Extraction of Cell-Cell interaction by DMR topic modeling

Takaho Tsuchiya

Ozaki Laboratory

University of Tsukuba

Projects in our laboratory

https://sites.google.com/view/ozakilab-jp/research

Outline

Background:

- Single cell spatial transcriptome technology is emerging
- Previous analysis methods can't capture Cell-Cell interaction fully
- Our results suggested importance of spatial Cell-Cell interaction

Methods:

- Topic model in natural language processing can be effective
  What is topic model?? How can it work for transcriptome data??
- Our idea to extract cell-cell interaction - DMR topic model

Research plan

Spatial single cell transcriptome technology

seqFISH+ paper and so on

Spatial single cell transcriptome data is emerging.

LETTER

https://doi.org/10.1038/s41586-019-1049-y

Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+

Chee-Huat Linus Eng[1], Michael Lawson[2], Qian Zhu[3], Ruben Dries[3], Noushin Koulena[2], Yodai Takei[2], Jina Yun[2], Christopher Cronin[2], Christoph Karp[2], Guo-Cheng Yuan[3] & Long Cai[2]*

HTS

GENE EXPRESSION

Spatial transcriptomics coming of age

In their study, Rodriques, Stickels et al. enabled high-resolution spatial capture with Drop-seq beads. In the original Drop-seq method, single cells

PNAS

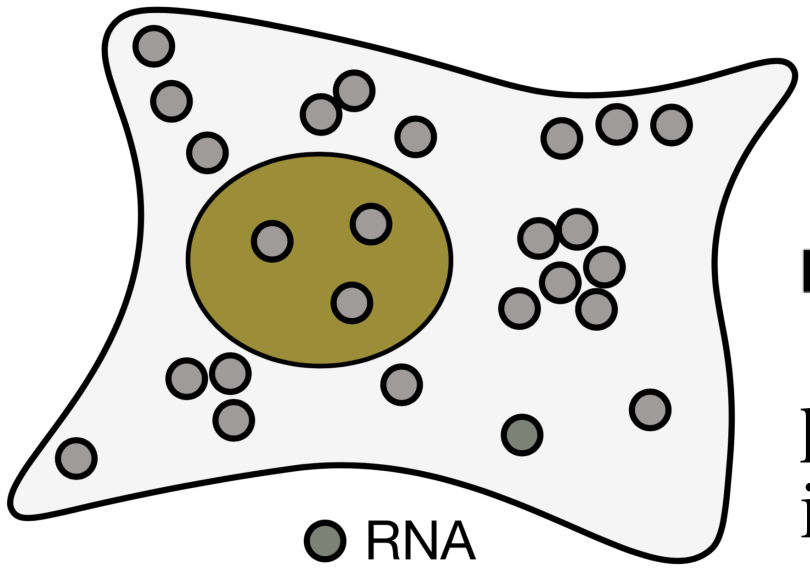
Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression

Chenglong Xia[a,b,c,1], Jean Fan[a,b,c,1], George Emanuel[a,b,c,1], Junjie Hao[a,b,c], and Xiaowei Zhuang[a,b,c,2]

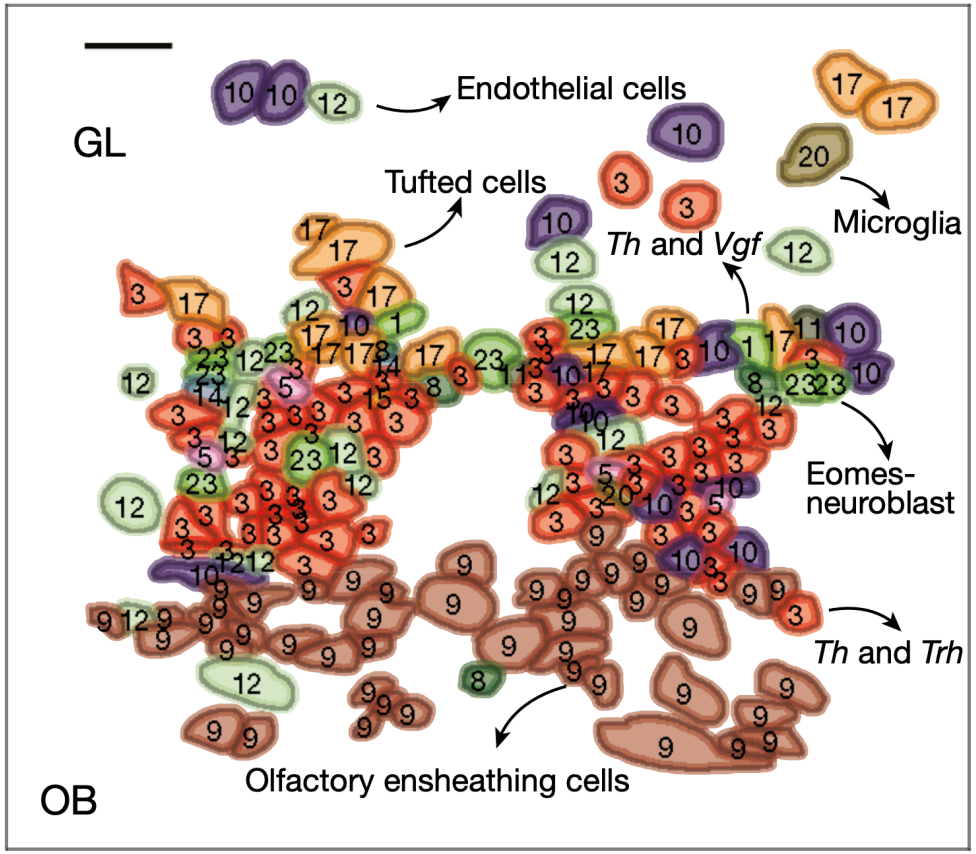
[a]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138; [b]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138; and [c]Department of Physics, Harvard University, Cambridge, MA 02138

seqFISH+ paper and so on

Spatial single cell transcriptome data appeared.



