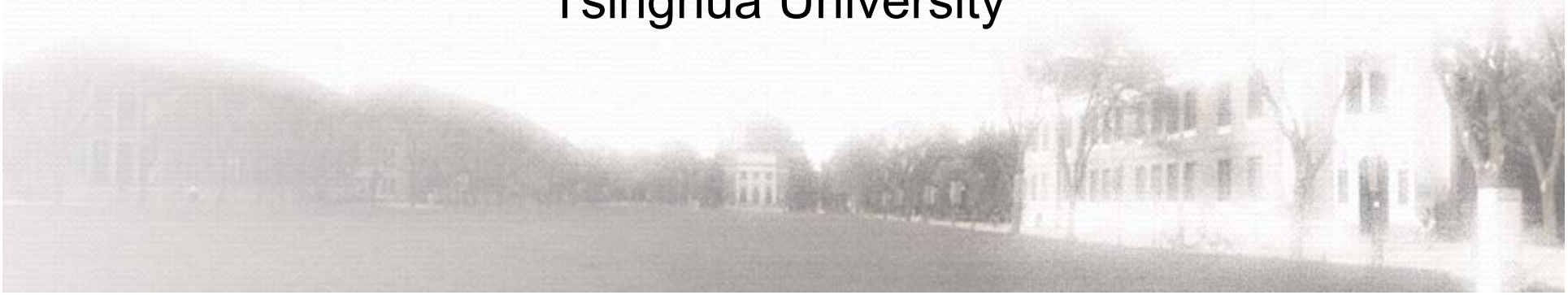
清華大學
Tsinghua University

Research Summary

Ge Liu

Machine Learning and Computational Biology Group, IIIS,

Tsinghua University



Research Experiences

- 2013.10 – 2014.2 Morikawa Lab, Tokyo University

  - Design of Analog CDMA communication system

- 2014. 7 – 2014.9 Murphy's Lab, CMU

  - Developing Multiple Object Types Generative Models for Protein Pattern from 3D Fluorescence Microscopy Images

- 2014.5 – present ML & Compbio Group, IIIS, Tsinghua

  - Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach

  - Analyzing Resting State Brain Connectivity and its Applications on Psychiatric Diagnosis from Electro-magnetic Medical Images

Developing Multiple Object Types Generative Models for Protein Pattern from 3D Fluorescence Microscopy Images

Lab:

Murphy's Lab, Lane Center of Computational Biology, CMU

Advisor:

Professor Robert F. Murphy

Motivation

- Building generative models for cells

  - Study the biochemical properties of cells and the mechanism of cell metabolism

  - Time and spatial resolved simulation

  - Cellorganizer: nuclear, cell shape and protein models

- Protein model

  - Study the patterns of organelles by studying patterns of proteins that label a certain type of organelles

  - Key elements: shapes and spatial distribution

  - There were some limitations in previous model in Cellorganizer.

Previous protein model in Cellorganizer

- Shape and size
  - Gaussian Mixture Model
- Spatial distribution

The objects do not always look like Gaussian objects

$$P(s, \varphi, \theta) = \frac{e^{\beta_0+\beta_1 s+\beta_2 s^2+\beta_3 \cos\varphi \sin\theta+\beta_4 \sin\varphi \sin\theta+\beta_5 \cos\theta}}{1 + e^{\beta_0+\beta_1 s+\beta_2 s^2+\beta_3 \cos\varphi \sin\theta+\beta_4 \sin\varphi \sin\theta+\beta_5 \cos\theta}}$$

where *s* is the ratio of the distance of a given object's center to the nuclear surface over the sum of that distance and the distance to the cell surface; ϑ and ϕ is the inclination angle and azimuth angle of the vector from the nuclear center to the object center.[1]

