Generative models and how it can be used in bioinformatics

2022-05-03

Ping-Han Hsieh

What if we all lived in a virtual reality

- Let us assume we all live in a virtual reality.

What if we all lived in a virtual reality

- Let us assume we all live in a virtual reality.

- Where I am the creator of this world.
  - I get to decide the physical law
  - I get to decide what people can observe.

What if we all lived in a virtual reality

- Let us assume we all live in a virtual reality.

- Where I am the creator of this world.
  - I get to decide the physical law
  - I get to decide what people can observe.

- I am trying to make sure nobody realizes they live in the virtual reality.
  - The physical law need to be understandable, so people find this world interesting.
  - But not too simple, so no one realizes they live in the virtual reality.

What if we all lived in a virtual reality

- Let us assume we all live in a virtual reality.

- Where I am the creator of this world.
  - I get to decide the physical law
  - I get to decide what people can observe.

- I am trying to make sure nobody realizes they live in the virtual reality.
  - The physical law need to be understandable, so people find this world interesting.
  - But not too simple, so no one realizes they live in the virtual reality.

- So, I can keep my reign forever.

Virtual Universe #01

- Let me put the hint to infer the physical law in the virtual universe.

- And the law is very simple:
  - Depend on the expression of Transcription Factor A, the expression of Gene A will change in a linear way.
  - There are only 16 cells in this universe, each has only Transcription Factor A, Transcription Factor B, Transcription Factor C and Gene A and you can observe the exact expression of each molecule.
  - And I let you know transcription factors regulate genes.
  - Gene expression follows Gaussian distribution.

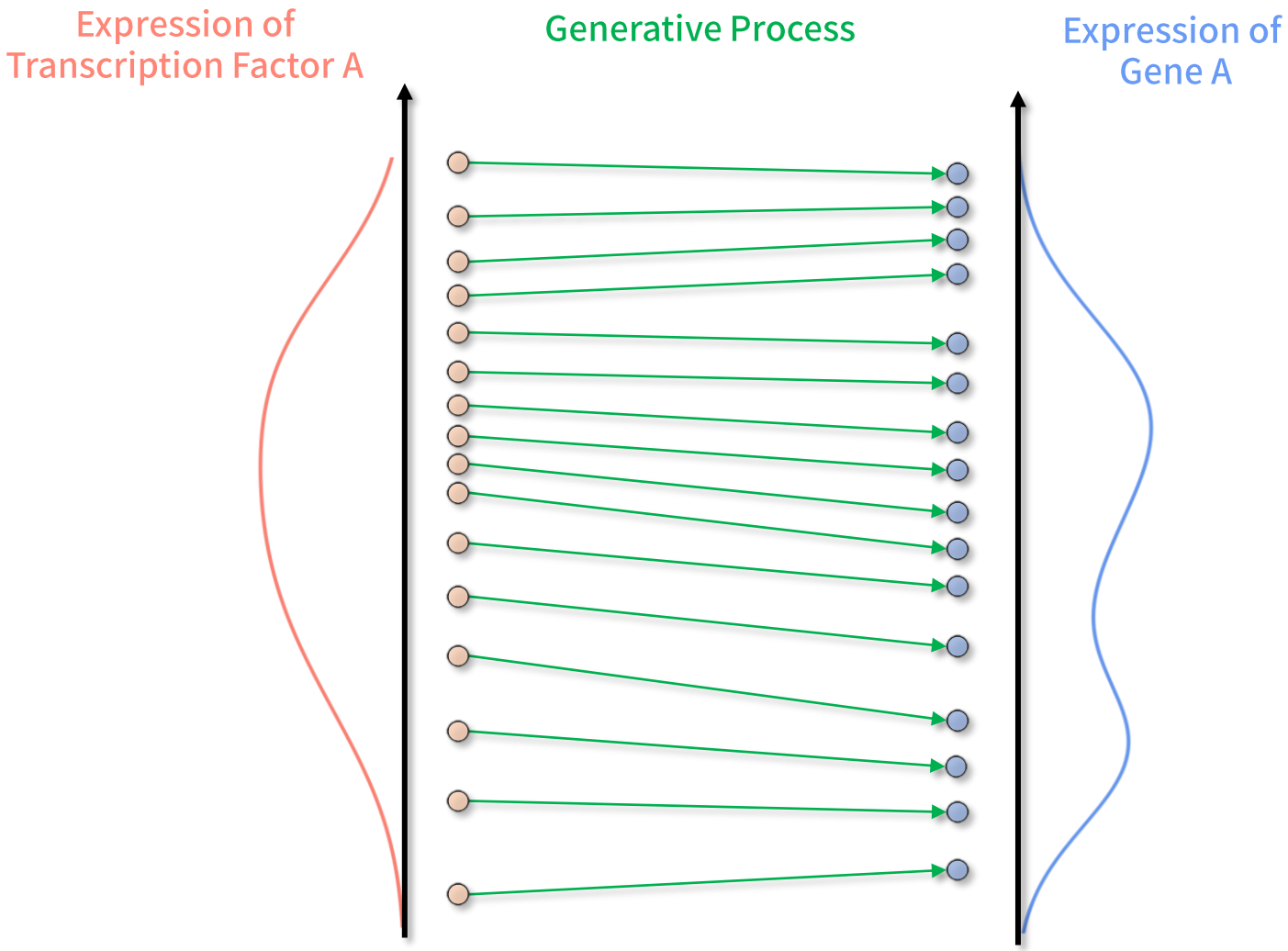
Virtual Universe #01

- Let me put the hint to infer the physical law in the virtual universe.

- And the law is very simple:
  - Depend on the expression of Transcription Factor A, the expression of Gene A will change in a linear way.
  - There are only 16 cells in this universe, each has only Transcription Factor A, Transcription Factor B, Transcription Factor C and Gene A and you can observe the exact expression of each molecule.
  - And I let you know transcription factors regulate genes.
  - Gene expression follows Gaussian distribution.