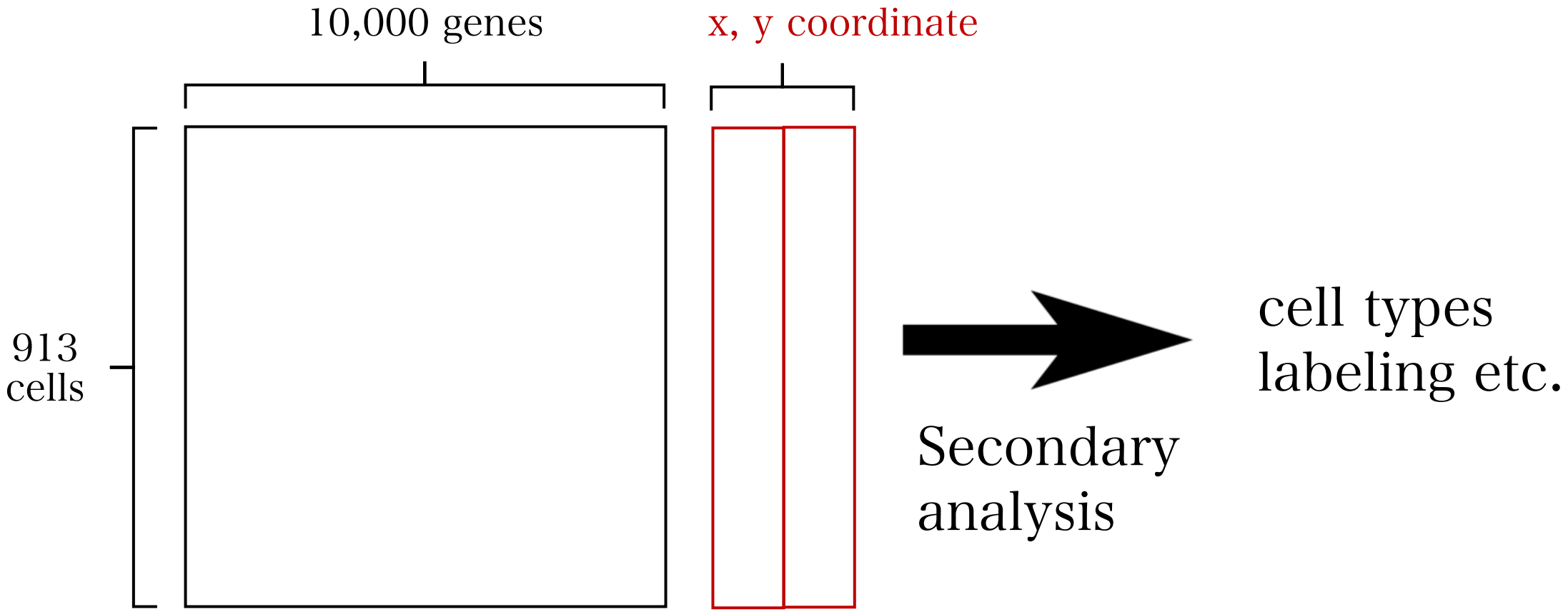
probing
in each cell



[Eng et al., Nature, 2019 April]

seqFISH+ paper and so on

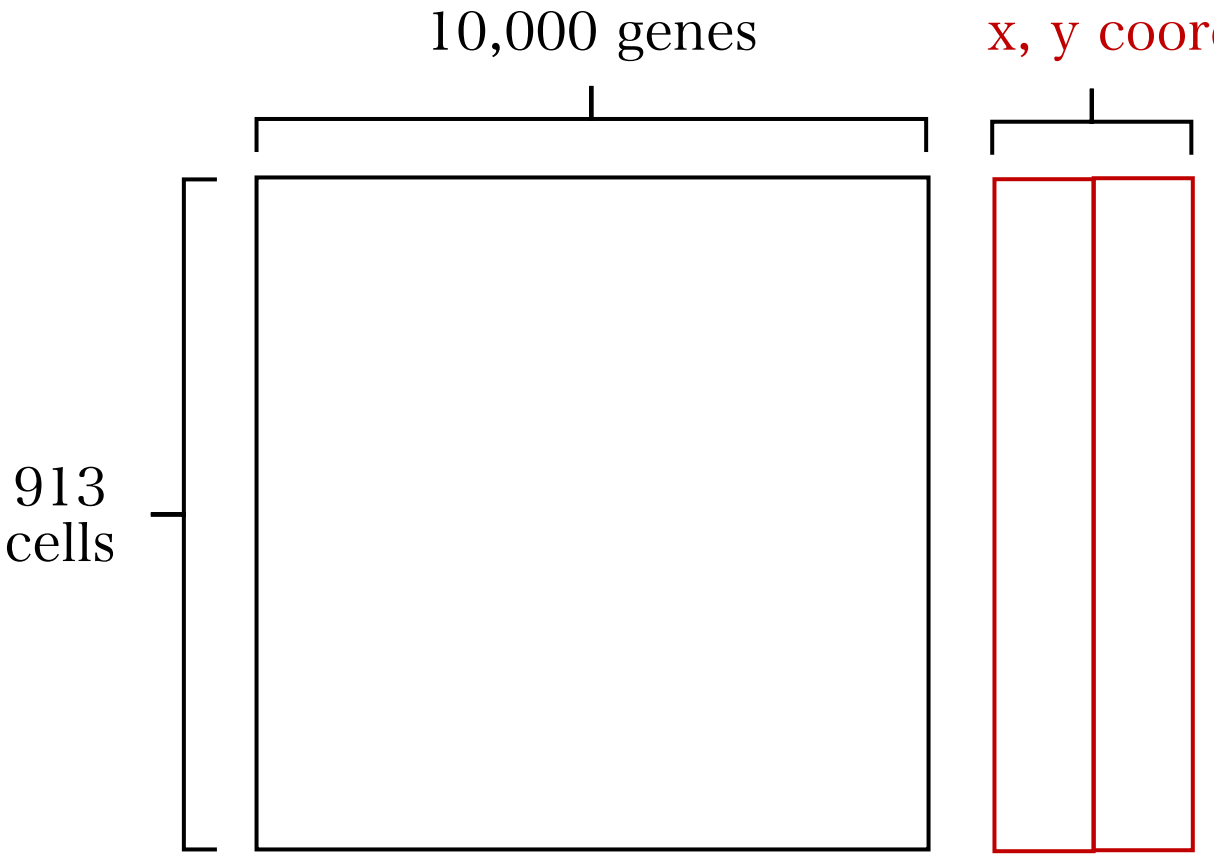
Spatial single cell transcriptome data appeared.



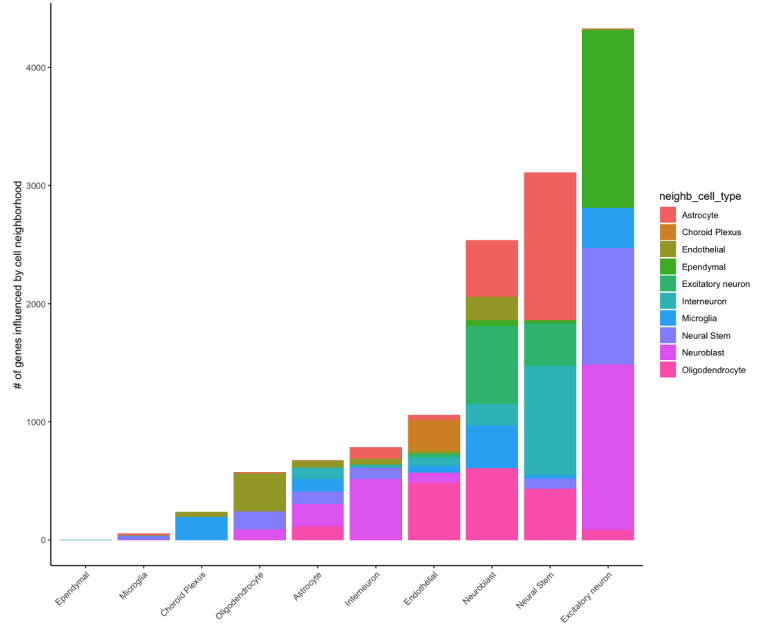
[Eng et al., Nature, 2019 April]

Our motivation

We've already suggested importance of Cell-Cell interaction.