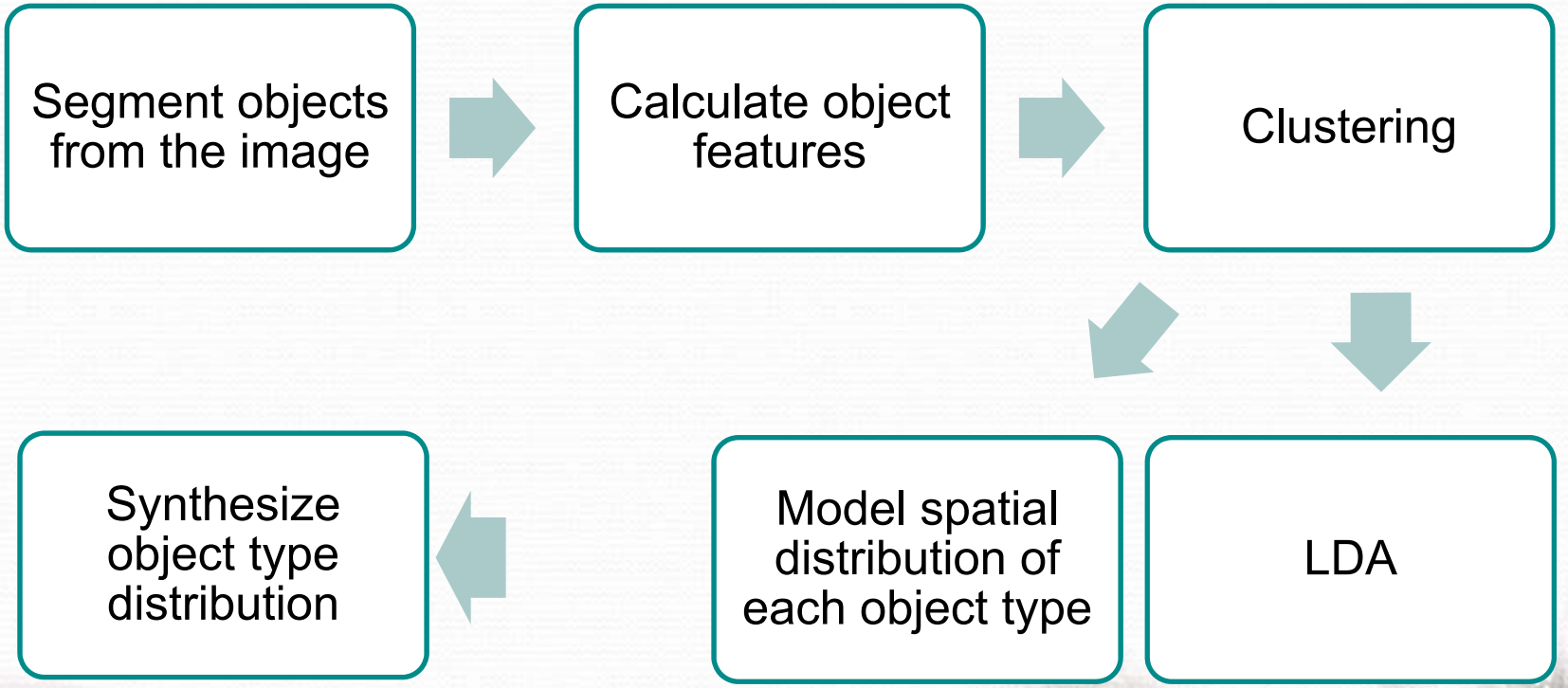
- Spatial distribution of the objects is independent to the shape of the objects.

The spatial distribution may depends on the features of objects.

