Non-Gaussian Factor Model for Interpretable Ordinal Time Series Analysis

Yang Ni and Scott A. Bruce

Department of Statistics
Texas A&M University

Research overview (Scott)

Develop computationally efficient, adaptive methodology for analyzing large, complex time series data and producing optimal summary measures.

Complexity:

- Nonstationarity
- Covariate-dependent
- High-dimensionality
- Replicated data
- Spatio-temporal data

Themes:

- Frequency and time domain
- Nonparametric
- Learning tasks
- Bayesian computation
- Theoretical support

Transdisciplinary Collaboration: Sleep research, kinesiology, neuroscience, climate science.

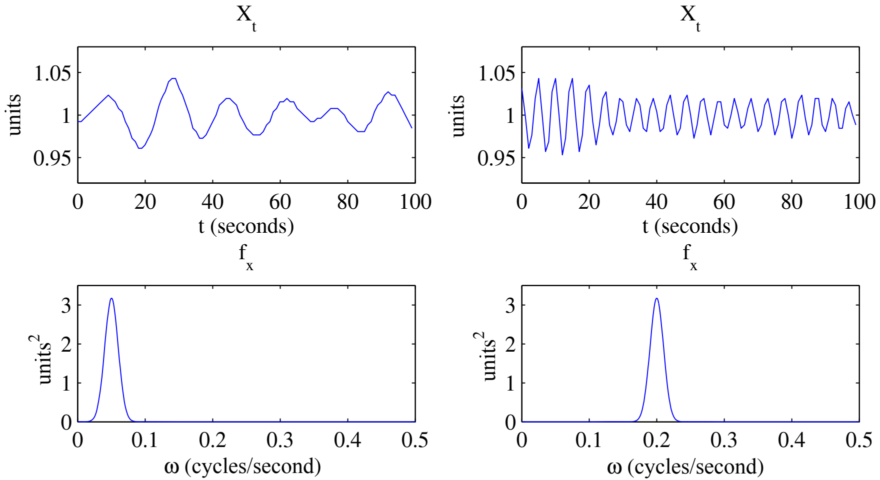
Frequency domain approach

- Consider a stationary univariate time series X_t.
- Cramér Representation:

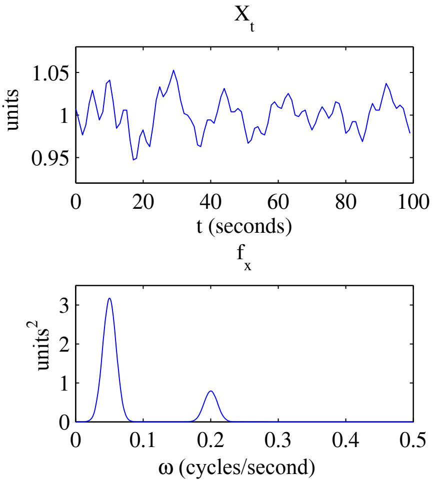
$$X_t = \int_{-1/2}^{1/2} A(\omega) \exp(2\pi i \omega t) dZ(\omega).$$

- Power spectrum: $f(\omega) = |A(\omega)|^2$.
- Type of frequency ANOVA: $\text{var}(X_t) = \int_{-1/2}^{1/2} f(\omega) d\omega$.
- Interpretation: variability of X_t attributable to periodic signals at frequency $\omega \in \mathbb{R}$.
- Can be extended to accommodate changes across time and/or covariate.

Two simulated examples



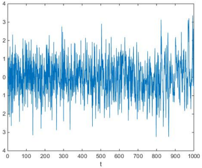
Two more simulated examples

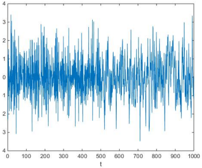


Adaptive Bayesian spectral analysis

Goal: Develop flexible methodology for adaptively estimating complex frequency domain structure and dependencies.

- Capture smooth and abrupt changes over different dimensions.
- Adaptive partitioning via reversible-jump Markov chain Monte Carlo.
- General framework for modeling nonstationary and covariate-dependent power spectra*.



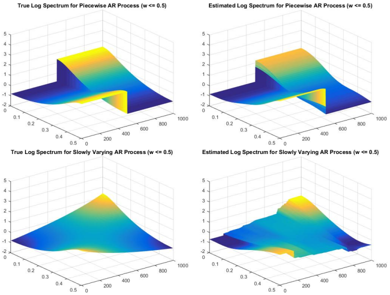


*Scott A. Bruce, Martica H. Hall, Daniel Buysse, and Robert T. Krafty (2018), Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series. *Biometrics*, 74:260-269.

Adaptive Bayesian spectral analysis

Goal: Develop flexible methodology for adaptively estimating complex frequency domain structure and dependencies.

- Capture smooth and abrupt changes over different dimensions.
- Adaptive partitioning via reversible-jump Markov chain Monte Carlo.
- General framework for modeling nonstationary and covariate-dependent power spectra*.

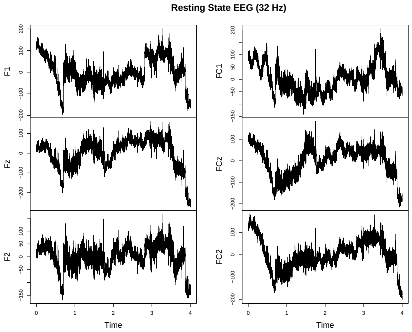


*Scott A. Bruce, Martica H. Hall, Daniel Buysse, and Robert T. Krafty (2018), Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series. *Biometrics*, 74:260-269.

What's new: Rebecca Lee developing hierarchical extension of this work.

Adaptive frequency band learning

Goal: Develop a quantitative framework for identifying frequency bands that optimally preserve characteristics of power spectra.

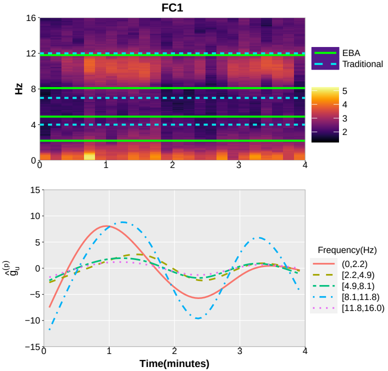


- Traditional frequency bands used in practice are often subjectively determined.
- Offer consistent estimation of frequency partition points*.
- Preserve important features:
    - time-varying dynamics
    - variability across replicates
    - discriminative power
    - cluster structure
    - local extrema

*Scott A. Bruce, Cheng Yong Tang, Martica H. Hall, and Robert T. Krafty (2020) Empirical Frequency Band Analysis of Nonstationary Time Series, *Journal of the American Statistical Association*, 115(532):1933-1945.

Adaptive frequency band learning

Goal: Develop a quantitative framework for identifying frequency bands that optimally preserve characteristics of power spectra.



- Traditional frequency bands used in practice are often subjectively determined.
- Offer consistent estimation of frequency partition points*.
- Preserve important features:
  - time-varying dynamics
  - variability across replicates
  - discriminative power
  - cluster structure
  - local extrema

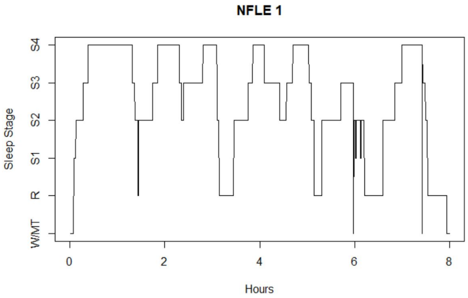
*Scott A. Bruce, Cheng Yong Tang, Martica H. Hall, and Robert T. Krafty (2020) Empirical Frequency Band Analysis of Nonstationary Time Series, *Journal of the American Statistical Association*, 115(532):1933-1945.

What's new: Connor Brubaker developing extension for multiple time series.

Categorical time series learning

Goal: Develop frequency-domain framework for supervised and unsupervised learning for categorical time series data.

- Motivation: studying sleep stage time series for different sleep disorders.
- Idea: use spectral envelope for feature-based learning.
- Offer consistent classification using frequency domain structure*.