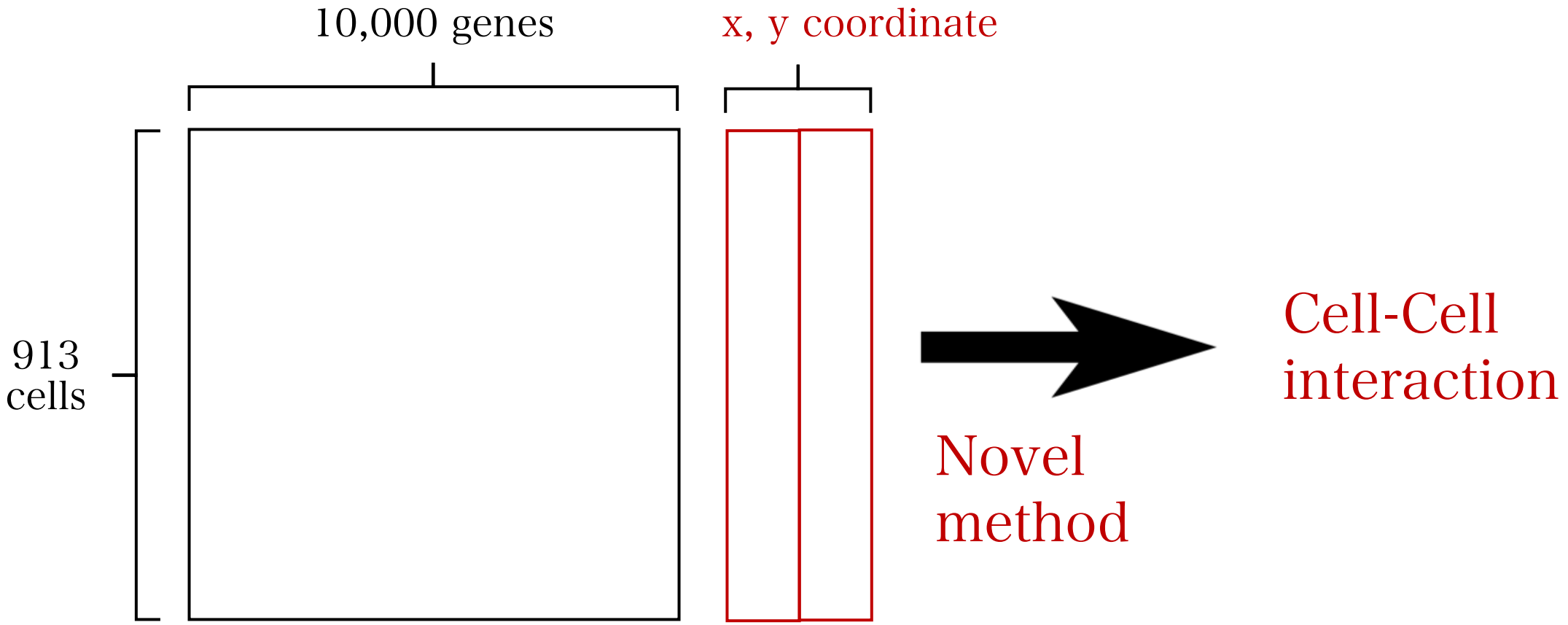


Secondary analysis



Our motivation

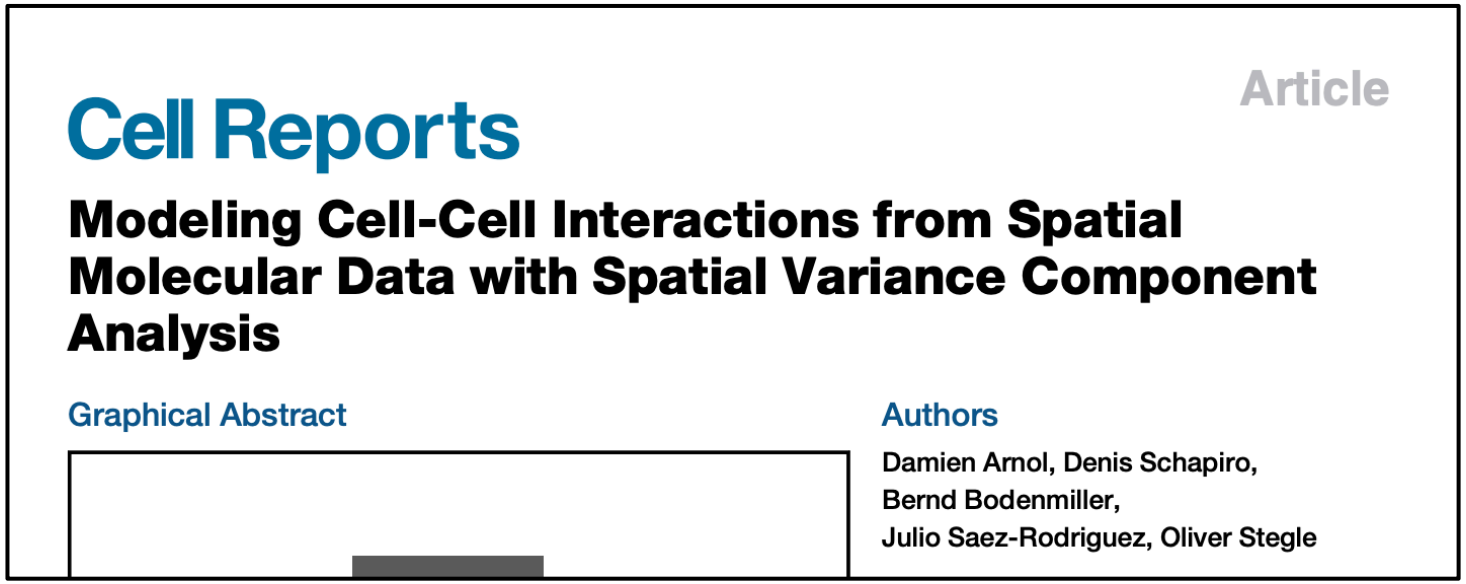
Develop a novel method to extract Cell-Cell interaction



Single cell RNA-seq can't capture Cell-Cell interaction fully.

Our motivation

Oliver group developed methods to analyze cell-cell interaction by using spatial information...
However, interpretability is not sufficient...



Article

Cell Reports

Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis

Graphical Abstract

Authors

Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, Oliver Stegle

Our approach “DMR topic modeling”

Latent Dirichlet Allocation (LDA) topic model

LDA is a generative probabilistic model

1. Data are assumed to be observed from a generative probabilistic process that includes hidden variables.
   - *In text, the hidden variables are the thematic structure.*
2. Infer the hidden structure using posterior inference
   - *What are the topics that describe this collection?*
3. Situate new data into the estimated model.
   - *How does a new document fit into the topic structure?*

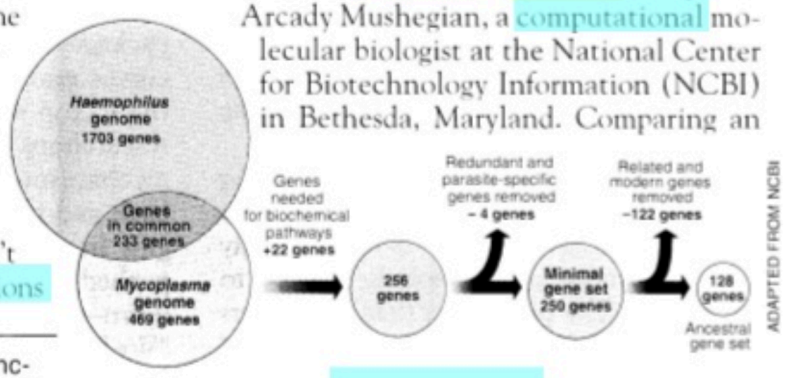
https://www.slideshare.net/hustwj/probabilistic-topic-models