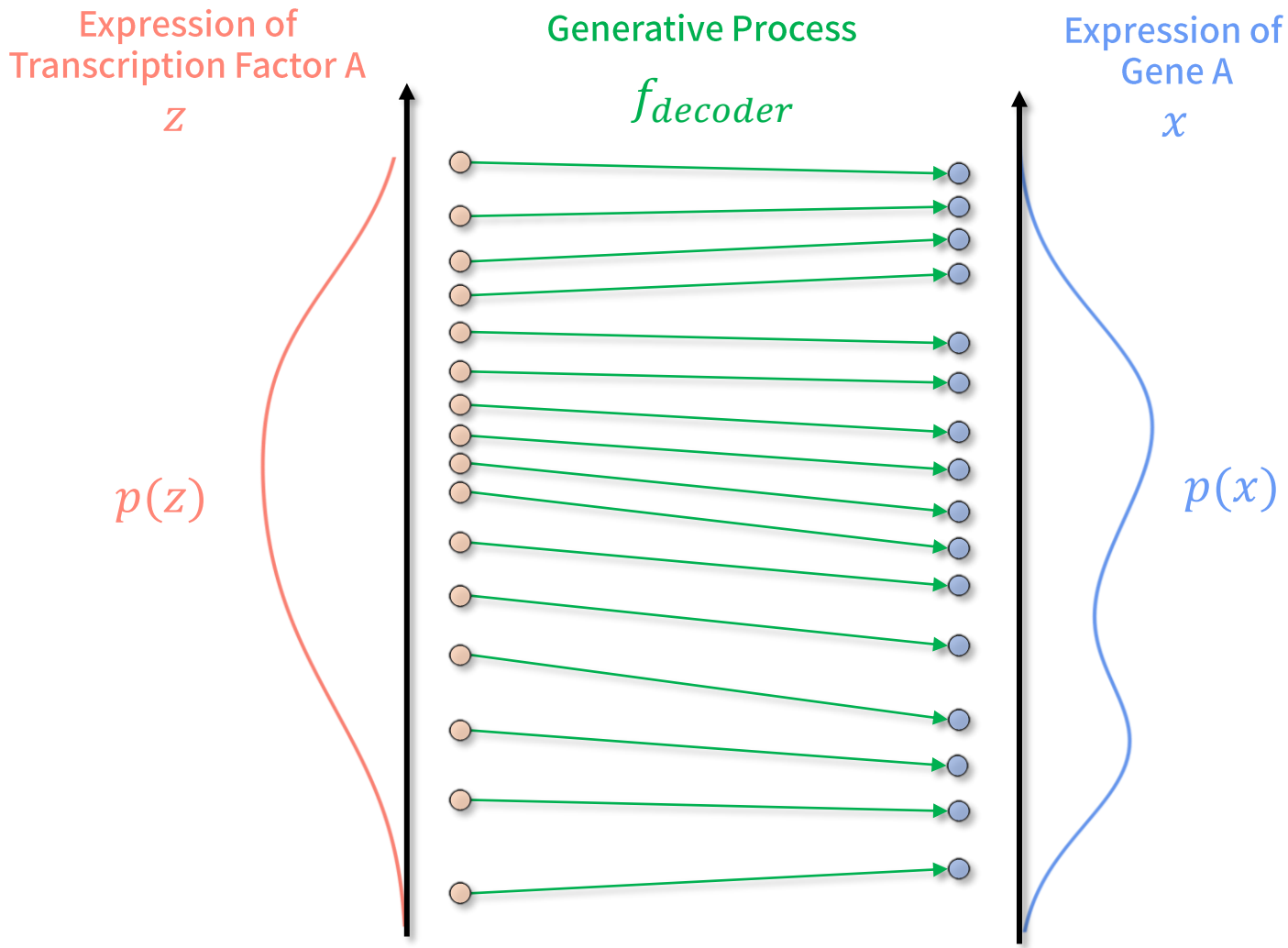
- Hope no one will find out the secret of this universe.