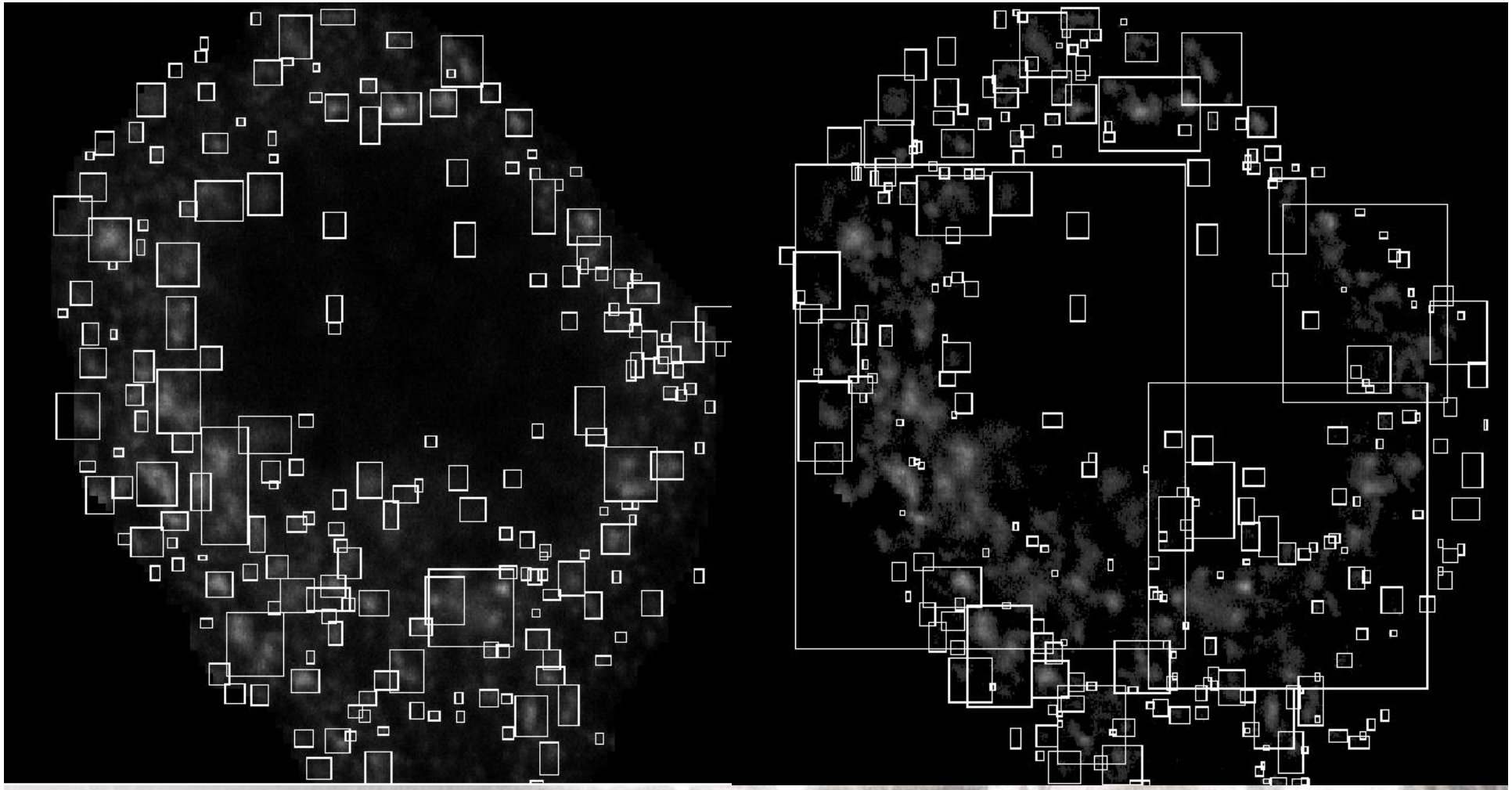
Multiple object types model

- More morphological features of each object should be taken into consideration, and the objects can be classified into different types based on the features.
- A cell level pattern can be studied based on the distribution of object types
- The dependency between spatial distribution and object features should be considered

Pipeline for model building



Improved object segmentation method



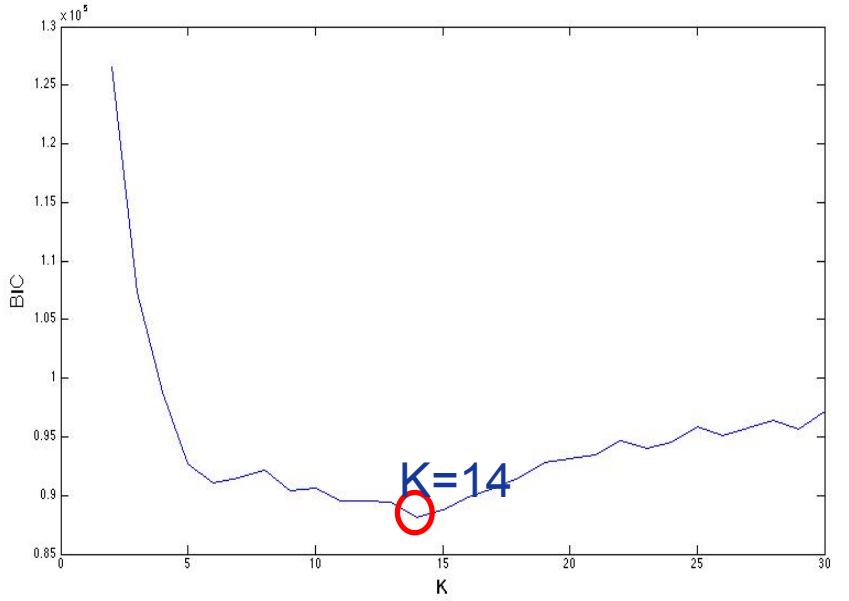
Now

Before

3D Sub-cellular Object Features

- SOF1.1 volumes V
- SOF1.2 surface area S
- SOF1.3 integral of mean curvature M
- SOF1.4 Euler number
- SOF1.5 Shape factor1 $f_1 = 6\sqrt{\pi}\frac{V}{\sqrt{S^3}}$
- SOF1.6 Shape factor2 $f_2 = 48\pi^2\frac{V}{M^3}$
- SOF1.7 Shape factor3 $f_3 = 4\pi\frac{S}{M^2}$
- SOF1.8 ~ 1.11 :Skeleton features

