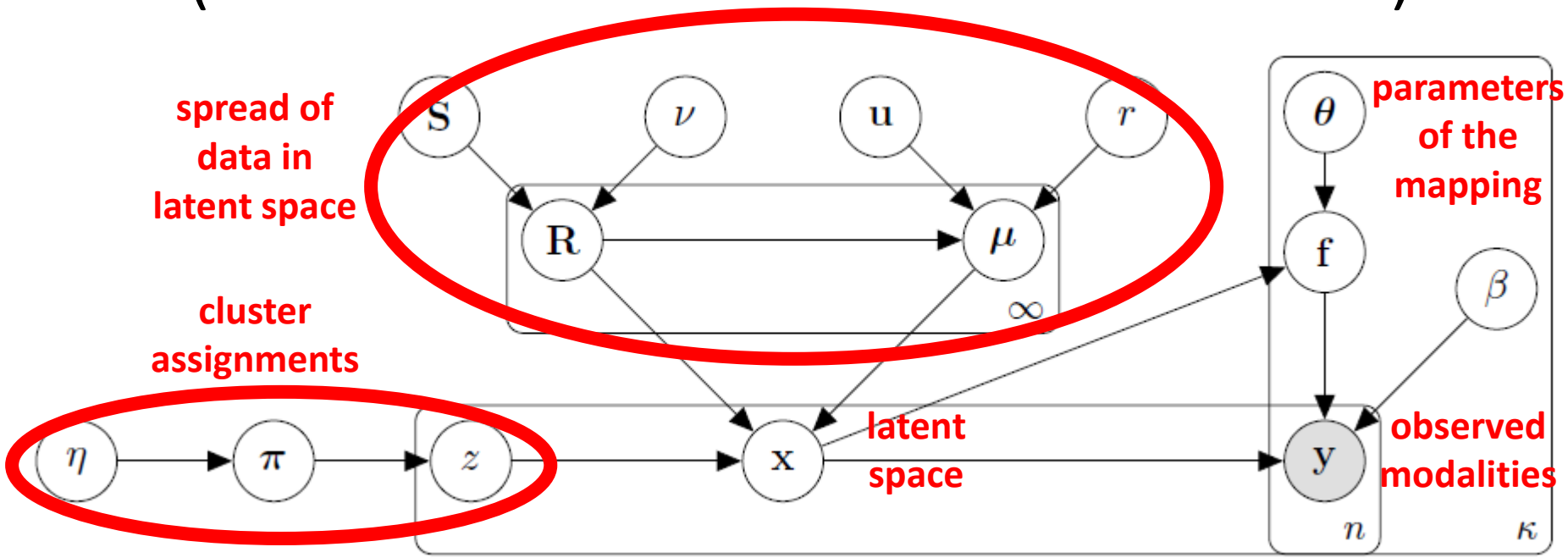
WYSS INSTITUTE

d dimensional latent ("hidden") space of frogs



multiple "observed" modalities

(aside on the mathematics – 1)



1. Draw mixture weights $\boldsymbol{\pi} \sim \mathcal{DP}(\eta)$
2. For each component $c = 1, \ldots, \infty$
   (a) Draw precision matrix $\mathbf{R}_c \sim \mathcal{W}(\boldsymbol{S}^{-1}, \nu)$
   (b) Draw mean $\boldsymbol{\mu}_c \sim \mathcal{N}(\boldsymbol{u}, (r\mathbf{R}_c)^{-1})$
3. For each entity $i = 1, \ldots, n$
   (a) Draw latent assignment $z_i \sim \text{Multinomial}(\boldsymbol{\pi})$
   (b) Draw latent coordinates $\mathbf{x}_{i,:} \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \mathbf{R}_{z_i}^{-1})$
4. For each view $k = 1, \ldots, \kappa$
   (a) Compute kernel K^k
   (b) For each observed dimension $j = 1, \ldots, p^k$
      i. Draw function $\mathbf{f}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^k)$
      ii. For each observation $i = 1, \ldots, n$
         A. Draw feature $y_{ij} \sim \mathcal{N}(\mathbf{f}_{:,j}(\mathbf{x}_{i,:}), (\beta^k)^{-1})$

WYSS INSTITUTE

(aside on the mathematics – 2)

- MCMC to evaluate posterior probabilities

$$p(\mathbf{X}|\mathbf{z}, \mathcal{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{u}, r, \mathbf{S}, \nu) \qquad p(\mathbf{z}|\mathbf{X}, \mathbf{u}, r, \mathbf{S}, \nu, \eta)$$

- Find out cluster assignments (integrating over the latent space and parameter space)

WYSS INSTITUTE

Advantages

- Principally clusters across modalities
  - by allowing modalities to “supervise” each other through a shared latent space
- Inherent dimensionality reduction
  - The low dimensional latent space (small d) can represent a high level humanistic understanding of how frogs cluster; thus a form of manifold learning
- Works for even few data points
  - Non-parametric Bayesian methods scale with data
- Requires no supervision
  - High level of abstraction
  - Automatically discovers number of clusters
- Can discover a global optima
  - Given enough time, an MCMC converges to the global optima
  - Integrating over other unknown variables gives a more robust clustering
- Allows fantasising new data
  - Given data in one modality, we can fantasise data in another modality through the shared latent space
- Can rank features by importance (both latent and observed)

Disadvantage: Very slow!

(Next step: move towards variational inference)

WYSS INSTITUTE

Results and Next Steps

- Frogs were clustering by development stages
- Interpret the latent dimensions
- Find out feature significances
- Relate Omics and HSI