Virtual Universe #01



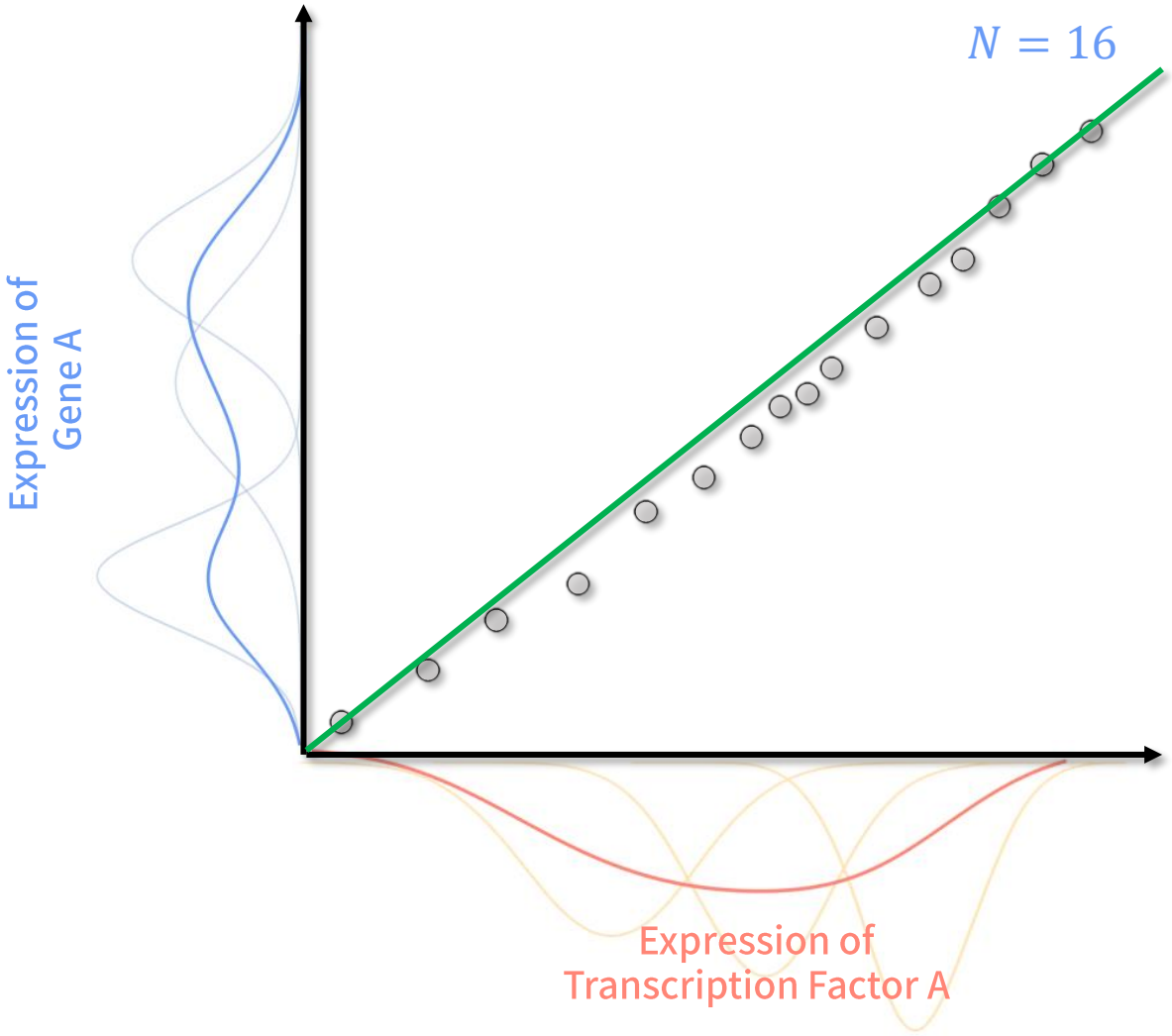
Virtual Universe #01

Term Definition



Virtual Universe #01

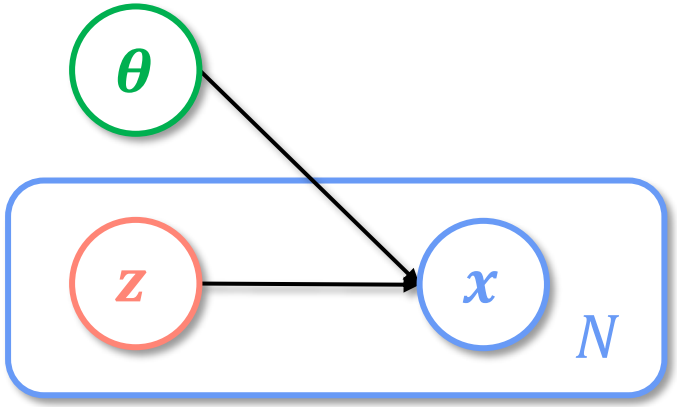
Linear Generative Process



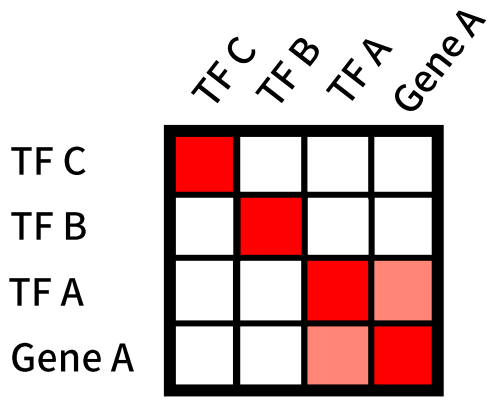
$$x = f_{decoder}(z)$$

$$x = zw + b$$

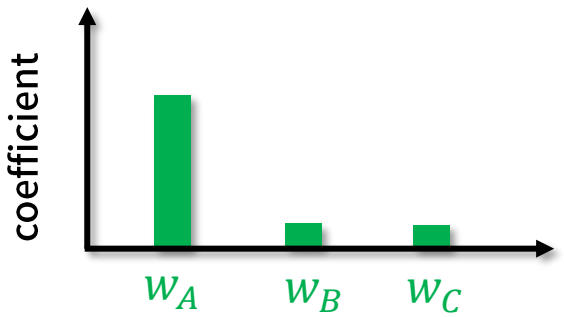
$$\theta = \{w, b\}$$



From the Data Analysist Point of View



$$Gene\ A = z_A w_A + z_B w_B + z_C w_C + b$$



- Gather information about the molecular systems (features)
- Build machine learning model or perform correlation analysis.
- Found that TF A can be used to predict the expression of Gene A (or the highest correlation coefficient).
- Conclude that in this molecular system, TF A might relate to the production of Gene A.

Virtual Universe #02

- Soon people will find out they are controlled by me. Let me make the world more complicated.

Virtual Universe #02

- Soon people will find out they are controlled by me. Let me make the world more complicated.

- The new law:
  - Depend on the expression of Transcription Factor A, the expression of Gene A will change in a non-linear way (but in groups).
  - There are only 16 cells in this universe, each has only Transcription Factor Aand Gene A and you can observe the exact expression of each molecule.
  - And I let you know transcription factors regulate genes.
  - Gene expression follows Gaussian distribution.