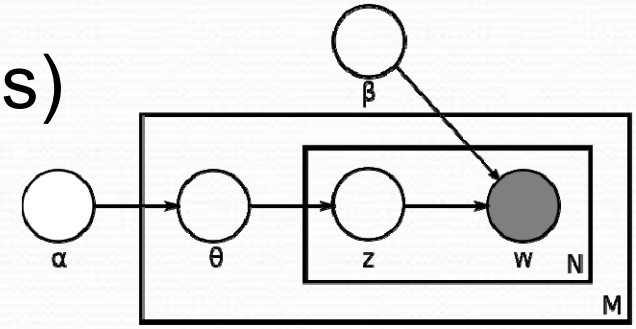
K-means cluster defined 14 object types



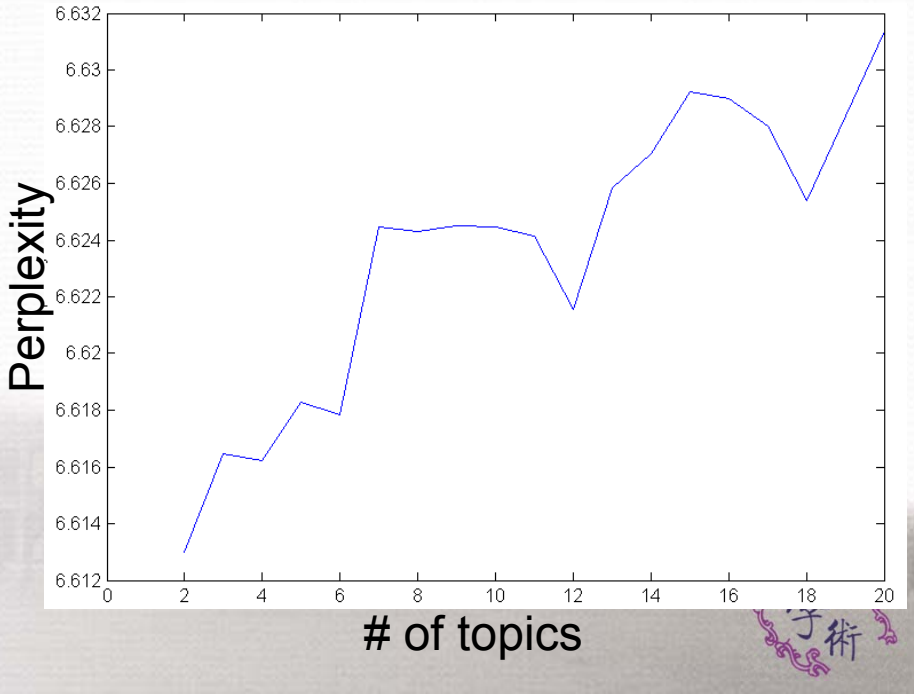
Using LDA for cell level pattern analysis

- Words: object types
- Documents: bags of words (Cells)
- Topics: distributions over words