LDA was developed for analyzing texts

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

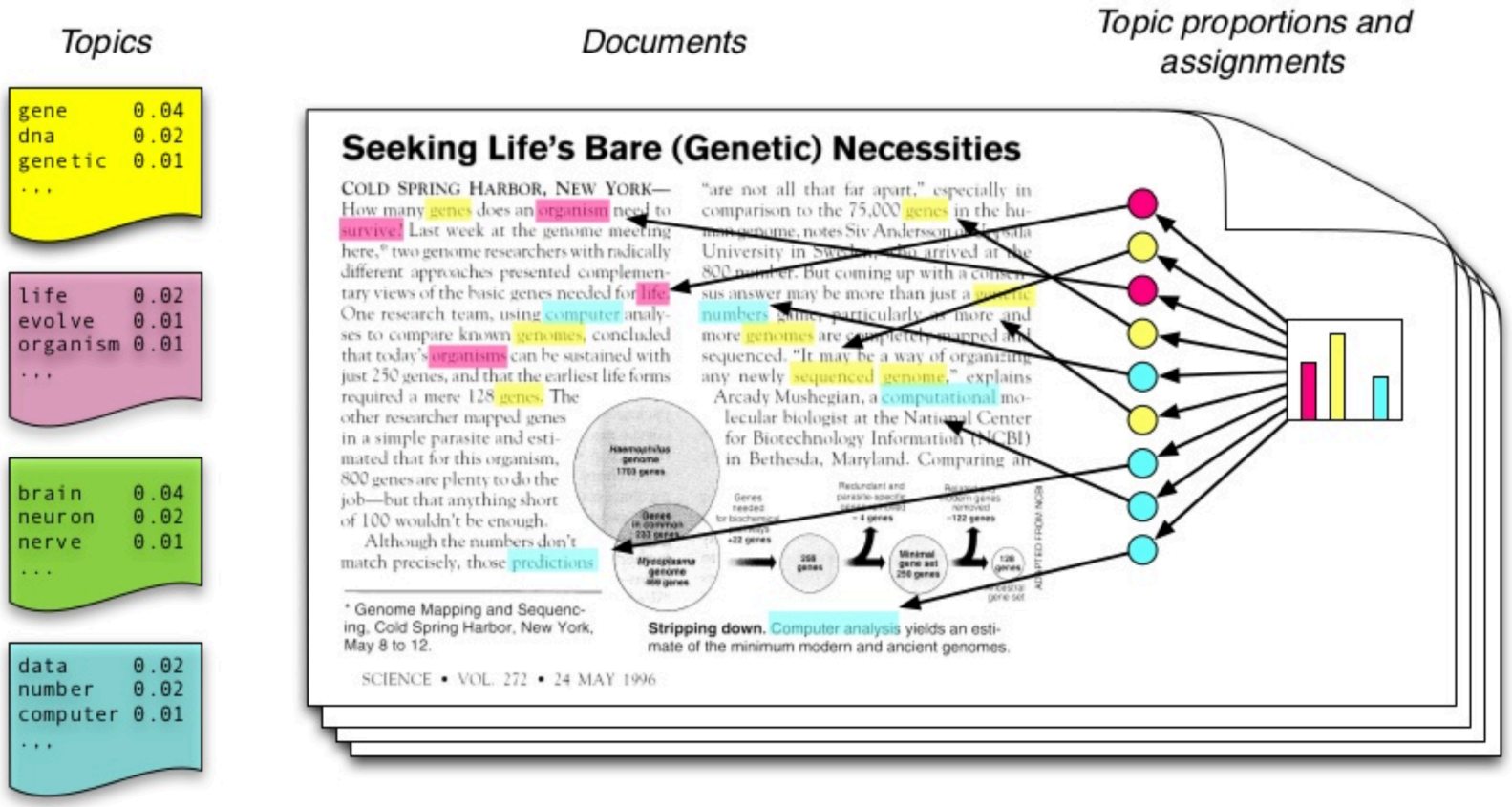
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

https://www.slideshare.net/hustwj/probabilistic-topic-models

Simple intuition: Documents exhibit multiple topics.

Generative process of LDA



https://www.slideshare.net/hustwj/probabilistic-topic-models

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Generative process of LDA

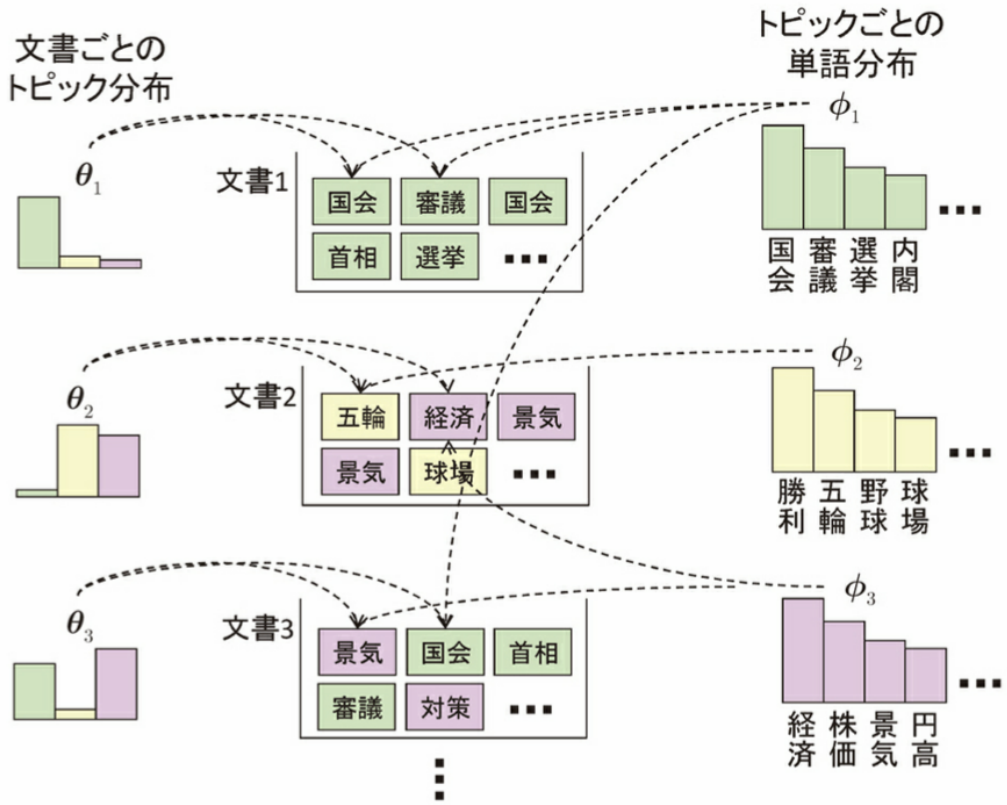


図 4.2 トピックモデルによる文書集合の生成例．単語ごとにトピックが割り当てられます．文書ごとのトピック確率 $\boldsymbol{\theta}_d$ に従って単語のトピック（色）を決め，そのトピックの単語分布 $\boldsymbol{\phi}_k$ に従って語彙を決めます．