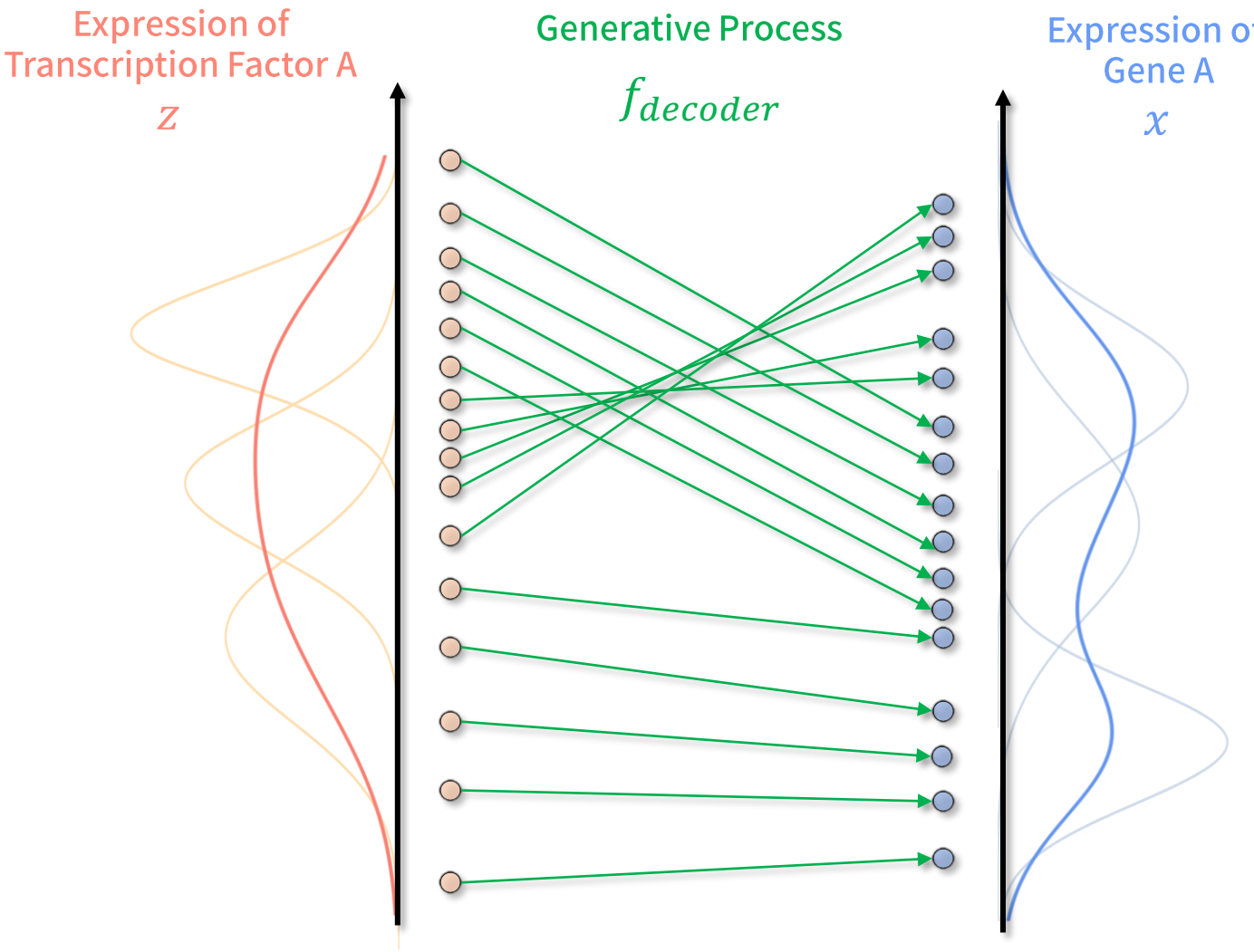
Virtual Universe #02

- Soon people will find out they are controlled by me. Let me make the world more complicated.

- The new law:
  - Depend on the expression of Transcription Factor A, the expression of Gene A will change in a non-linear way (but in groups).
  - There are only 16 cells in this universe, each has only Transcription Factor A and Gene A and you can observe the exact expression of each molecule.
  - And I let you know transcription factors regulate genes.
  - Gene expression follows Gaussian distribution.

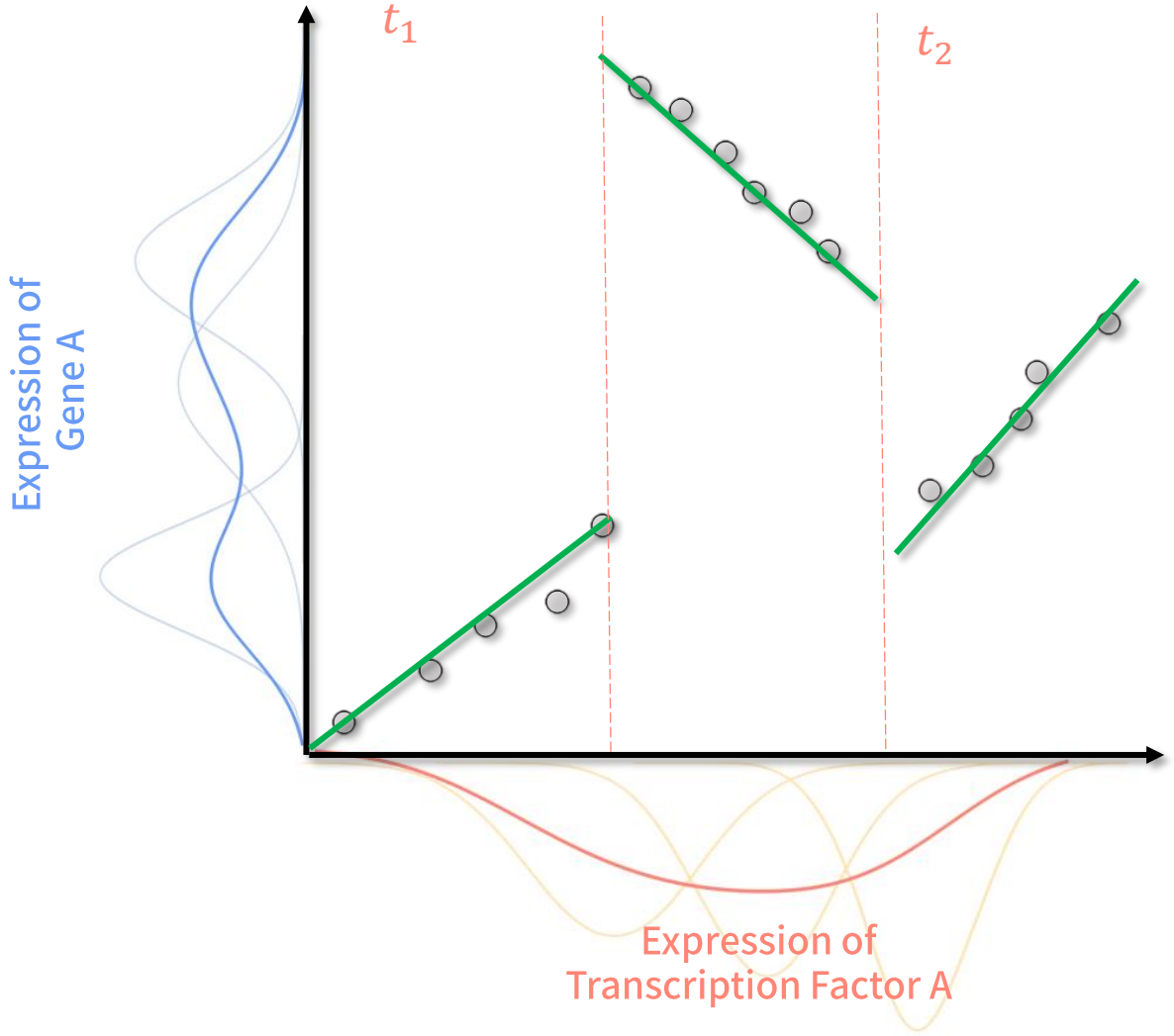
- I guess no one can figure this out.