Dirichlet-polynomial distribution

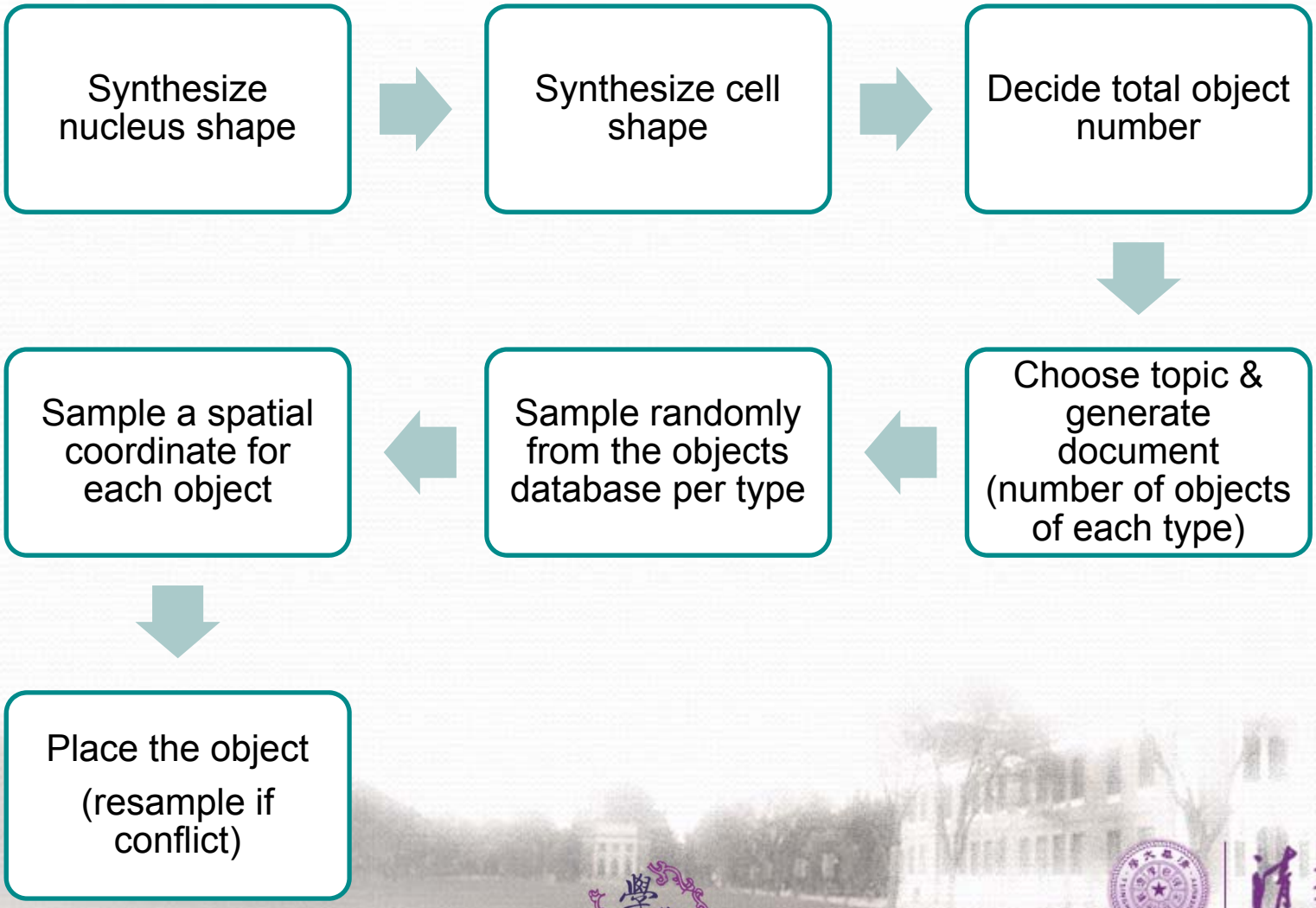


(http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)

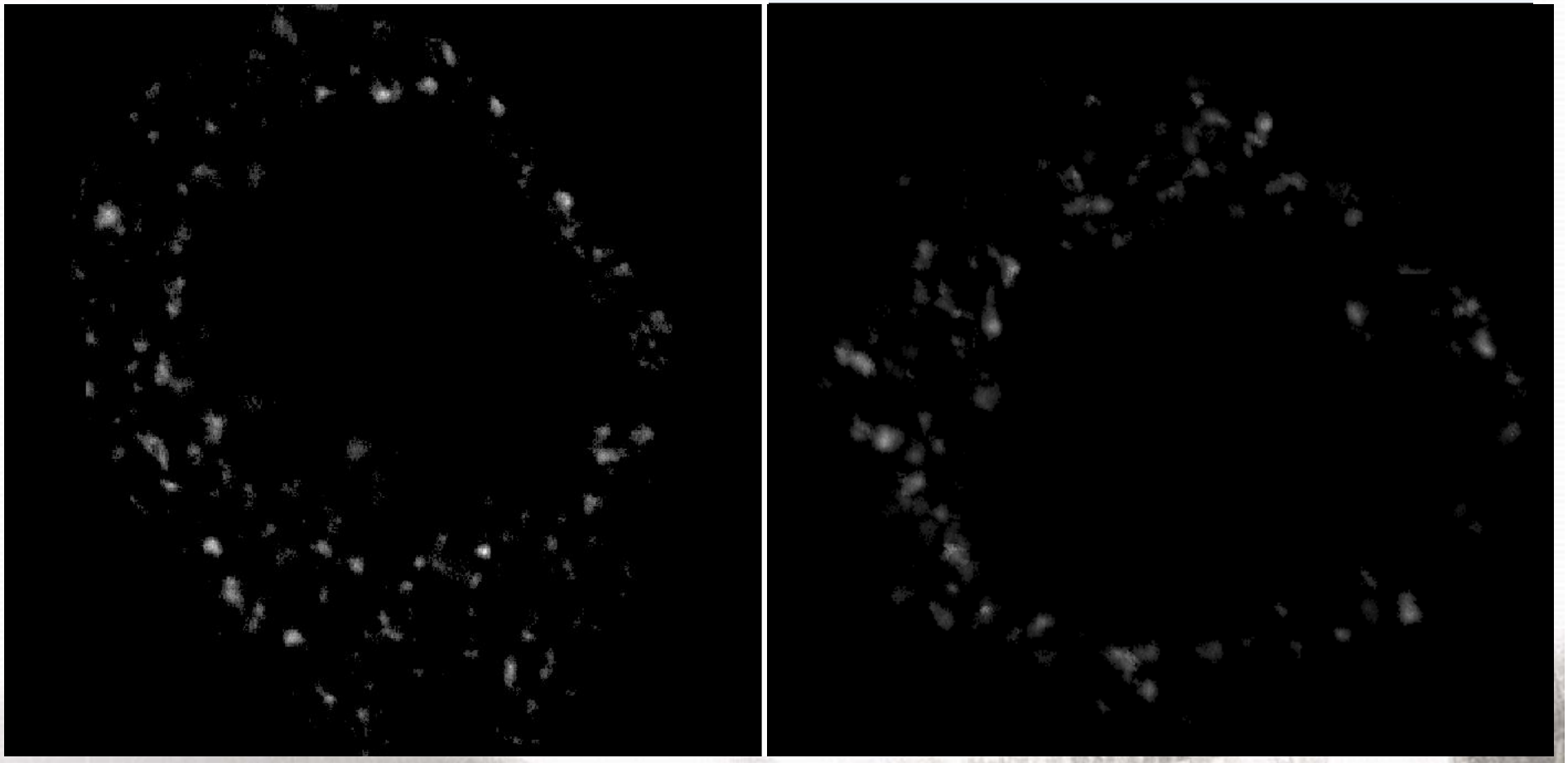


Perplexity = exp(-log likelihood/document length)