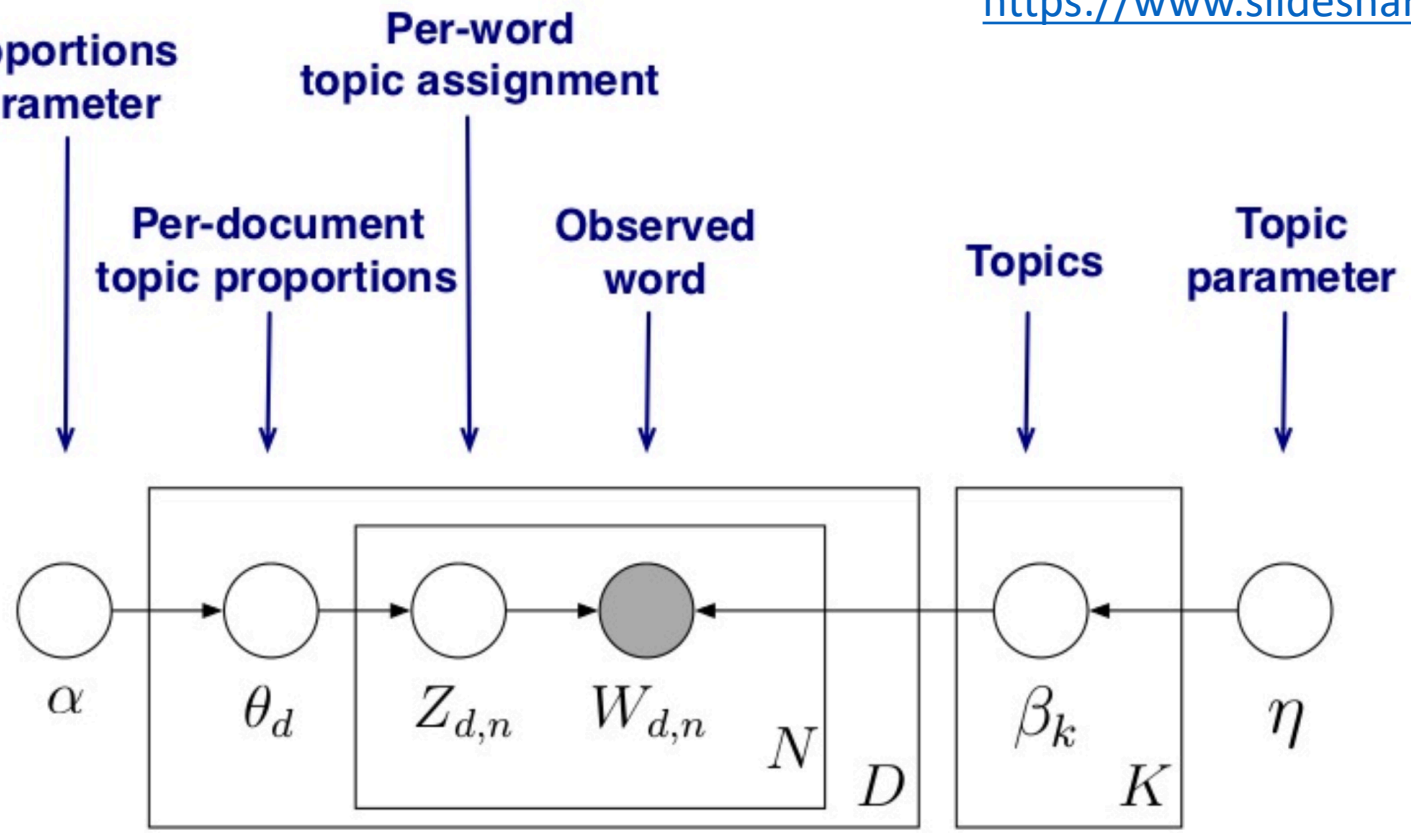
表 4.1 本節で用いる記号．表 2.1 に記載されている記号も再掲します．

記号	説明
D	文書数
N_d	文書 d に含まれる単語数（文書長）
V	全文書のなかで現れる単語の種類数（語彙数）
$\mathbf{W}$	文書集合
$\mathbf{w}_d$	文書 d
w_{dn}	文書 d の n 番目の単語
K	トピック数
N_k	文書集合全体でトピック k が割り当てられた単語数
N_{dk}	文書 d でトピック k が割り当てられた単語数
N_{kv}	文書集合全体で語彙 v にトピック k が割り当てられた単語数
θ_{dk}	文書 d でトピック k が割り当てられる確率
ϕ_{kv}	トピック k のとき語彙 v が生成される確率

岩田具治『トピックモデル』講談社 p.57

Graphical model of LDA

https://www.slideshare.net/hustwj/probabilistic-topic-models

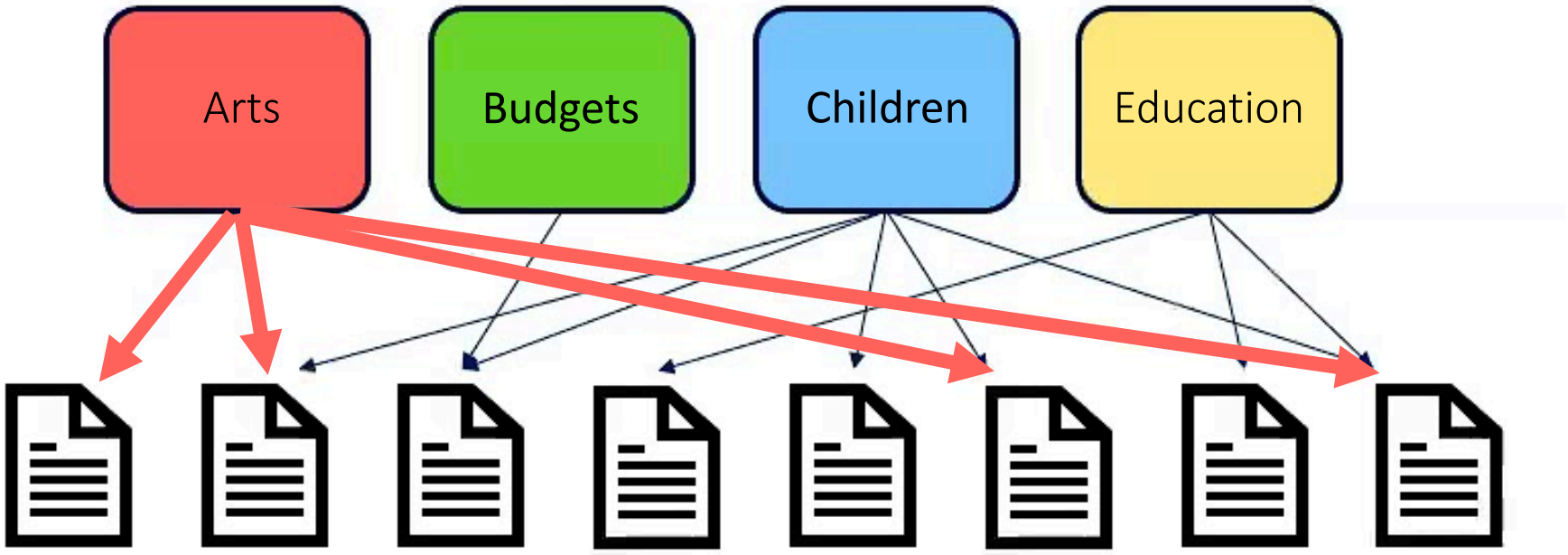


- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

Intuitive visualization

"A topic model is
a type of statistical model for discovering the abstract "topics"
that occur in a collection of documents." - Wikipedia

Case of documents:

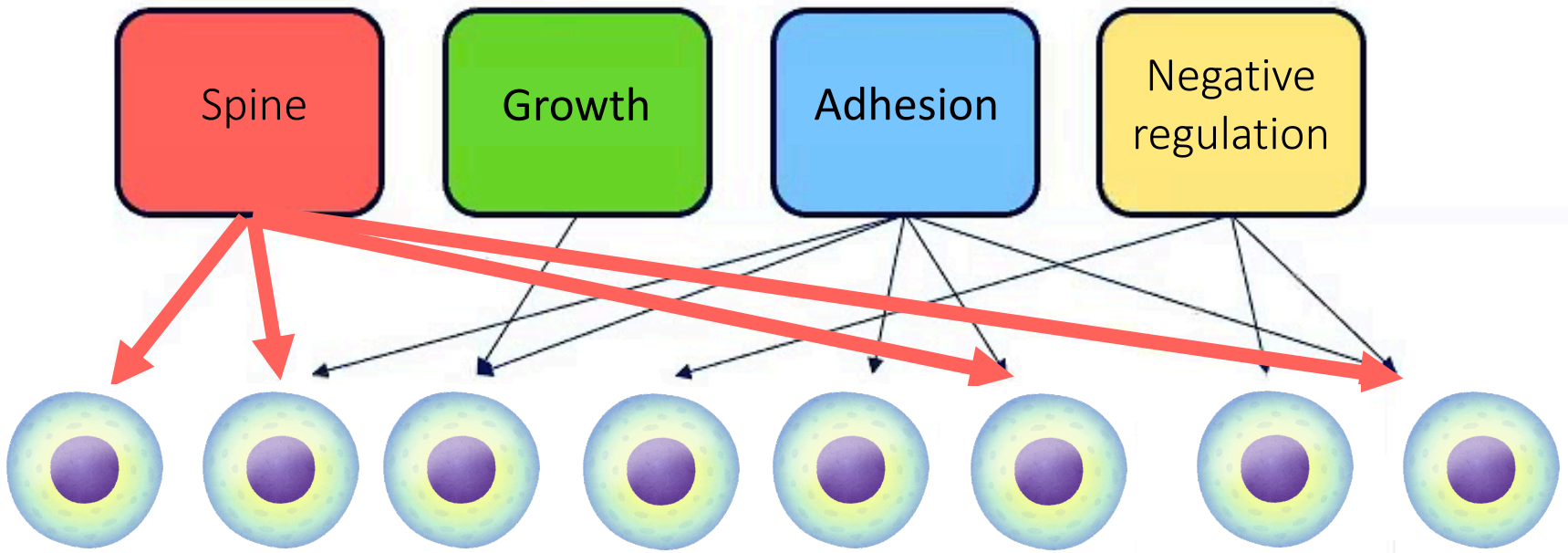


These documents have relatively abundant words related to "Arts".

How about biology??

"A topic model is
a type of statistical model for discovering the abstract "topics"
that occur in a collection of documents." - Wikipedia

Case of cells: # document → cell / words → expressed gene counts



These cells have relatively abundant expressed gene counts related to "Spine".

Topic model is used in bioinformatics

Liu *et al. SpringerPlus (2016) 5:1608*
DOI 10.1186/s40064-016-3252-8

SpringerPlus

REVIEW Open Access

An overview of topic modeling and its current applications in bioinformatics

Lin Liu[1,2], Lin Tang[3], Wen Dong[1], Shaowen Yao[4*] and Wei Zhou[4*]

SOFTWARE Op

CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data

David A. duVerle[1,5*], Sohiya Yotsukura[2], Seitaro Nomura[3], Hiroyuki Aburatani[3] and Koji Tsuda[1,4,5*]

A survey of topic model in bioinformatics

WACODE#3

東京大学　新領域創成科学研究科
情報生命科学専攻　岩崎研究室　博士三年
福永　津嵩

1 of 19

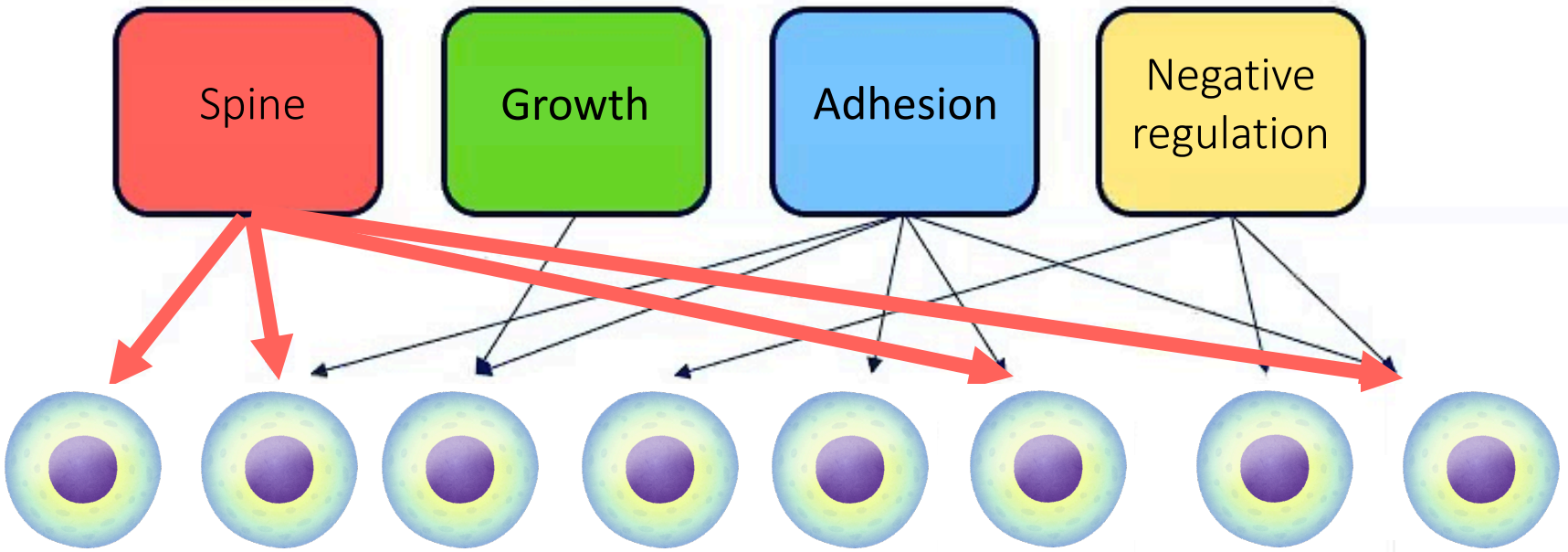
A survery of topic model in bioinformatics　1,465 views

https://www.slideshare.net/tsukasafukunaga5/a-survery-of-topic-model-in-bioinformatics-55031864

Deficiency of ordinary LDA

We can't extract Cell-Cell extraction...
No arrow exists between cells.