Virtual Universe #02



Virtual Universe #02

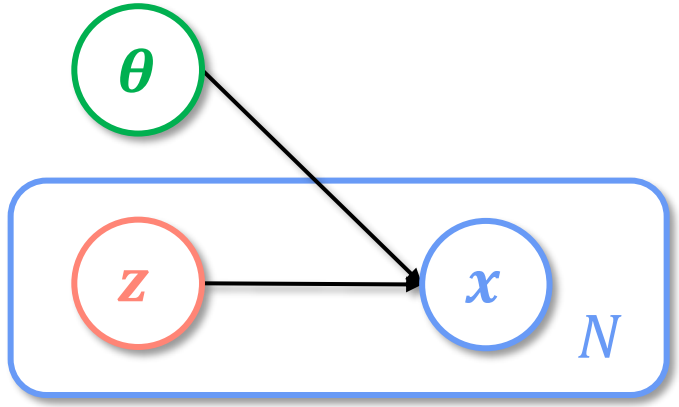
Non-linear Generative Process



$$x = f_{decoder}(z)$$

$$x = 1_{\{z<t_1\}}[zw_1 + b_1] +$$
$$1_{\{t_1 \le z \le t_2\}}[zw_2 + b_2] +$$
$$1_{\{z>t_2\}}[zw_3 + b_3]$$

$$\theta = \{w_1, w_2, w_3, b_1, b_2, b_3\}$$



Virtual Universe #03

- Maybe I should just remove all the researchers in this universe?

Virtual Universe #03

- Maybe I should just remove all the researchers in this universe?

- This time:
  - Depend on the expression of Molecule A, the expression of Gene A will change in a non-linear way (but in groups).
  - There are only 16 cells in this universe, each has only Molecule A and Gene A and you can only observe the exact expression of Gene A in part of these cells.
  - You do not know which molecule regulates Gene A, you can not observe expression of Molecule A neither.
  - Gene expression follows Gaussian distribution.