Synthesize protein image



Synthesizing new protein image



Original

Synthesized

Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach

Lab:

Machine Learning and Computational Biology Group, IIIS, Tsinghua University

Advisor:

Professor Jianyang Zeng

Background

- TCGA data -> cancer subtype identification
- Previous integrative clustering approaches for analyzing cancer data seldom exploit both intrinsic properties and cross-modality correlations.[2]
- Different types of data:

  - real valued data: gene expression, DNA Methylation, miRNA expression, etc.

  - discrete data: mutation, copy number, etc.

Mutation data

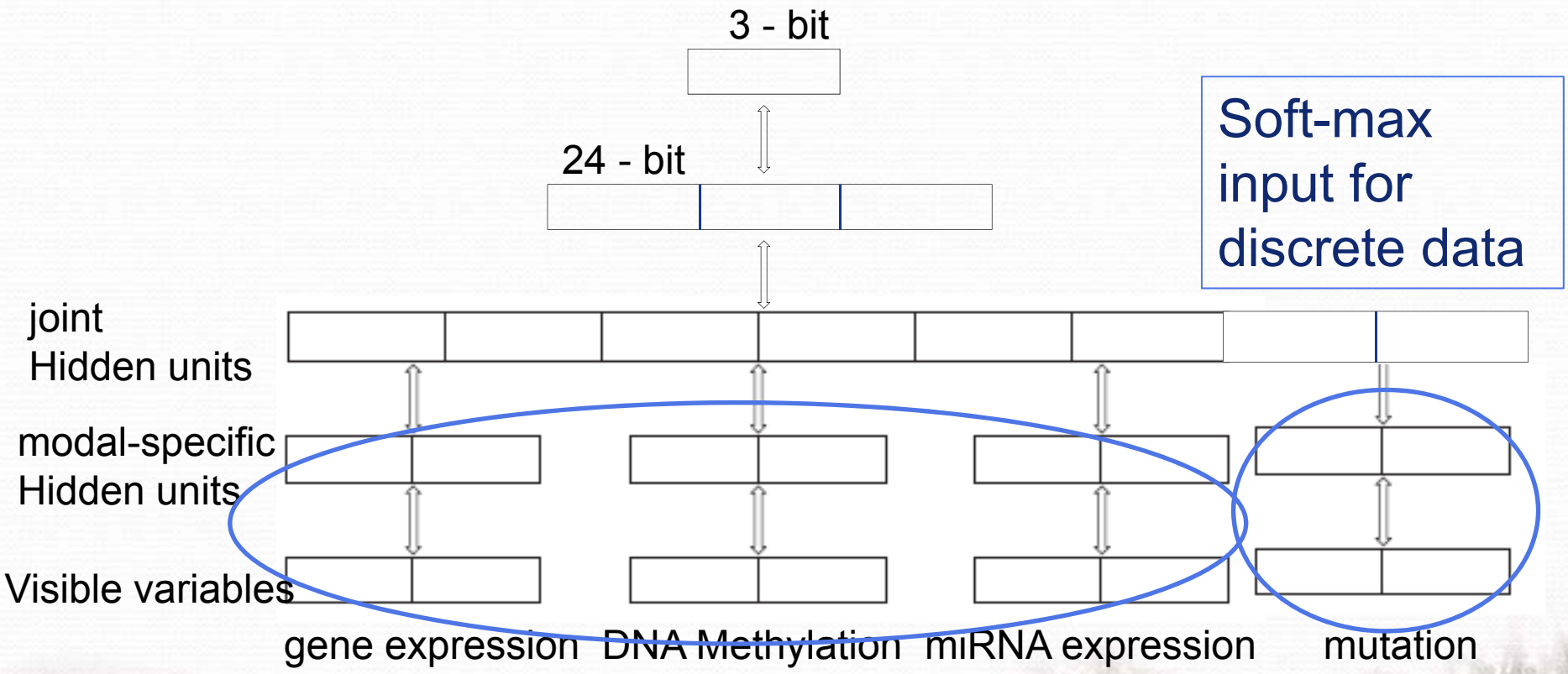
Hugo_Symbol	Variant_Classification	Variant_Type	Tumor_Sample_Barcode
ABHD13	Silent	SNP	TCGA-04-1331-01A-01W...
ACTRT1	Missense_Mutation	SNP	TCGA-04-1331-01A-01W...