Case of cells: # document → cell / words → expressed gene counts



→ Our approach "DMR topic modeling"

Dirichlet Multinomial Regression (DMR) topic model

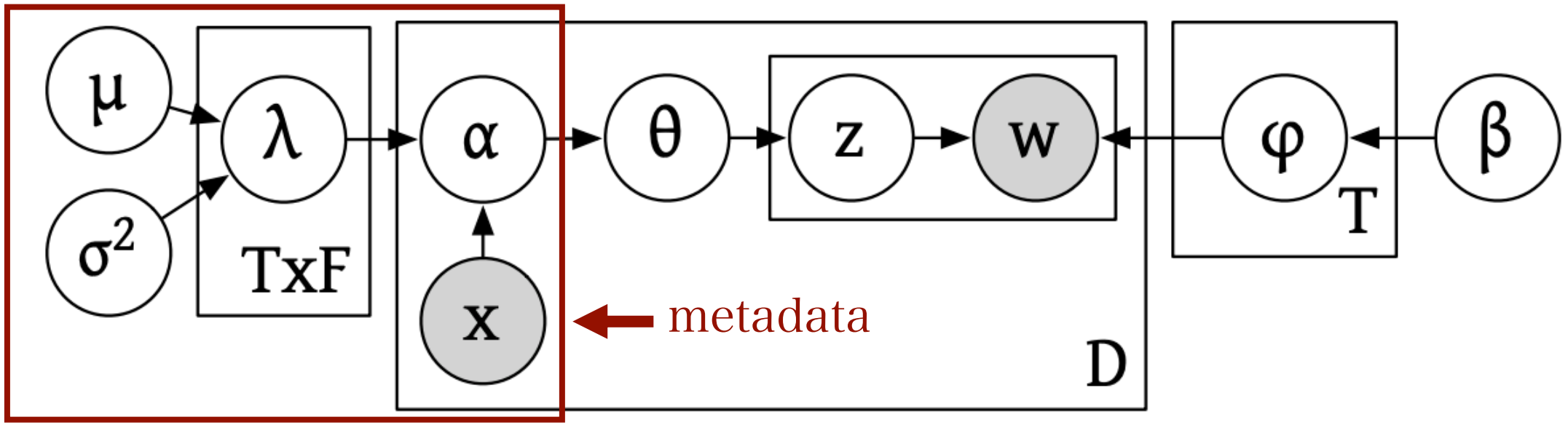
DMR topic model

We can capture weight derived from metadata.

[Mimmo and McCallium, arXiv, 2013]

Extension of LDA

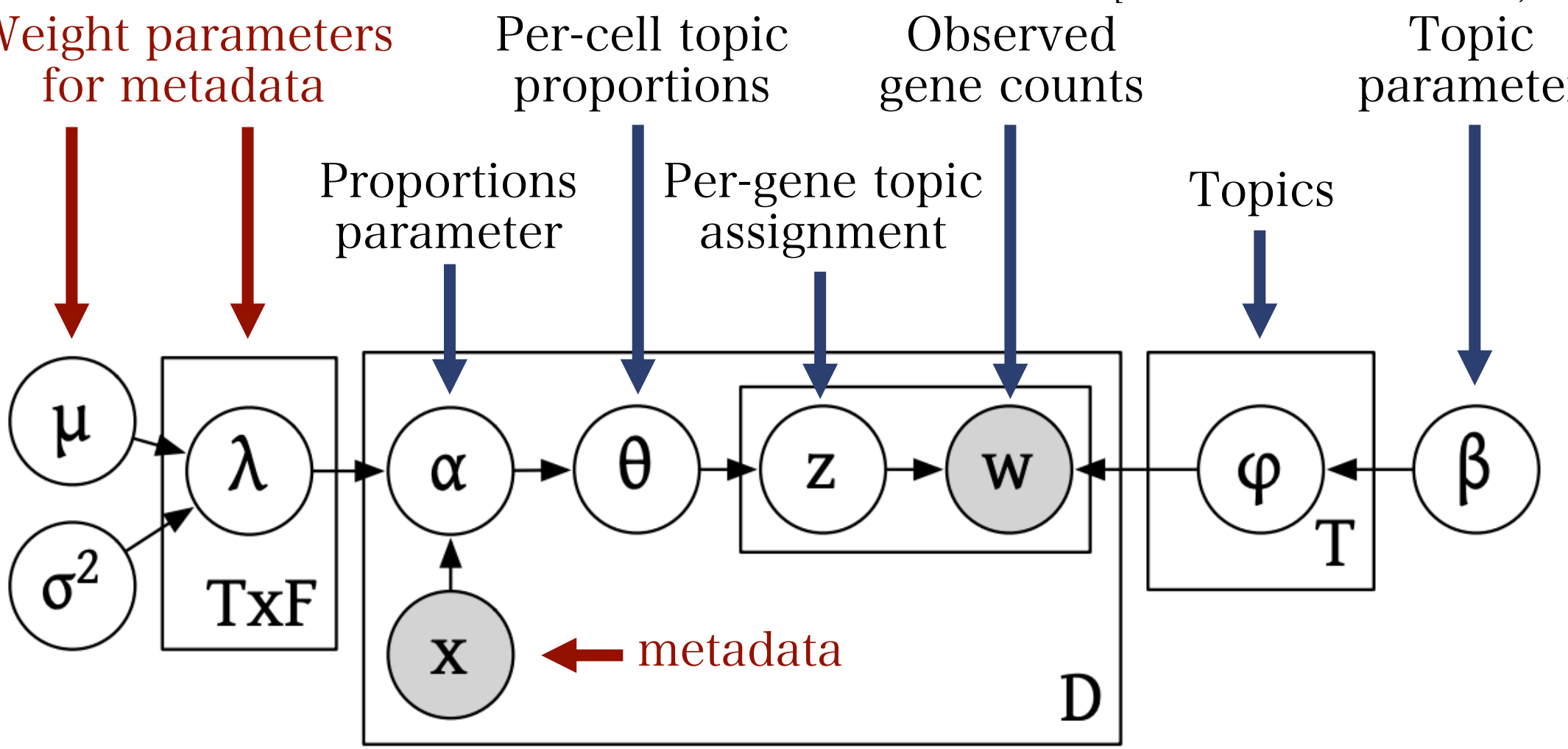
Weight parameters of metadata



DMR topic model

We can capture weight derived from metadata.

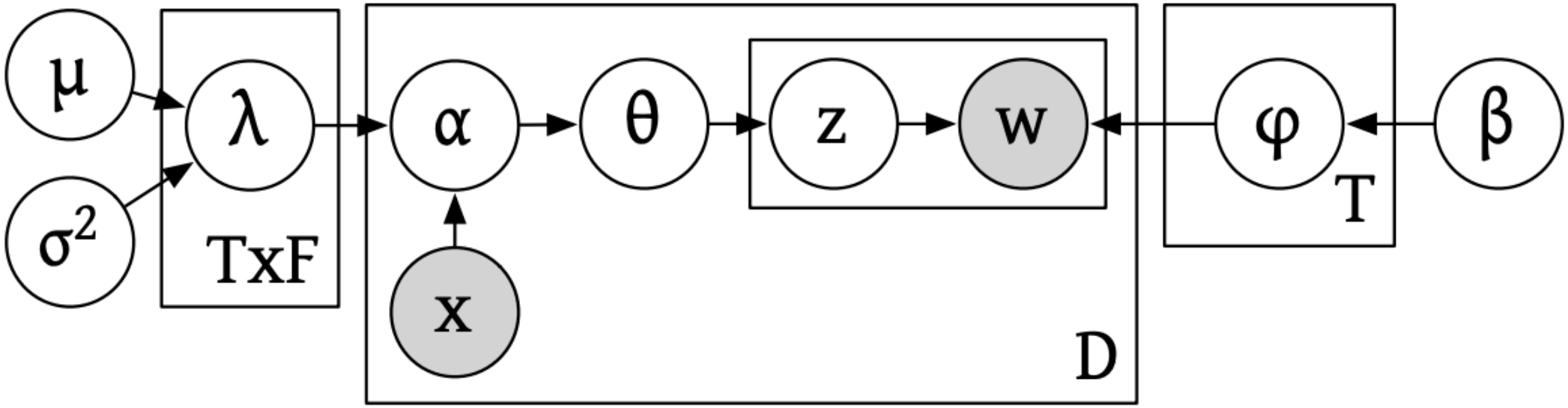
[Mimmo and McCallium, arXiv, 2013]