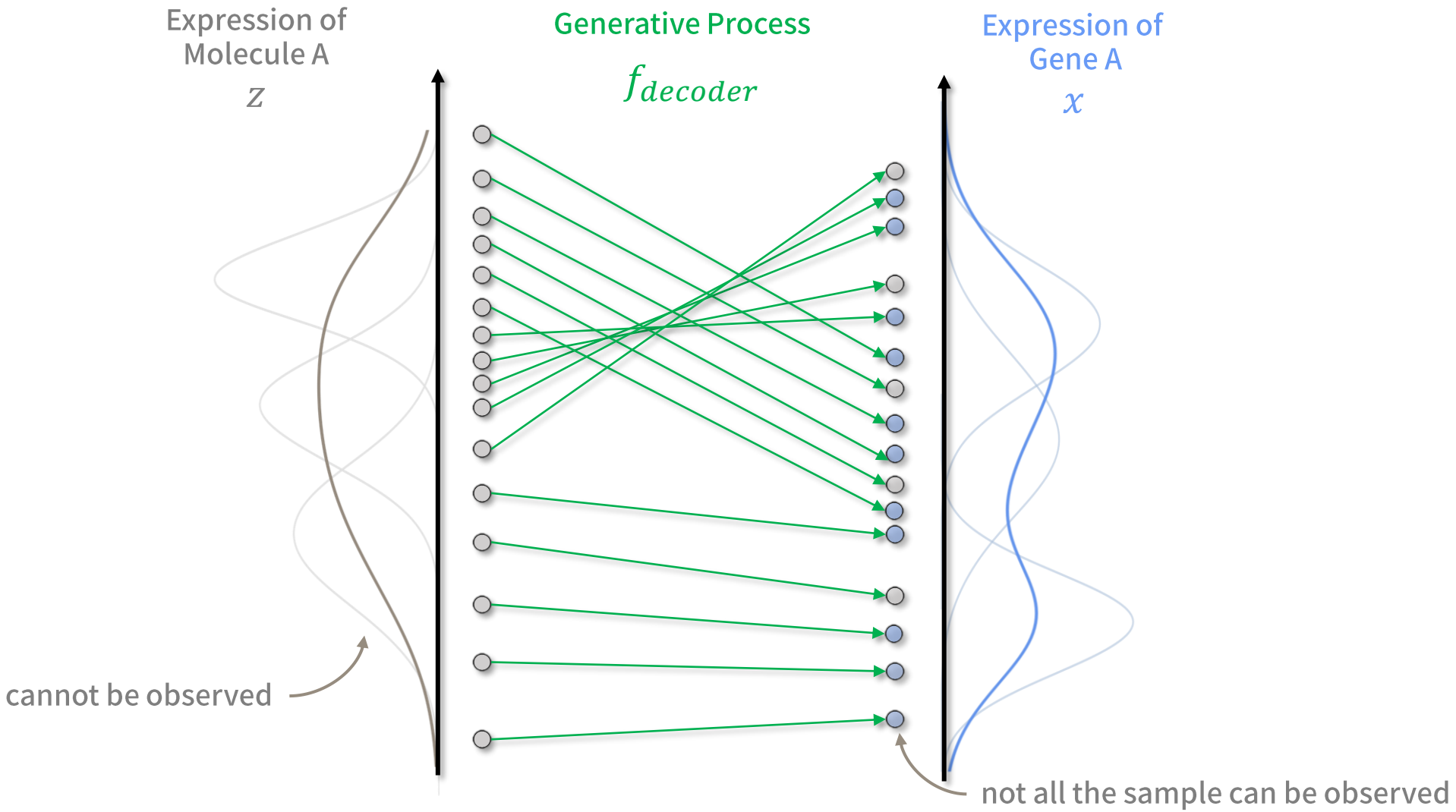
Virtual Universe #03

- Maybe I should just remove all the researchers in this universe?

- This time:
  - Depend on the expression of Molecule A, the expression of Gene A will change in a non-linear way (but in groups).
  - There are only 16 cells in this universe, each has only Molecule A and Gene A and you can only observe the exact expression of Gene A in part of these cells.
  - You do not know which molecule regulates Gene A, you can not observe expression of Molecule A neither.
  - Gene expression follows Gaussian distribution.

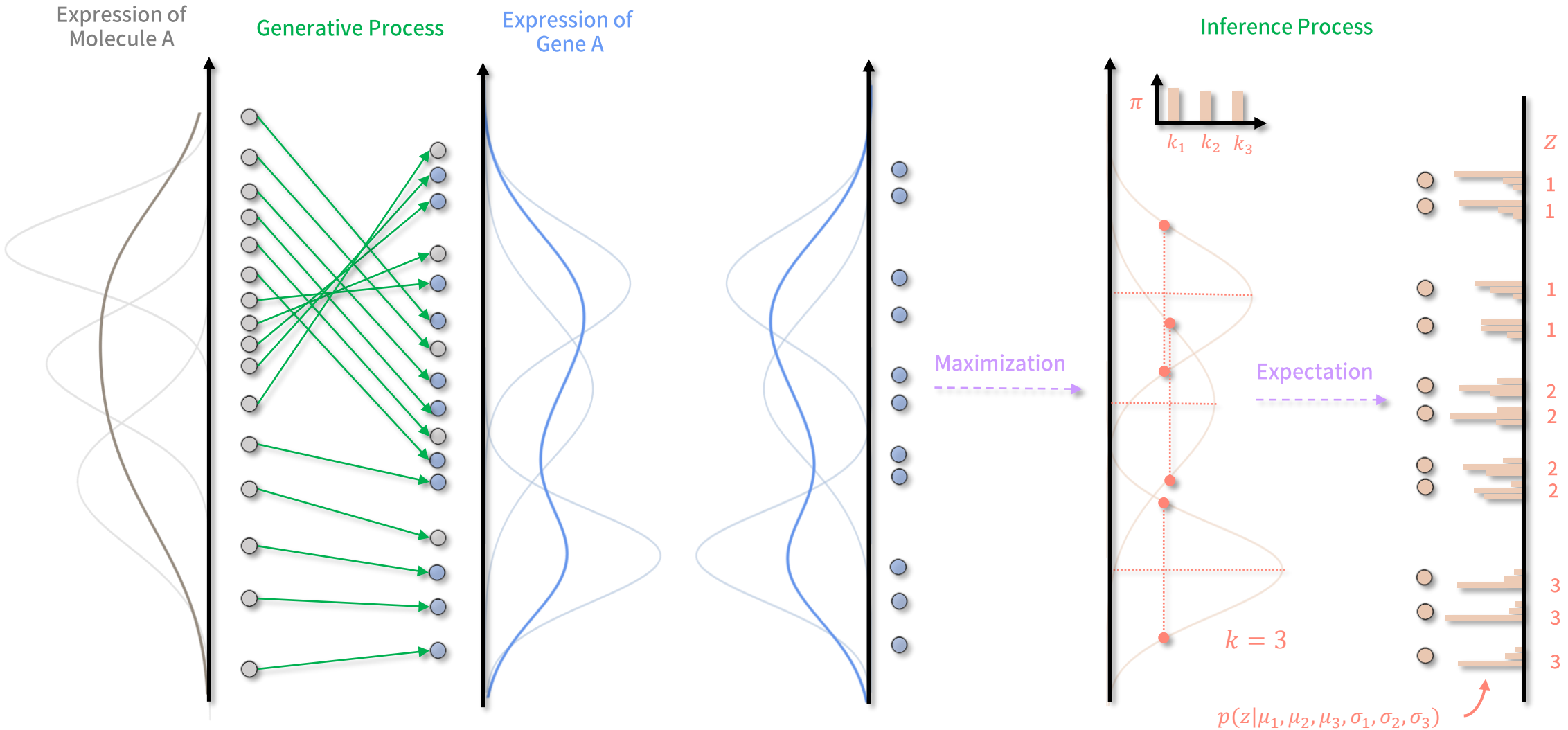
- Perfect!