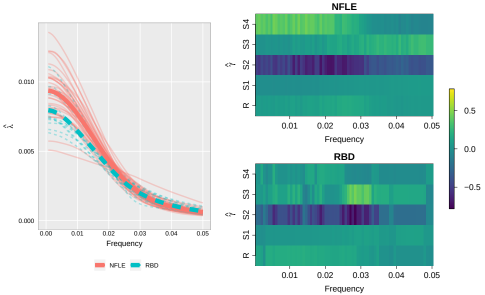
NFLE: Nocturnal Frontal Lobe Epilepsy, **RBD**: REM Behavior Disorder

*Zeda Li, Scott A. Bruce, and Tian Cai. (2022) "Classification of Categorical Time Series Using the Spectral Envelope and Optimal Scalings". Journal of Machine Learning Research. 23(299):1-31. `http://jmlr.org/papers/v23/21-0369.html`

Categorical time series learning

Goal: Develop frequency-domain framework for supervised and unsupervised learning for categorical time series data.

- Motivation: studying sleep stage time series for different sleep disorders.
- Idea: use spectral envelope for feature-based learning.
- Offer consistent classification using frequency domain structure*.



NFLE: Nocturnal Frontal Lobe Epilepsy, **RBD**: REM Behavior Disorder

*Zeda Li, Scott A. Bruce, and Tian Cai. (2022) "Classification of Categorical Time Series Using the Spectral Envelope and Optimal Scalings". Journal of Machine Learning Research. 23(299):1-31. `http://jmlr.org/papers/v23/21-0369.html`

What's new: Developing extension for nonstationary time series (interested students may contact me).

Identifying Granger Causal Structures and Stationarity Regions via Deep Learning

More to come...

Identifying Granger Causal Structures and Stationarity Regions via Deep Learning

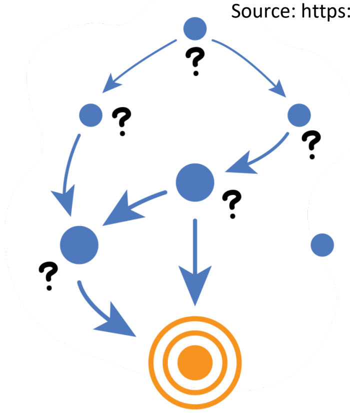
More to come...

What's new: Eric Gao testing feasibility and developing architecture.

Research Overview

Research Interests

Source: https://www.inguo.io/

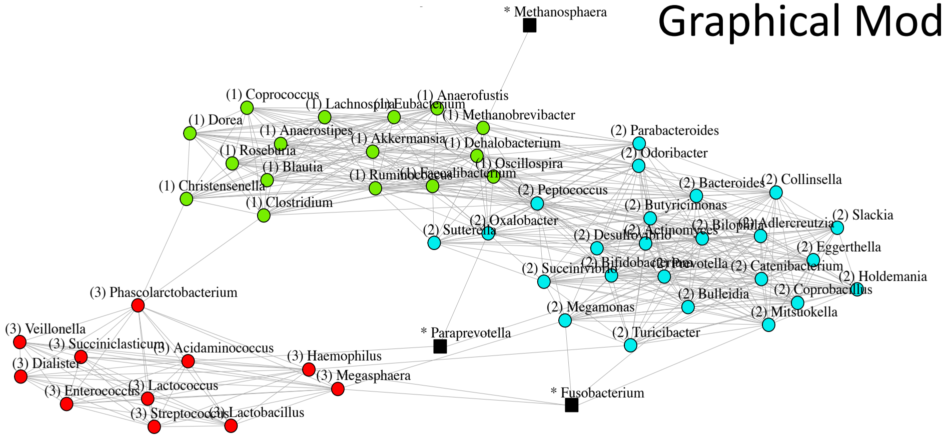


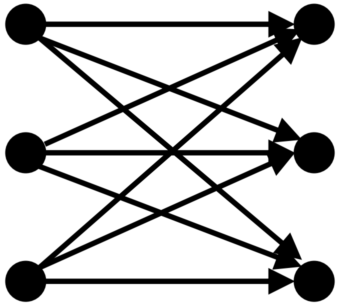
Methods

Causal Discovery

Clustering and Bayesian Nonparametrics

Graphical Models





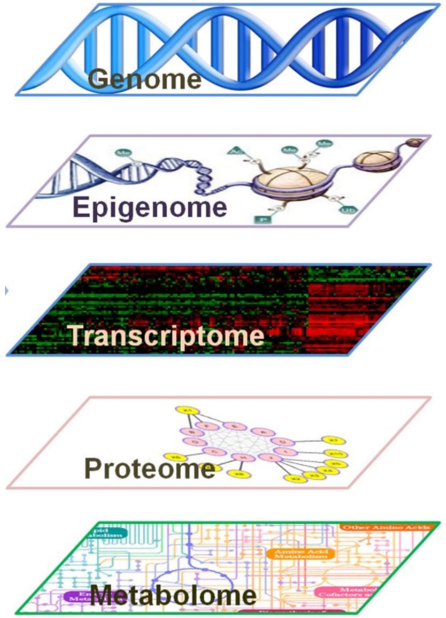
Applications

Single-Cell Multi-Omics

Microbiome Multi-Omics

Digital Health

Source: Yan V. Sun, Yijuan Hu 2016



Source: aamc.org

Causal Discovery

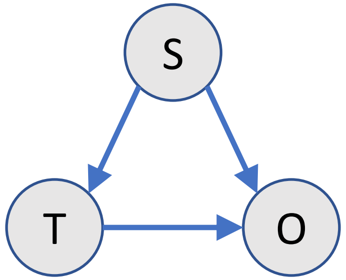
Simpson's Paradox

- Kidney stone treatment

	Treatment A	Treatment B
	78%(273/350)	83%(289/350)

Simpson’s Paradox

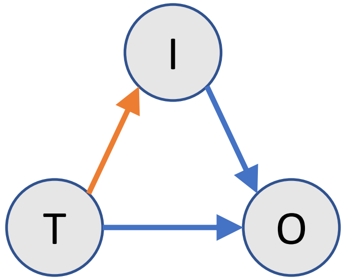
- Kidney stone treatment



	Treatment A	Treatment B
Small Stones	93%(81/87)	87%(234/270)
Large Stones	73%(192/263)	69%(55/80)
Total	78%(273/350)	83%(289/350)