4 Variant Type :
1-SNP, 2-DEL, 3-INS, 4-DNP

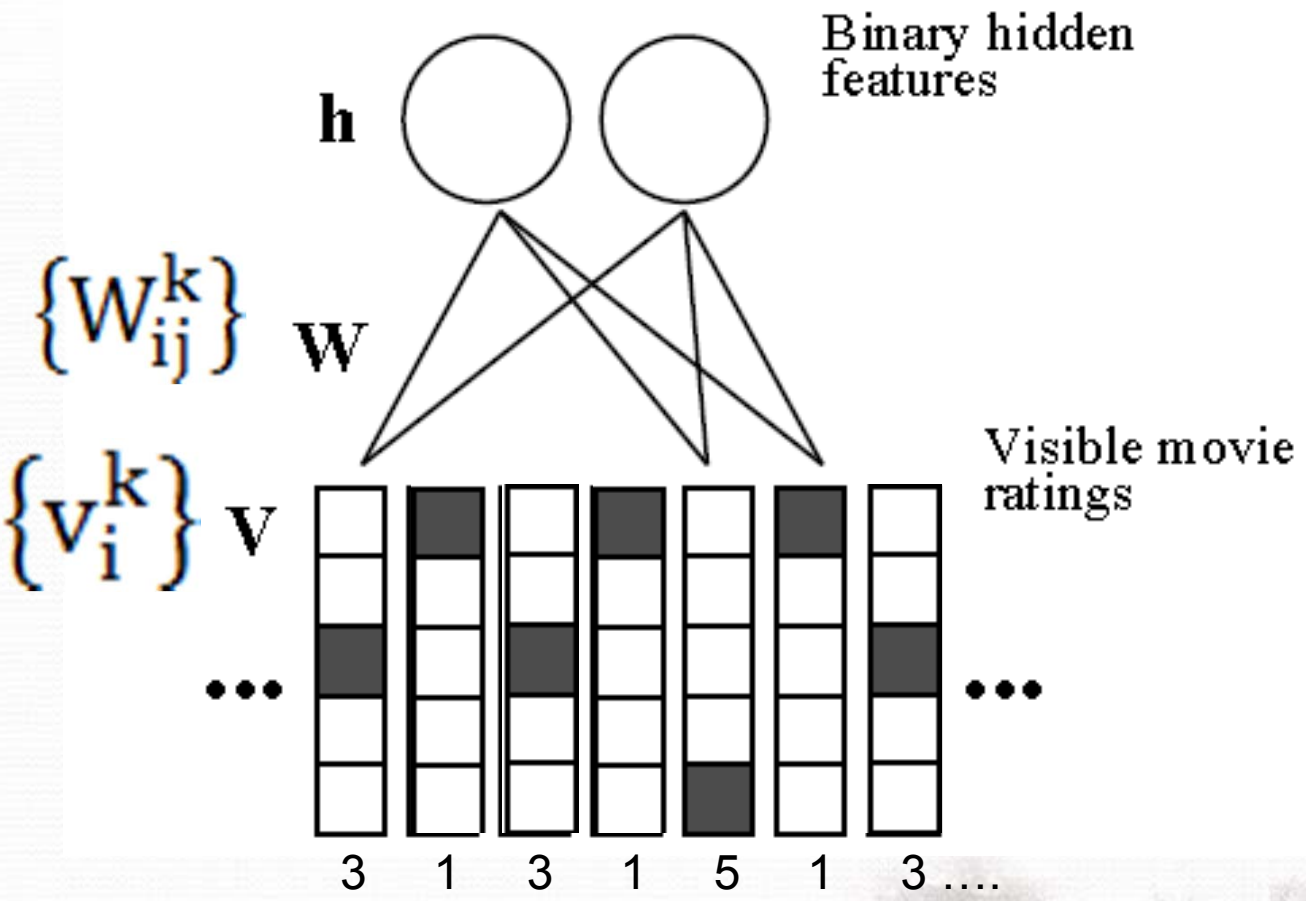
Discrete data

15 Variant Classification:
1-Missense_Mutation, 2-Silent, 3-Nonsense_Mutation, 4-Splice_Site, 5-RNA, 6-In_Frame_Del, 7-In_Frame_Ins, 8-Frame_Shift_Ins, 9-Frame_Shift_Del,10-Intron, 11-Nonstop_Mutation, 12-Translation_Start_Site, 13-IGR, 14- 5' Flank, 15-3'UTR

Multimodal Deep Belief Network



Soft-max unit



Multinomial Distribution

Training

- Greedy layer-wise Contrastive Divergence

$$\Delta W_{ij}^k = \epsilon(<v_i^k h_j>_{data} - <v_i^k h_j>_T)$$

- Gibbs Sampling

$$p(v_i^k = 1|\mathbf{h}) = \frac{\exp\left(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k\right)}{\sum_{l=1}^{K} \exp\left(b_i^l + \sum_{j=1}^{F} h_j W_{ij}^l\right)} \quad (1)$$