Virtual Universe #03



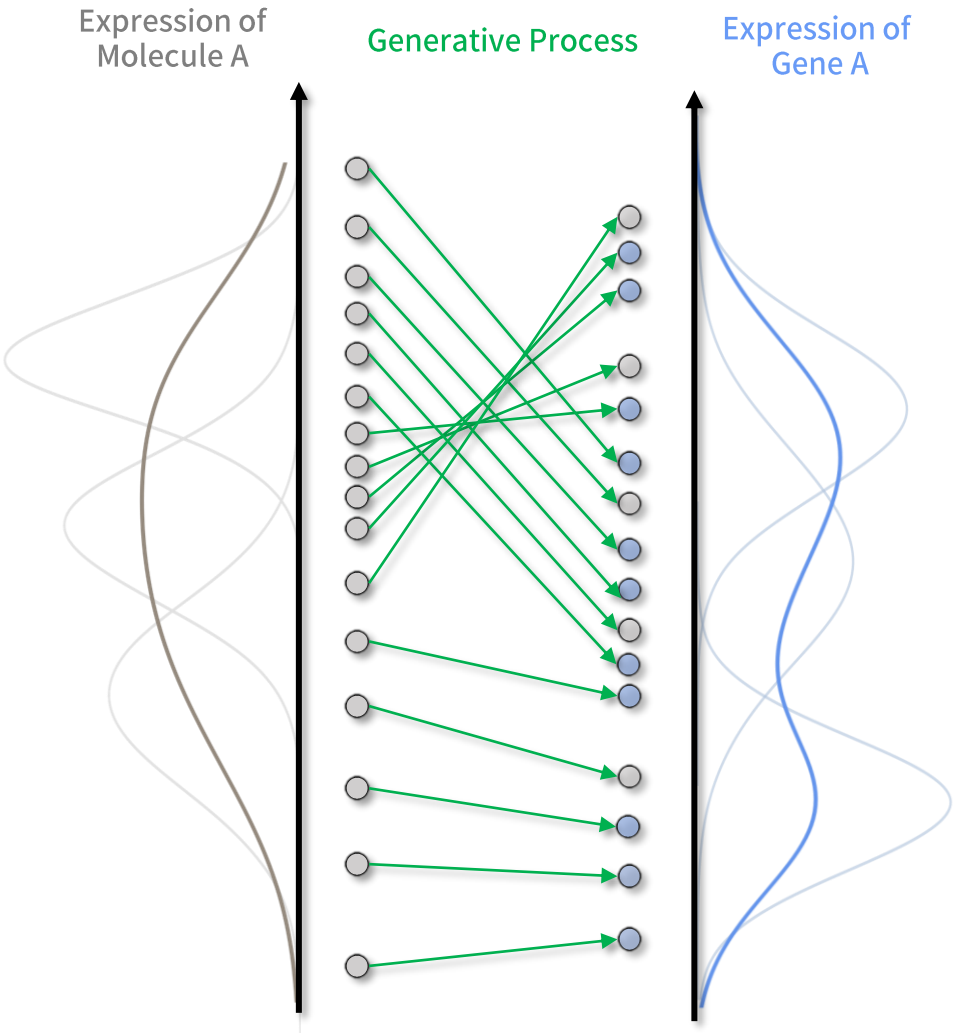
Virtual Universe #03

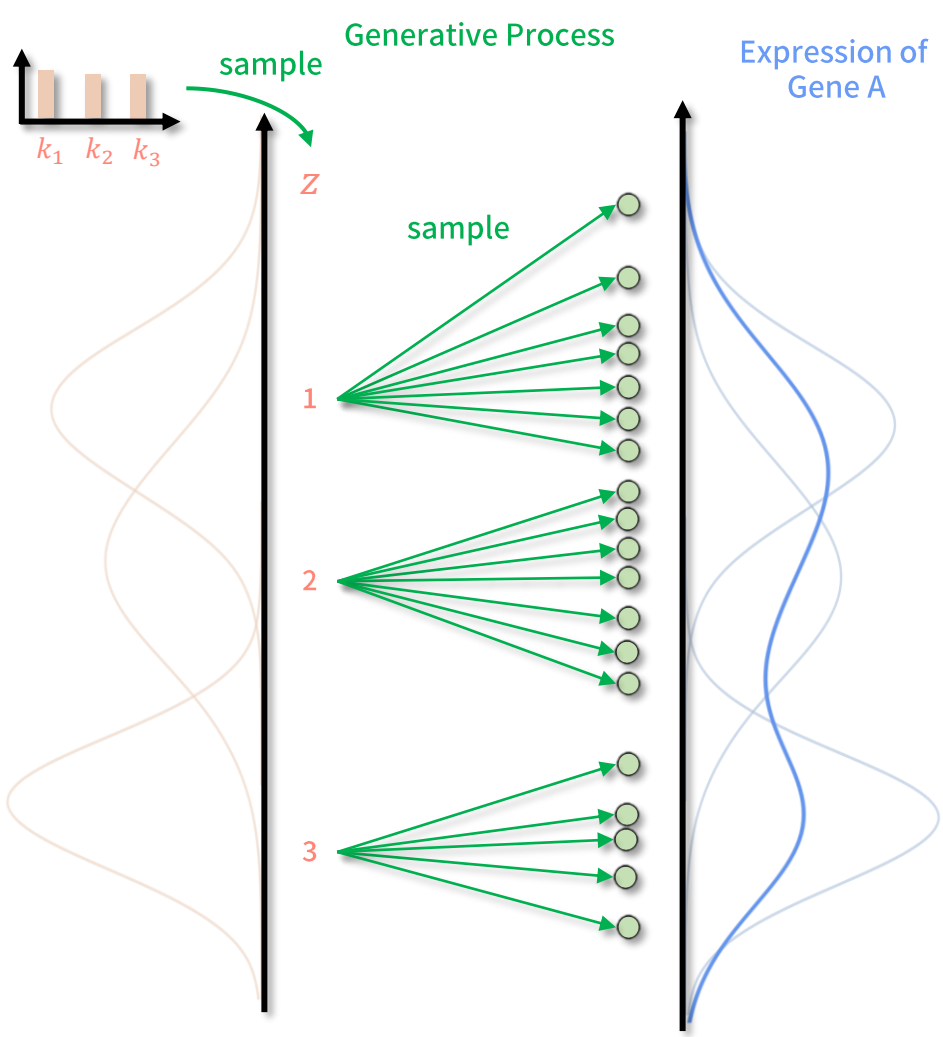
Gaussian Mixture Model (1)