Simpson's Paradox

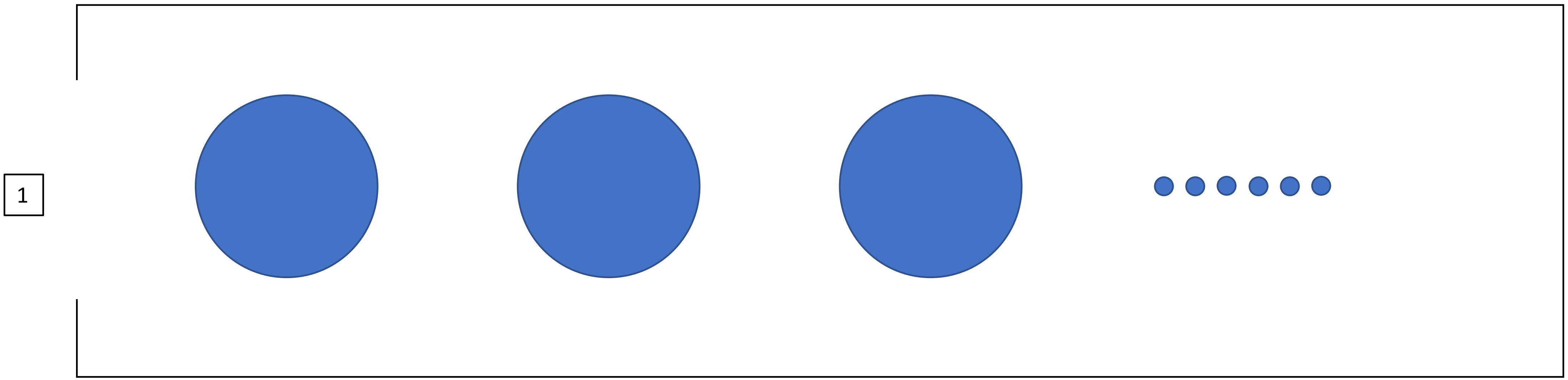
- Kidney stone treatment



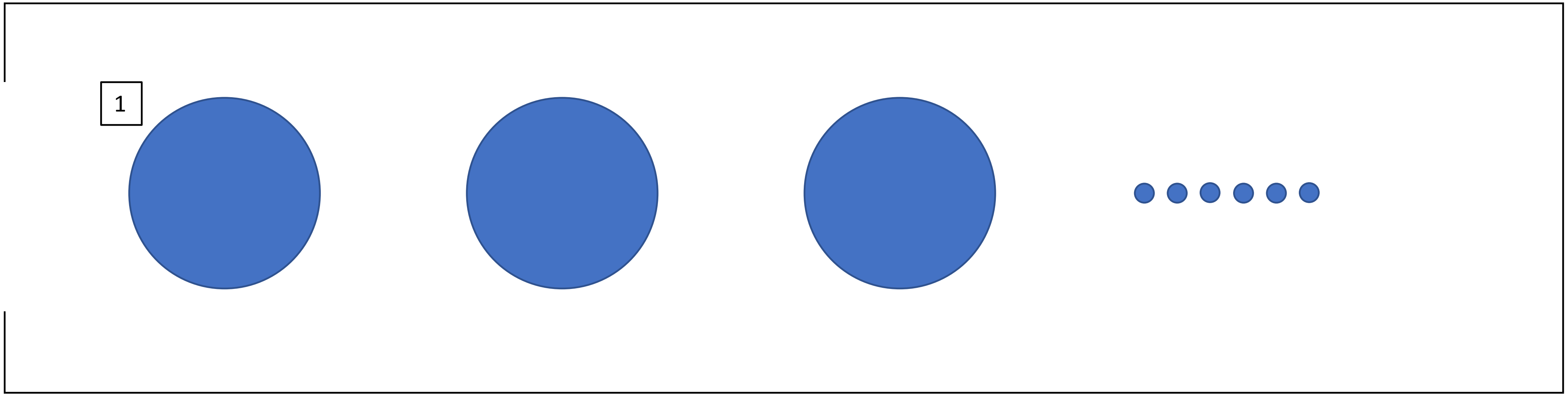
	Treatment A	Treatment B
Non-Infection	93%(81/87)	87%(234/270)
Infection	73%(192/263)	69%(55/80)
Total	78%(273/350)	83%(289/350)