Softmax units

$$p(h_j = 1|\mathbf{V}) = \sigma(b_j + \sum_{i=1}^{m}\sum_{k=1}^{K} v_i^k W_{ij}^k) \quad (2)$$

$$P(v_i|\mathbf{h}) = N(b_i + \sigma_i \sum_{j=1}^{g} W_{ij} h_j, \sigma_i^2),$$

Gaussian units

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{m} W_{ij} v_i - a_j\right)},$$

Survival time prediction (On going)

- Cox proportional hazards model with L1 penalized log partial likelihood (LASSO) for feature selection
- Random survival forest (RSF)

Data: hidden layer output of the multimodal Deep Belief Network

Analyzing Resting State Brain Connectivity and its Applications on Psychiatric Diagnosis from Electromagnetic Medical Images(Recently initiated, graduation thesis)

Lab:

Machine Learning and Computational Biology Group, IIIS, Tsinghua University,

Massachusetts General Hospital

Advisor:

Professor Jianyang Zeng

清華大學
Tsinghua University

Background

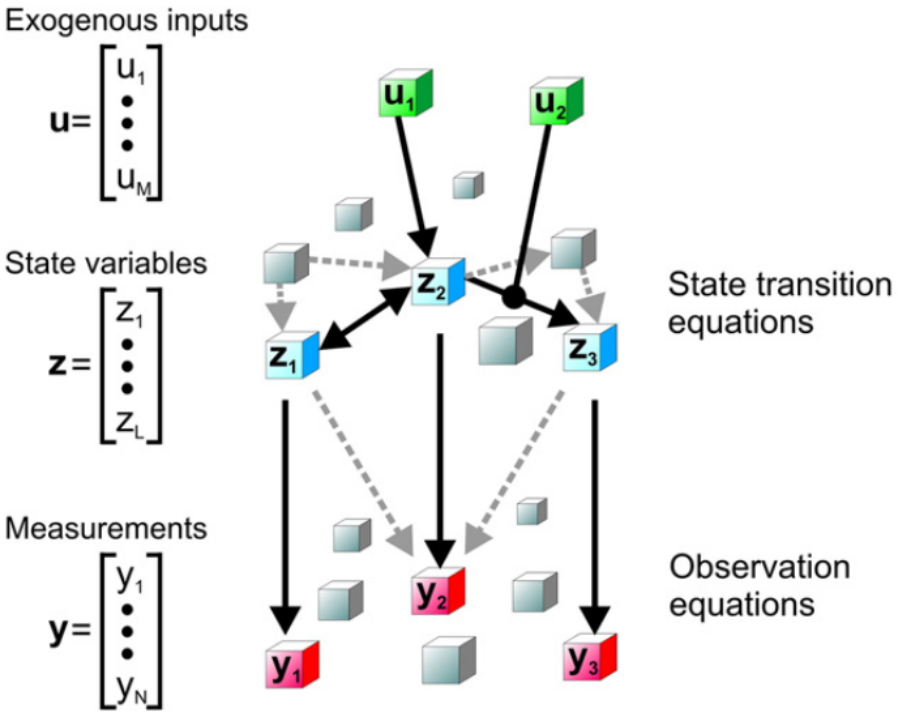
- Studies on human neural networks from fMRI, EEG, MEG data have gained significant progress in recent years. Among them, the studies on resting state neural network are becoming increasingly prevalent, for its strong relevance to anatomical neuroscience, cognitive neuroscience and psychiatric.
- Main focus: brain connectivity
- Granger Causality Analysis (GCA)

  - The seeding procedure does not yield full networks.

  - The pair-wise model does not give the most accurate estimation of edges.

Granger Causality Analysis

- GCA(Roebroeck, et al. 2005)
  - $z[t] = Az[t-1] + Cu[t-1] + \eta(t)$
  - Discrete time model
  - Linear
  - Stochastic
  - $y[t] = Hz[t] + Fu[t] + \varepsilon(t)$
  - Resting state



Goals

- Seeded -> all to all
- Visualize the temporal changes of the map.
- Improve the performance of GCA in speed and computational efficiency.
- Develop APP's on diseases diagnosis using electromagnetic medical images data and GCA.
- Develop crowd-sourcing approaches to solve brain connectivity problem.

Summary

Qualities

- Have a good foundation of machine learning theories and rich experience in implementing learning algorithms (e.g. deep learning, LDA, etc.)
- Familiar with biomedical images, clinical data and genomic data.
- Have good experience in coding and software developing (Cell-organizer).
- Have the capability to learn new things fast, to propose, implement, examine and solve problems independently.

Future research interests

- Developing good learning algorithms.
- Interaction of machine learning and biology.
- Computational genomics, especially the micro mechanism of gene regulation, gene variation and evolution.

清華大學
Tsinghua University

Thank you for your time and consideration!

Reference

[1] T. Peng and R.F. Murphy (2011) Image-derived, Three-dimensional Generative Models of Cellular Organization. Cytometry Part A 79A:383-391.

[2] Muxuan Liang, Zhizhong Li, Ting Chen and Jianyang Zeng*. Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2015.