DMR topic model

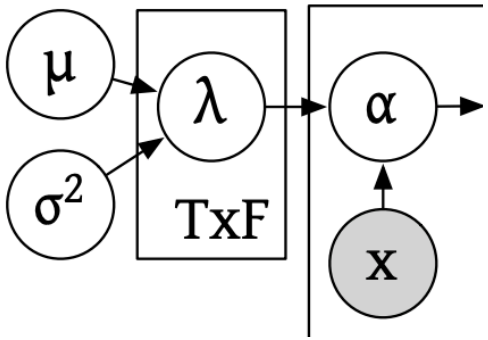
Generative process:

1. For each topic t,
   (a) Draw $\boldsymbol{\lambda}_t \sim \mathcal{N}(0, \sigma^2 I)$
   (b) Draw $\boldsymbol{\phi}_t \sim \mathcal{D}(\beta)$
2. For each document d,
   (a) For each topic t let $\alpha_{dt} = \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t)$.
   (b) Draw $\boldsymbol{\theta}_d \sim \mathcal{D}(\boldsymbol{\alpha}_d)$.
   (c) For each word i,
      i. Draw $z_i \sim \mathcal{M}(\boldsymbol{\theta}_d)$.
      ii. Draw $w_i \sim \mathcal{M}(\boldsymbol{\phi}_{z_i})$.



DMR topic model

We can capture weight derived from metadata.

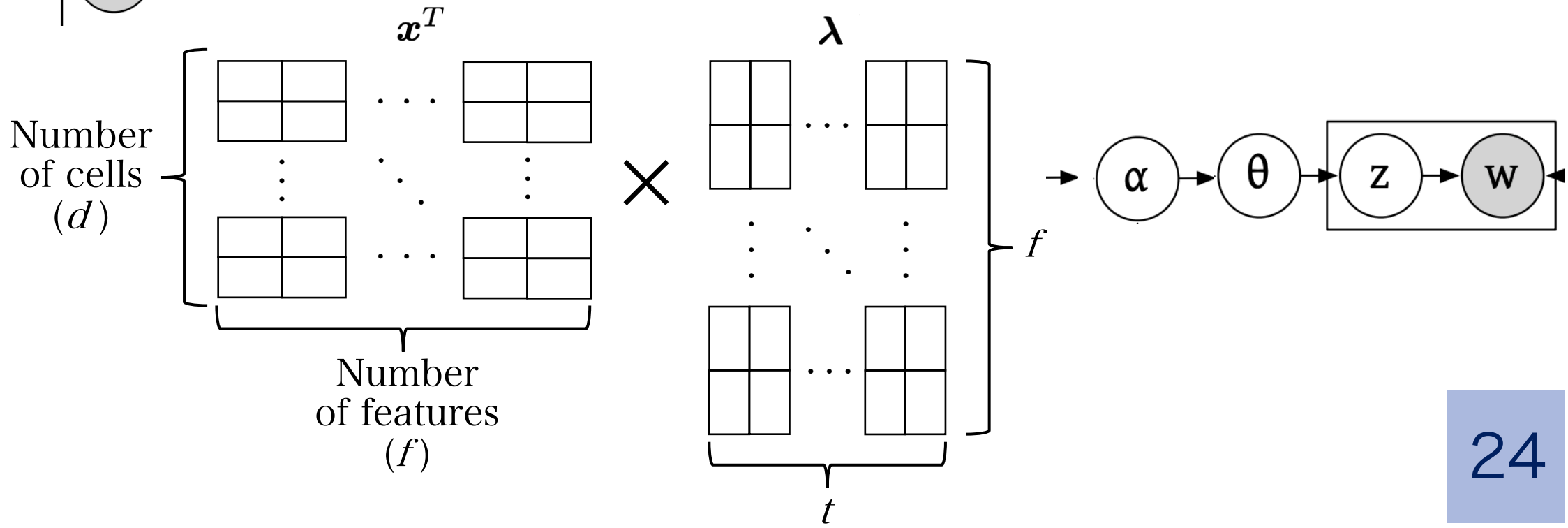


2. For each document d,

   (a) For each topic t let $\alpha_{dt} = \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t)$.

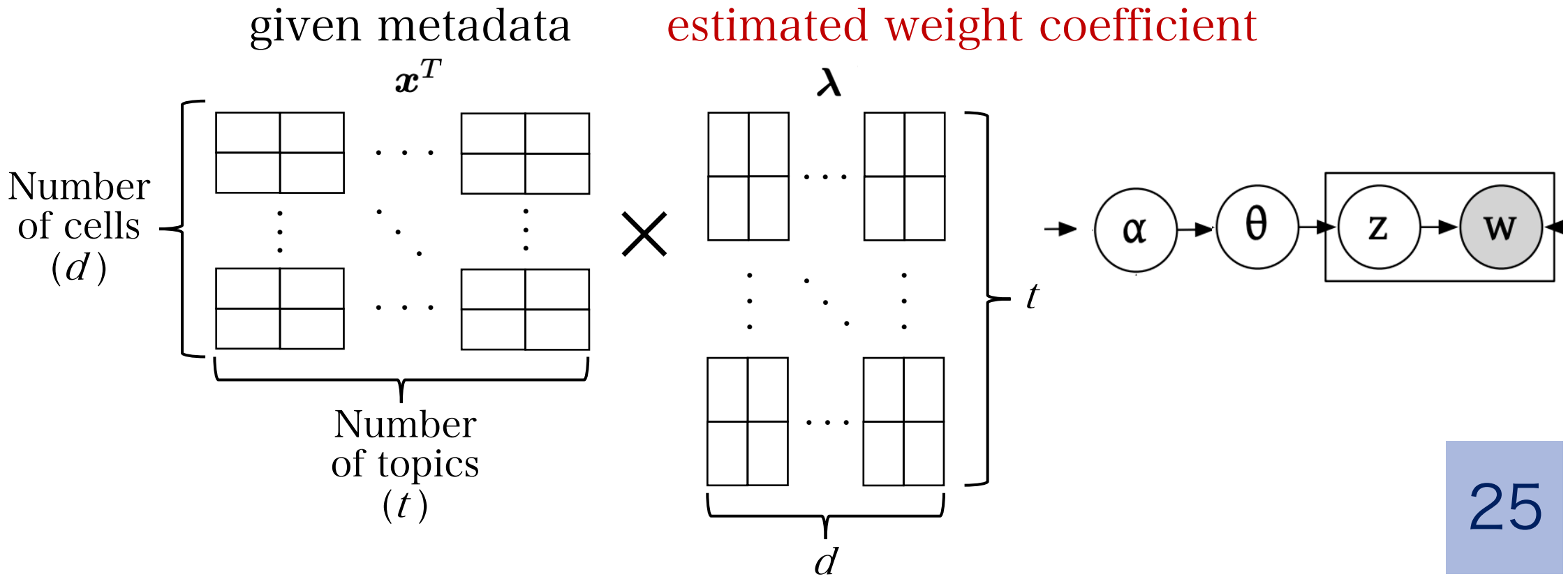
given metadata

estimated weight coefficient



DMR topic model

We can analyze relationships between given meta data and weight coefficient as cell-cell interaction.



Research plan

Research plan roughly

1. Evaluate DMR topic model by using simulated data
2. Apply DMR topic model to real data of seqFISH+ etc.
3. Compare to other methods in #1 and 2

4. Interpret biological meanings and blush up our method
   # additional analysis may be needed

5. Packaging our method

6. Write paper and submit

Acknowledgement

Dr. Ozaki – conceived this project
and all members of our laboratory