Bayesian Nonparametrics

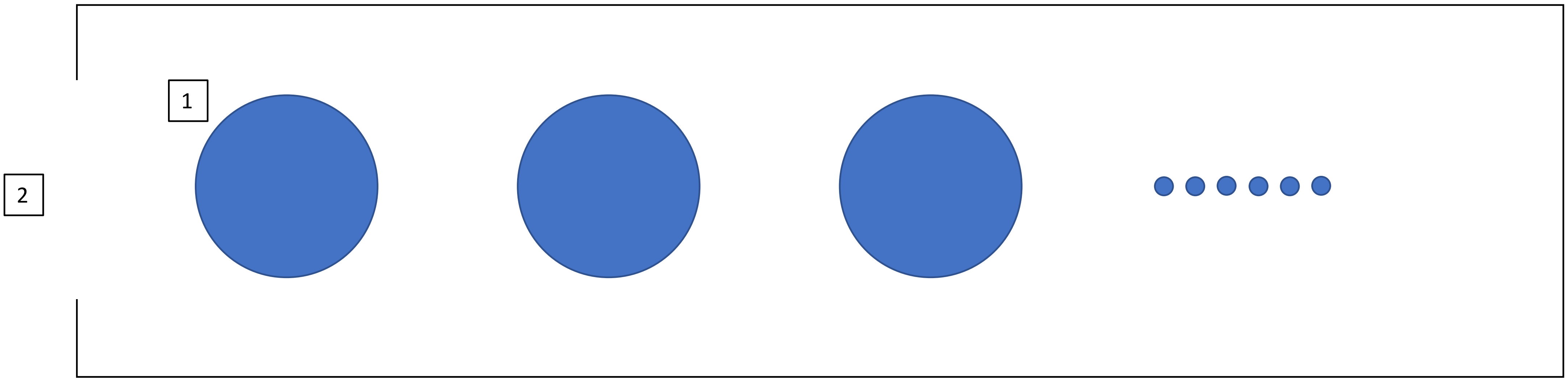
Chinese Restaurant Process



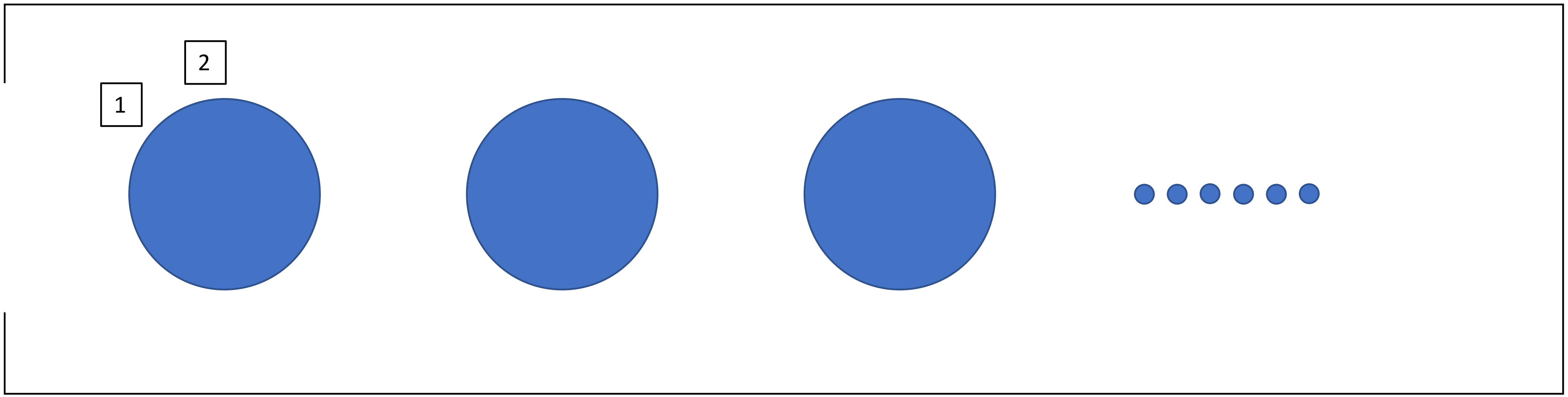
Chinese Restaurant Process



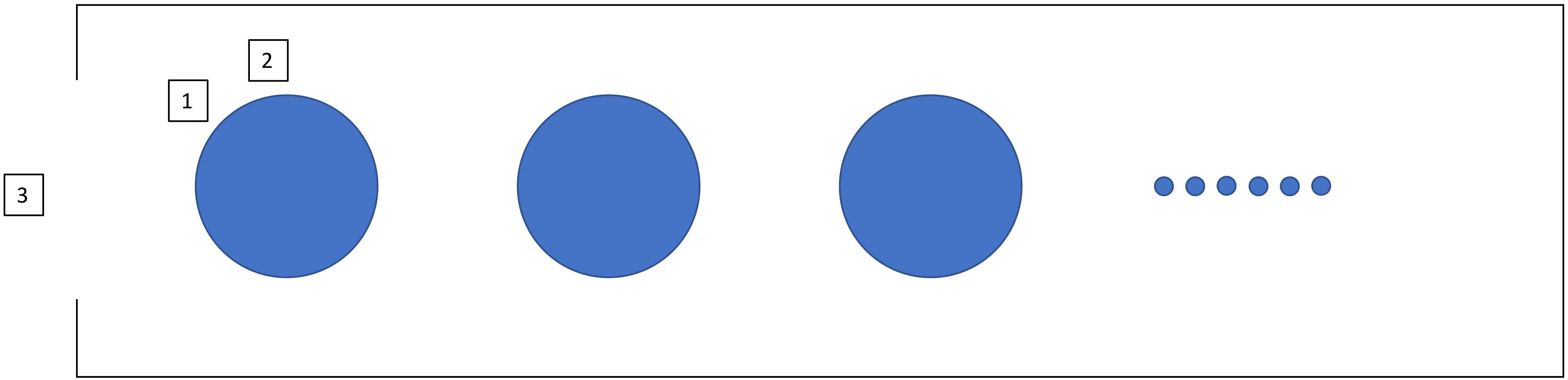
Chinese Restaurant Process



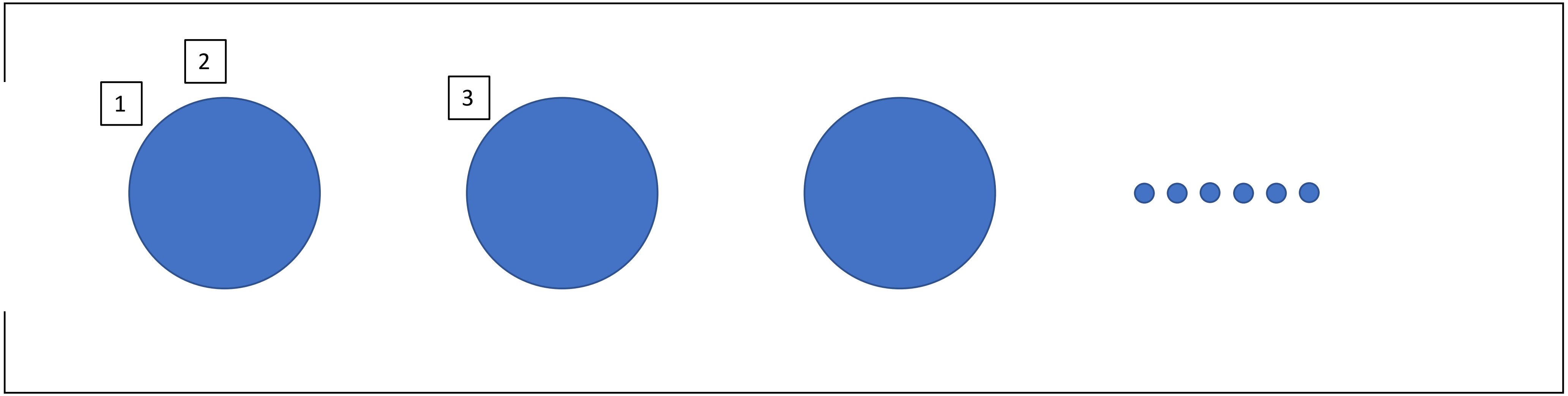
Chinese Restaurant Process



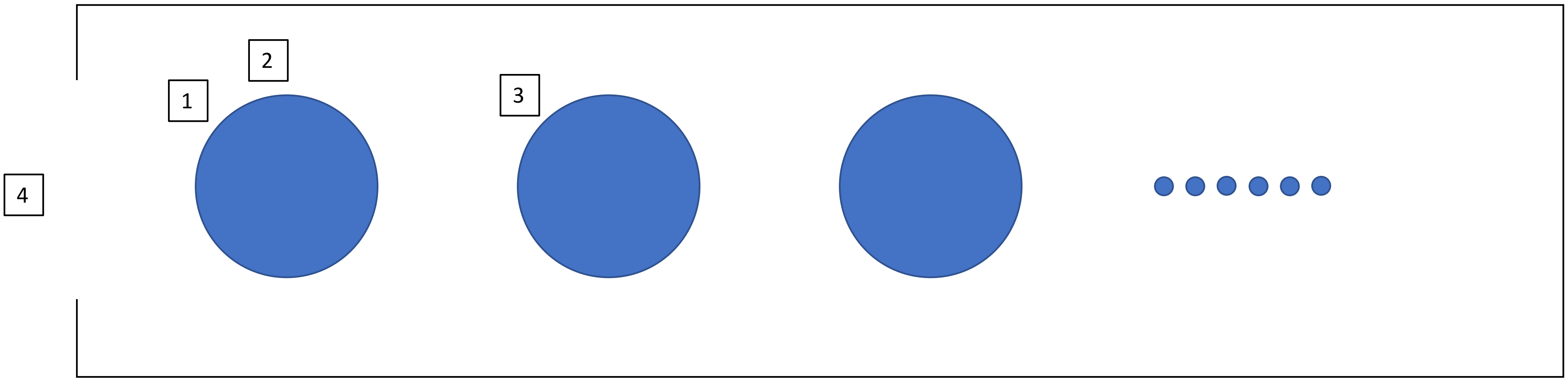
Chinese Restaurant Process



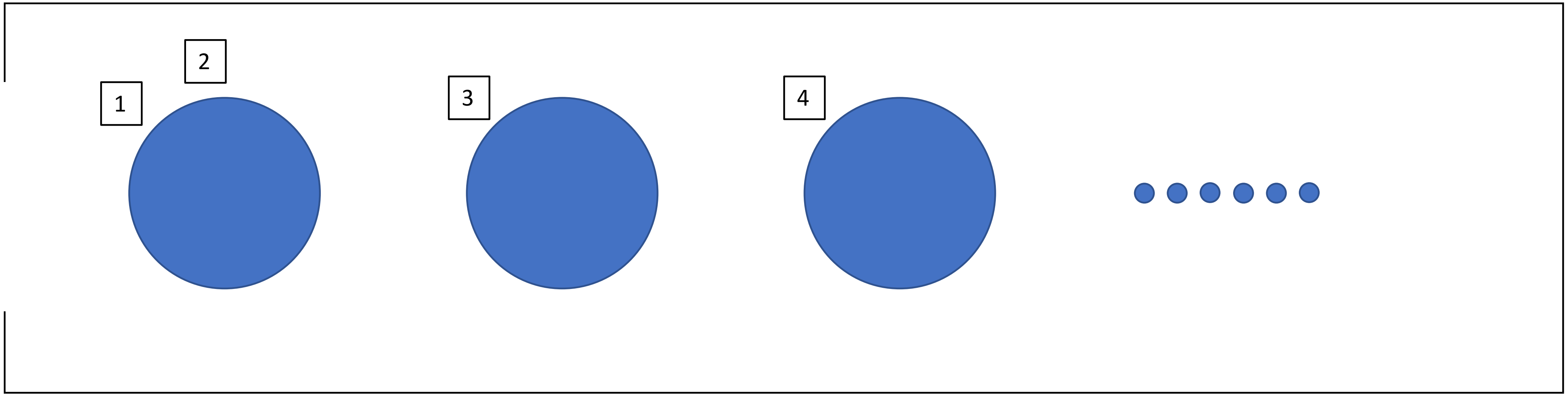
Chinese Restaurant Process



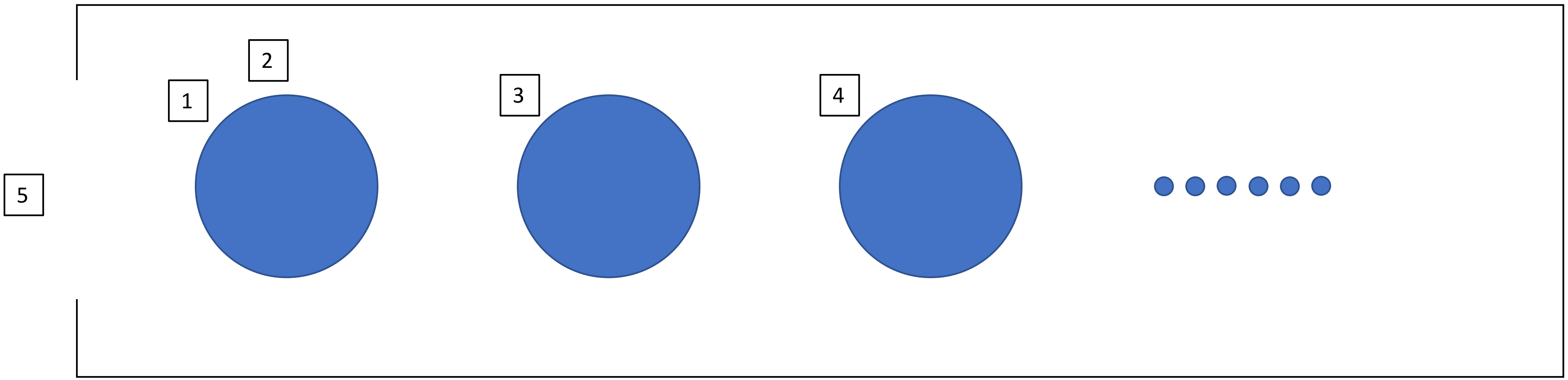
Chinese Restaurant Process



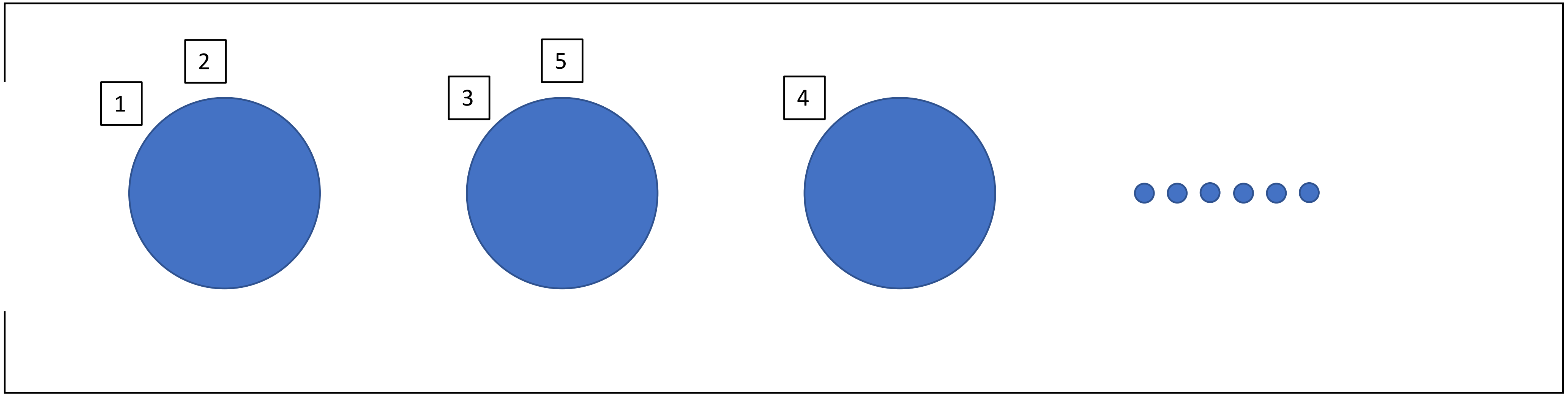
Chinese Restaurant Process



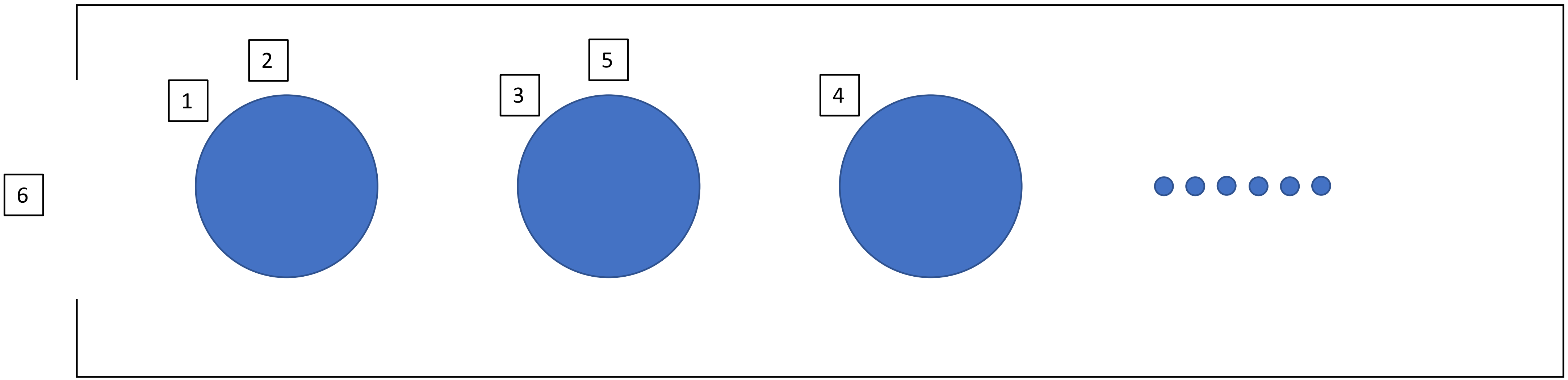
Chinese Restaurant Process



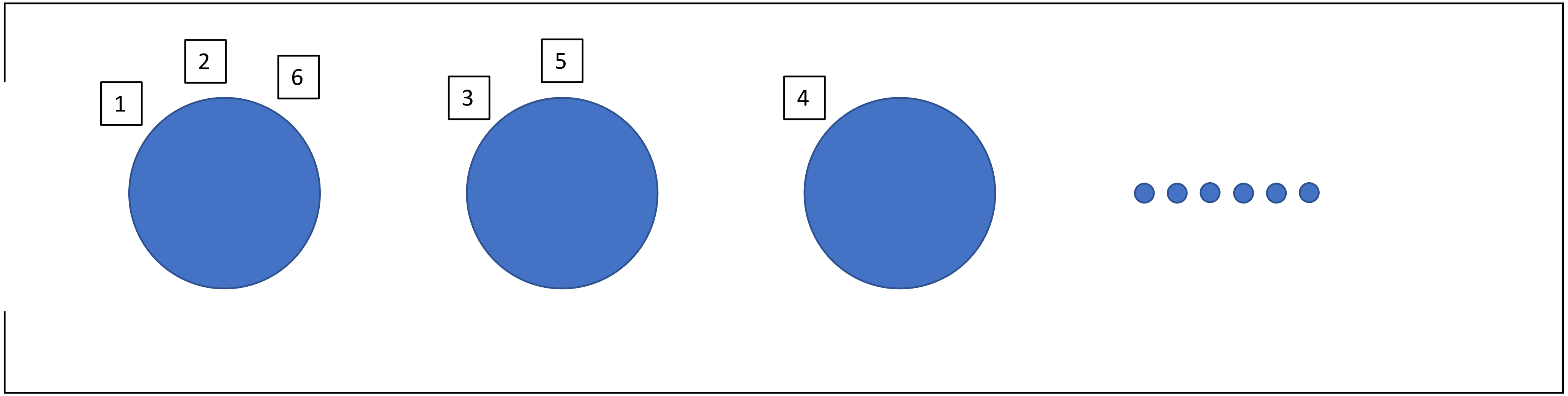
Chinese Restaurant Process



Chinese Restaurant Process



Chinese Restaurant Process



Indian Buffet Process

$\boxed{1}$

Indian Buffet Process

$\boxed{1}$

Indian Buffet Process

......

1

2

Indian Buffet Process

$\boxed{1}$

$\boxed{2}$ $\boxed{2}$

Indian Buffet Process

$\boxed{1}$

$\boxed{2}$ $\boxed{2}$

$\boxed{3}$

Indian Buffet Process

$\boxed{1}$

$\boxed{2}$ $\boxed{2}$

$\boxed{3}$

Indian Buffet Process

$\cdots\cdots$

$\boxed{1}$

$\boxed{2}$ $\boxed{2}$

$\boxed{4}$

$\boxed{3}$

Indian Buffet Process

$\boxed{1}$

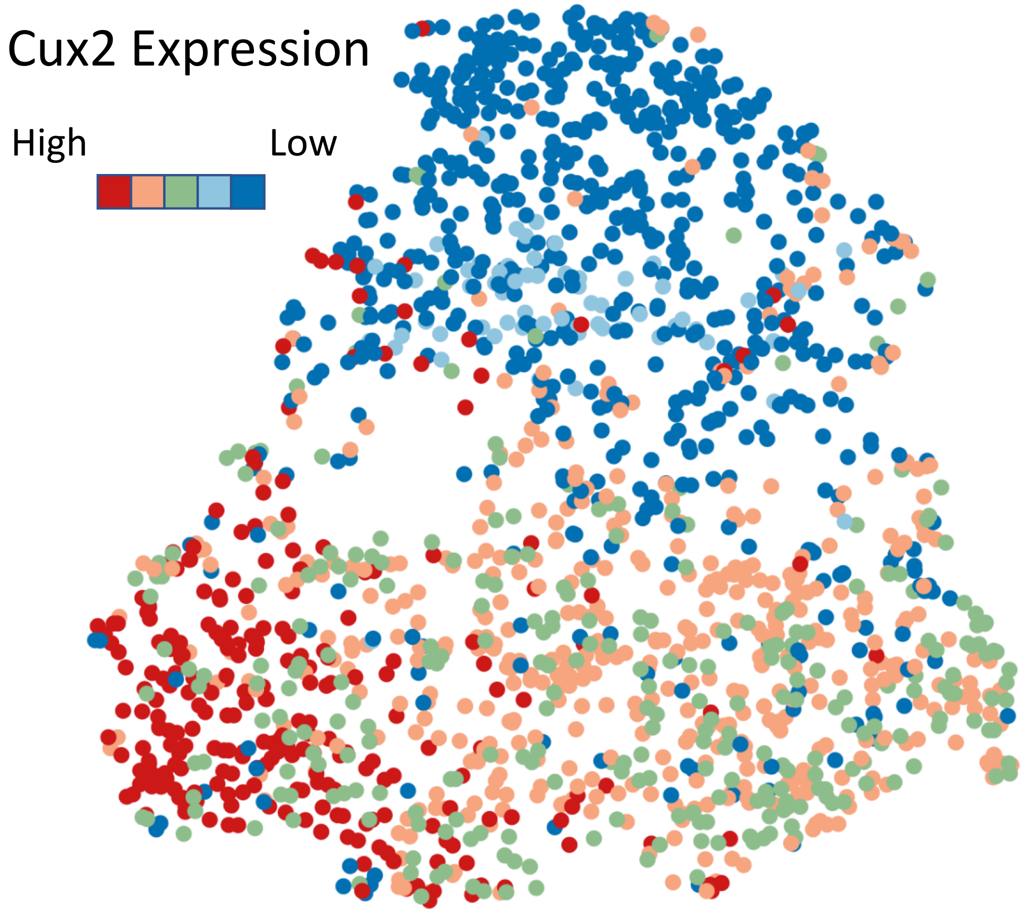
$\boxed{2}$ $\boxed{2}$

$\boxed{3}$

$\boxed{4}$ $\boxed{4}$ $\boxed{4}$

Single-Cell Multi-Omics

scRNA-seq



Features:

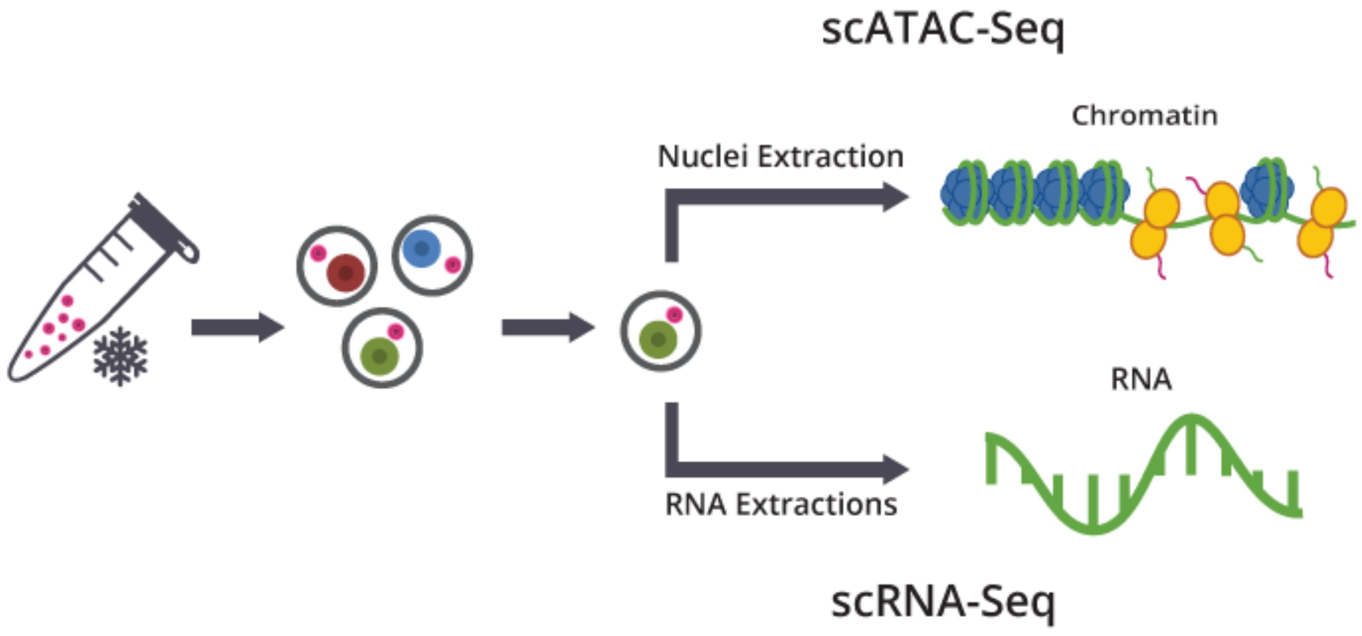
- Zero-inflated counts
- Cells may be spatially correlated (spatial transcriptomics)
- A tissue region may be non-convex

Problems:

- Find cell types – clustering
- Estimate gene regulatory networks – causal discovery and graphical models

scRNA-seq + scATAC-seq

Source: https://www.genewiz.com/Public/Services/Next-Generation-Sequencing/Single-Cell-ATAC-Seq/



Features:

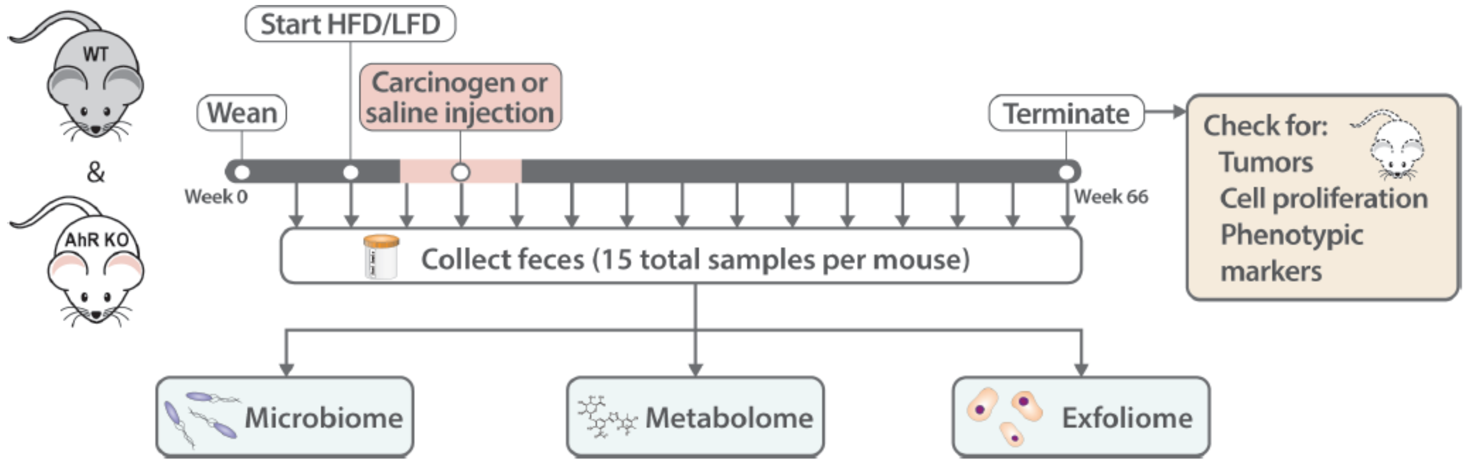
- Zero-inflated counts
- Chromatin accessibility + gene expression

Problems:

- Integrate these two types of data by using biological knowledge

Microbiome Multi-Omics

Microbiome + Metabolome + Exfoliome



Features:

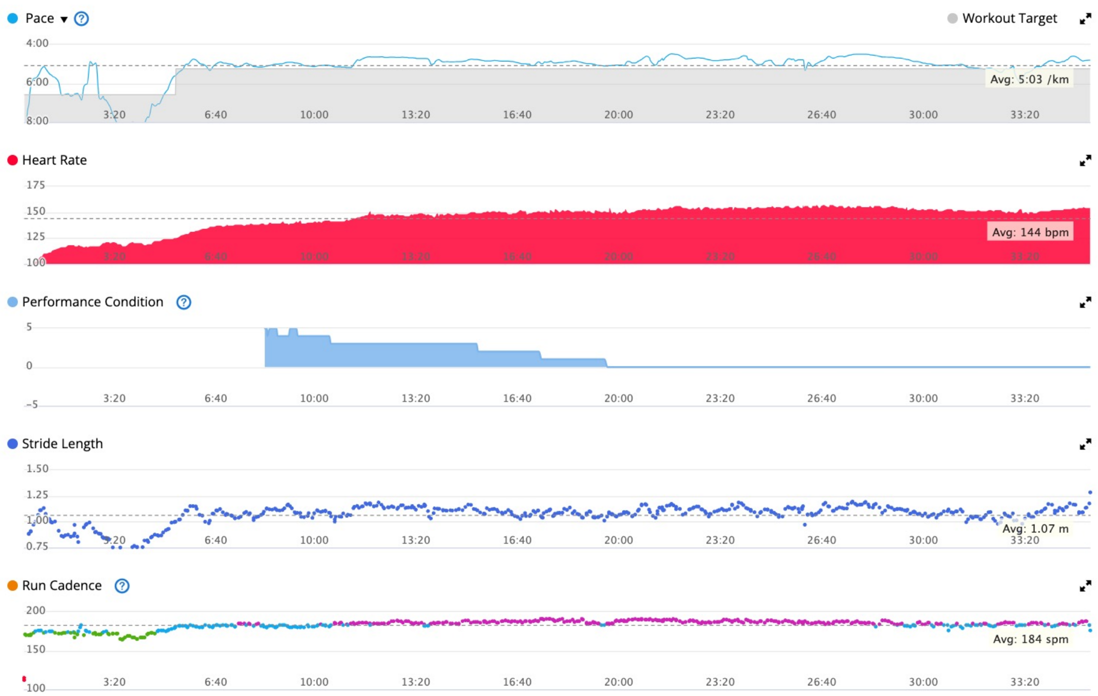
- Microbiome data are zero-inflated and compositional
- Longitudinal data
- Multiple data types

Problems:

- Find microbial communities
- Estimate microbial networks

Digital Health

Wearable Device Data



fenix 6X Pro
Software: 26.00
Elev Corrections: Disabled
Summary Data: Original

HRM-Pro
Software: 8.8

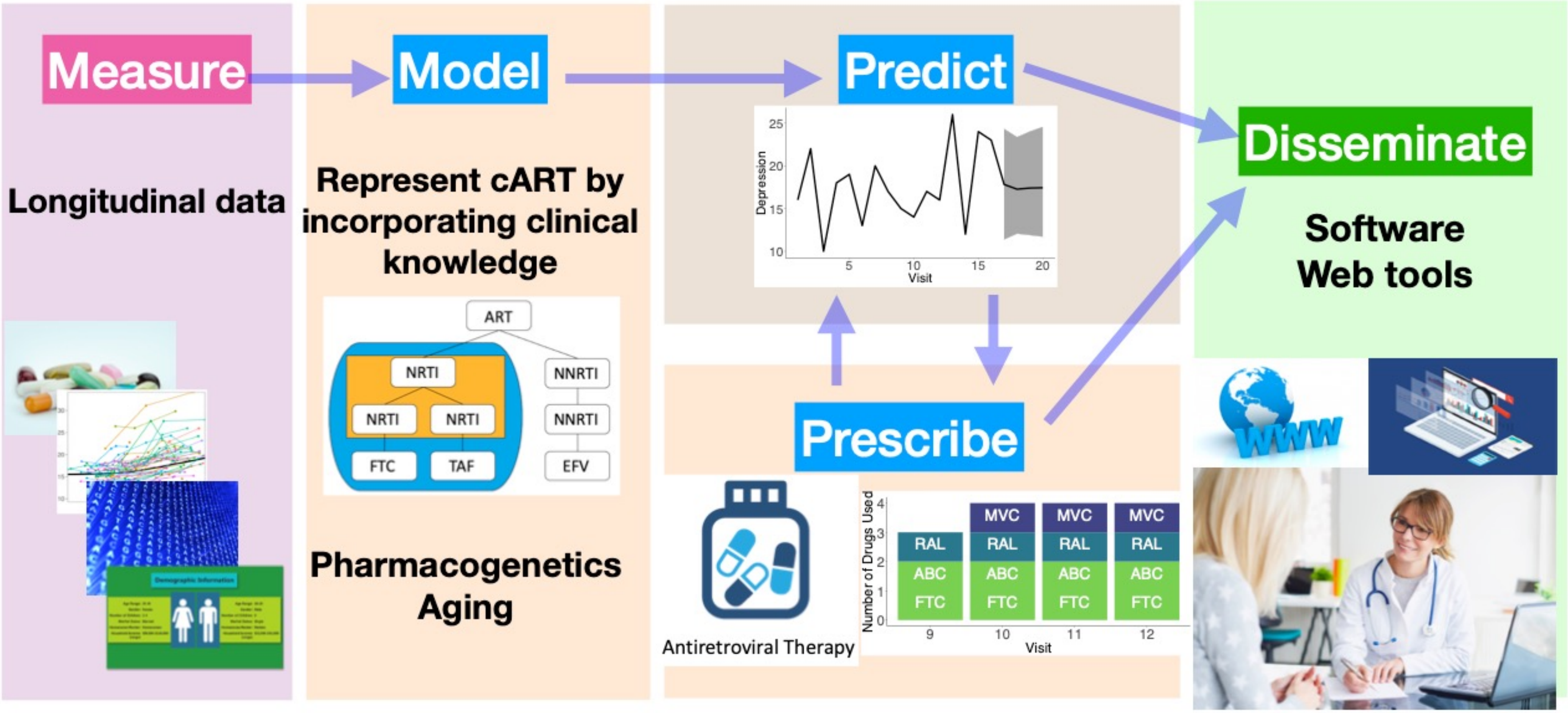
Features:

- Multivariate functional data are infinite dimensional
- Mixed type (count, categorical, etc)

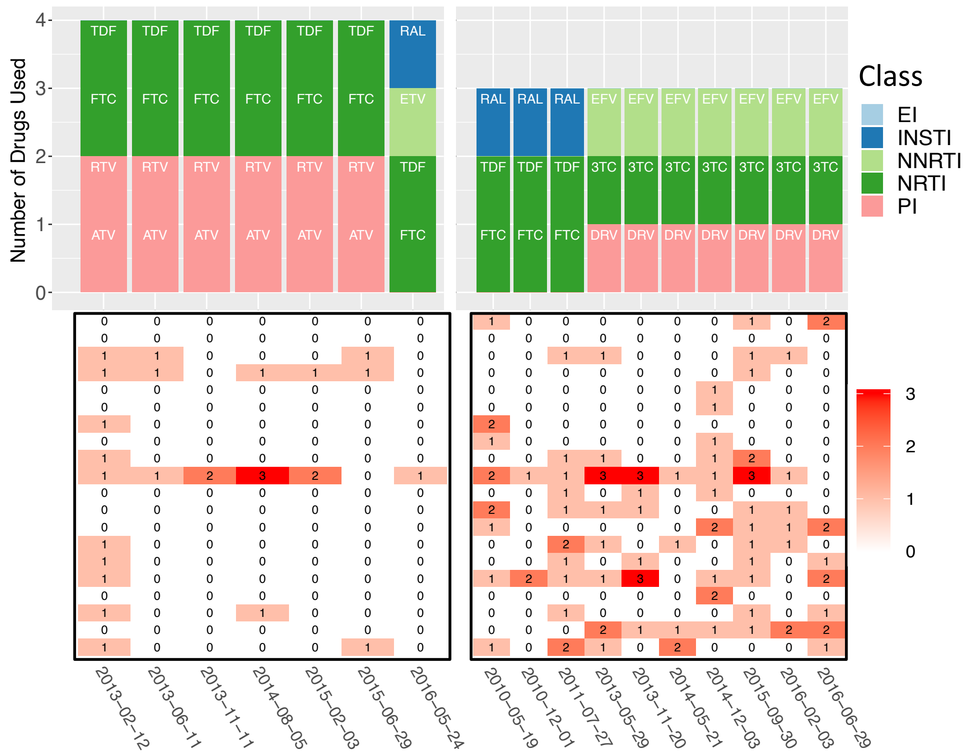
Problems:

- Identify causal relationships among random functions

HIV Longitudinal Study



HIV Longitudinal Study



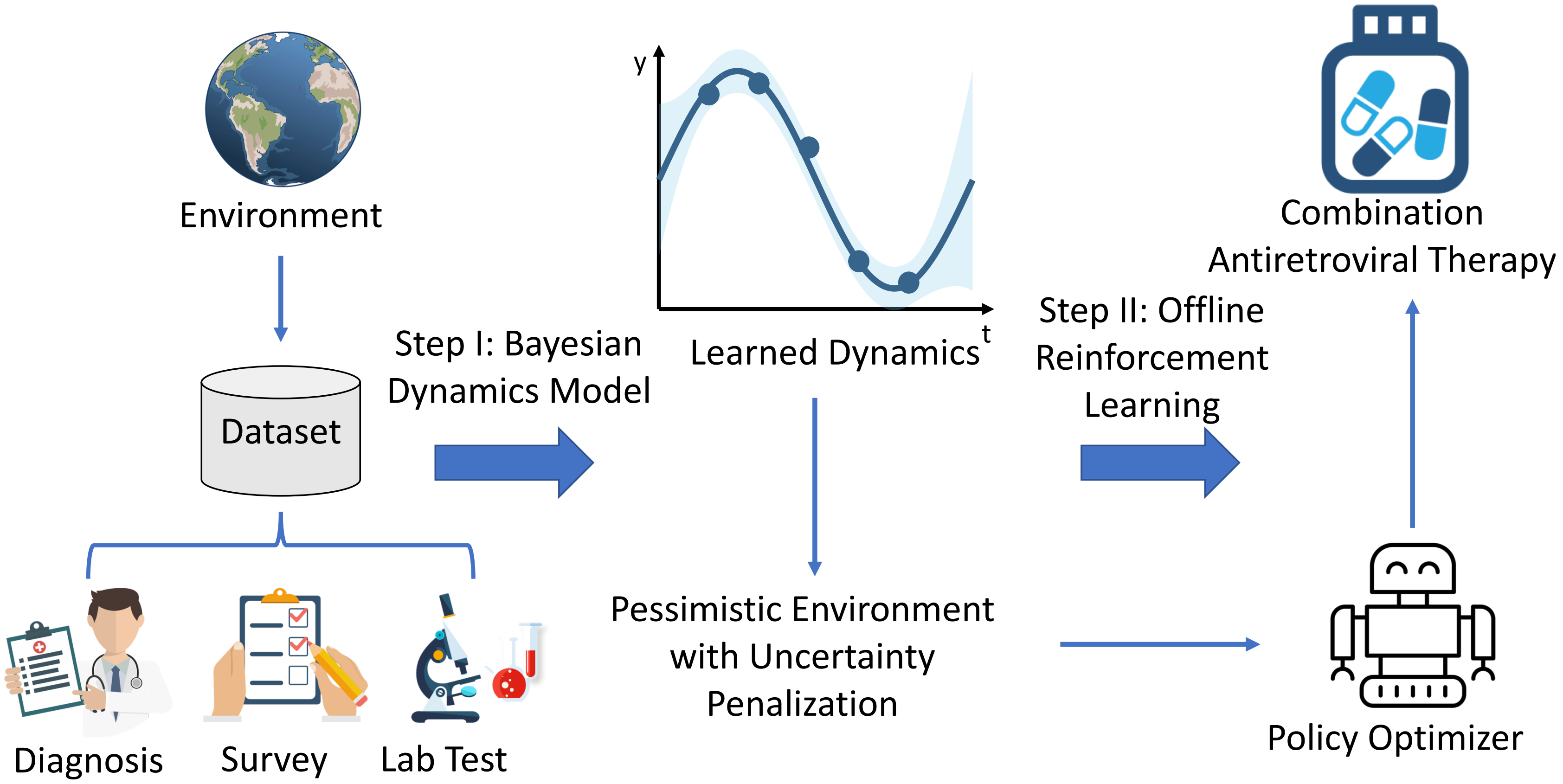
Center for Epidemiologic Studies Depression Scale (CES-D), NIMH

Below is a list of the ways you might have felt or behaved. Please tell me how often you have felt this way during the past week.

	During the Past Week			
	Rarely or none of the time (less than 1 day)	Some or a little of the time (1-2 days)	Occasionally or a moderate amount of time (3-4 days)	Most or all of the time (5-7 days)
1. I was bothered by things that usually don't bother me.	☐	☐	☐	☐
2. I did not feel like eating; my appetite was poor.	☐	☐	☐	☐
3. I felt that I could not shake off the blues even with help from my family or friends.	☐	☐	☐	☐
4. I felt I was just as good as other people.	☐	☐	☐	☐
5. I had trouble keeping my mind on what I was doing.	☐	☐	☐	☐
6. I felt depressed.	☐	☐	☐	☐
7. I felt that everything I did was an effort.	☐	☐	☐	☐
8. I felt hopeful about the future.	☐	☐	☐	☐
9. I thought my life had been a failure.	☐	☐	☐	☐
10. I felt fearful.	☐	☐	☐	☐
11. My sleep was restless.	☐	☐	☐	☐
12. I was happy.	☐	☐	☐	☐
13. I talked less than usual.	☐	☐	☐	☐
14. I felt lonely.	☐	☐	☐	☐
15. People were unfriendly.	☐	☐	☐	☐
16. I enjoyed life.	☐	☐	☐	☐
17. I had crying spells.	☐	☐	☐	☐
18. I felt sad.	☐	☐	☐	☐
19. I felt that people dislike me.	☐	☐	☐	☐
20. I could not get "going."	☐	☐	☐	☐

SCORING: zero for answers in the first column, 1 for answers in the second column, 2 for answers in the third column, 3 for answers in the fourth column. The scoring of positive items is reversed. Possible range of scores is zero to 60, with the higher scores indicating the presence of more symptomatology.

HIV Longitudinal Study



Non-Gaussian Factor Model for Interpretable Ordinal Time Series Analysis

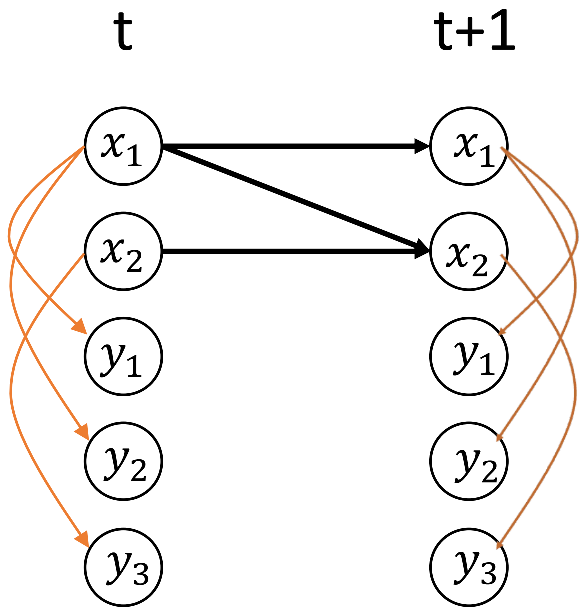
- $Y_t = (y_{1t}, \ldots, y_{pt})'$ observed ordinal responses to the questions at time $t = 1, \ldots, T$
- $y_{jt} \in \{1, \ldots, C_j\}$
- $X_t = (x_{1t}, \ldots, x_{Lt})'$ latent continuous factors/traits at time t
- Assume $L \leq p$

- Let $X_t^{(d)} = \{X_{t-1}, \ldots, X_{t-d}\}$

Non-Gaussian Factor Model for Interpretable Ordinal Time Series Analysis

$$X_t = g\left(X_t^{(d)}, \epsilon_t\right)$$ Causal model

$$\Pr\left(y_{jt} \leq c \middle| X_T\right) = \Phi(\eta_c - \beta_j' X_t - \alpha_j)$$ Measurement/factor model



Two sources of non-identifiability:

1. Latent factors
2. Causal graph

Our hypothesis:

If ϵ_t is non-Gaussian, both latent factors and the causal graph is identifiable