Virtual Universe #03

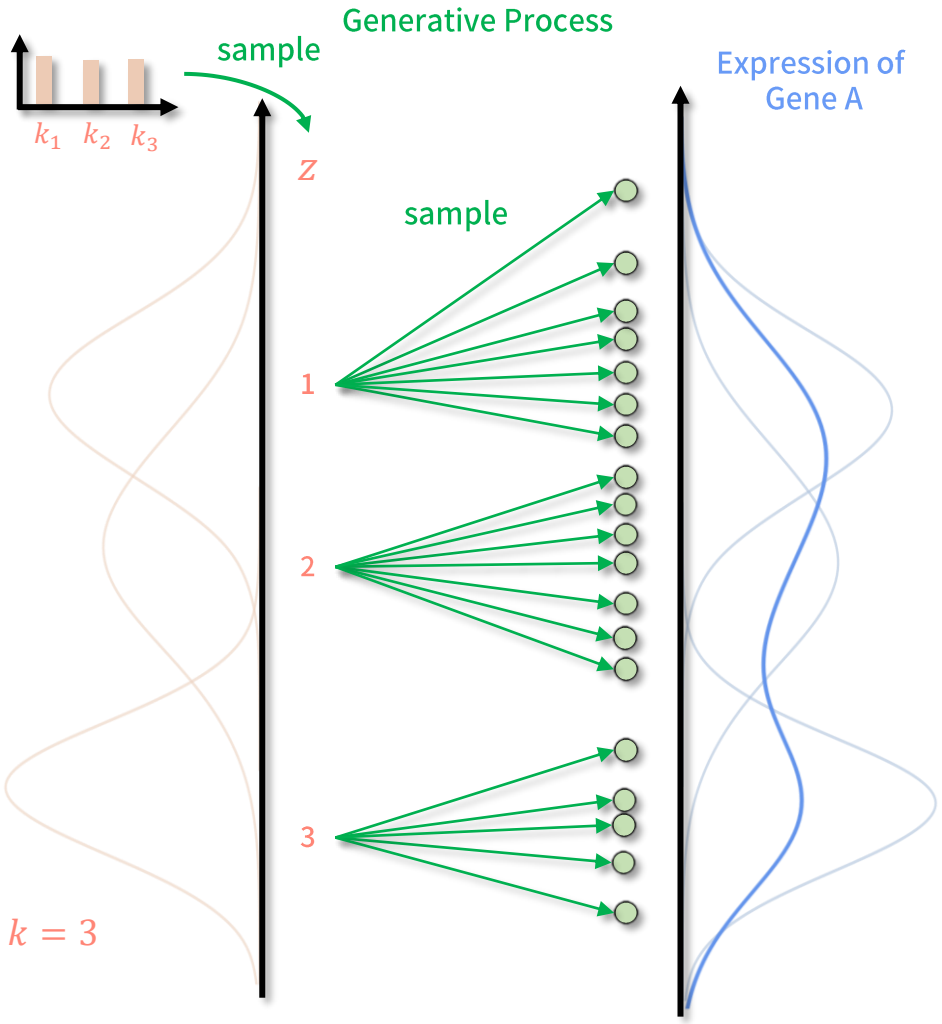
Gaussian Mixture Model (2)





Virtual Universe #03

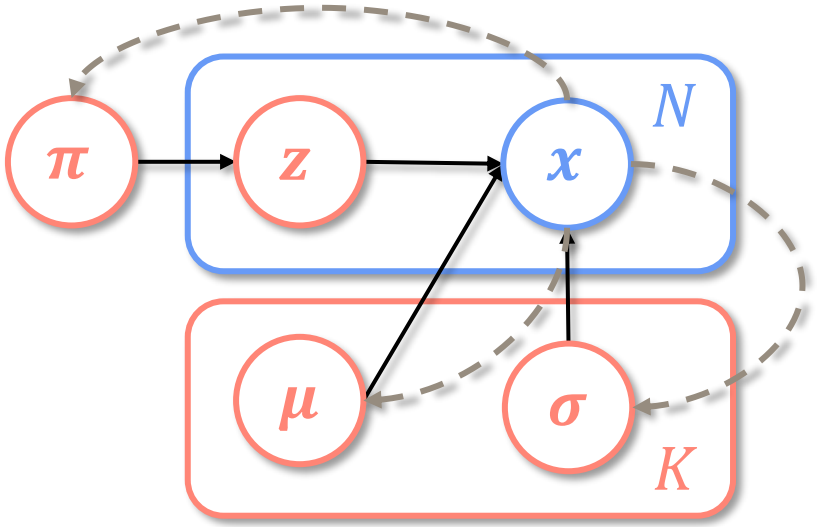
Gaussian Mixture Model (3)



$$x = f_{decoder}(z)$$

$$z \sim Cat(\pi)$$

$$x \sim N(\mu_z, \sigma_z)$$



Virtual Universe #03

Gaussian Mixture Model (4)

- Assumptions:
  - There are different clusters in the samples.
  - Samples from the same cluster undergo similar generative process.

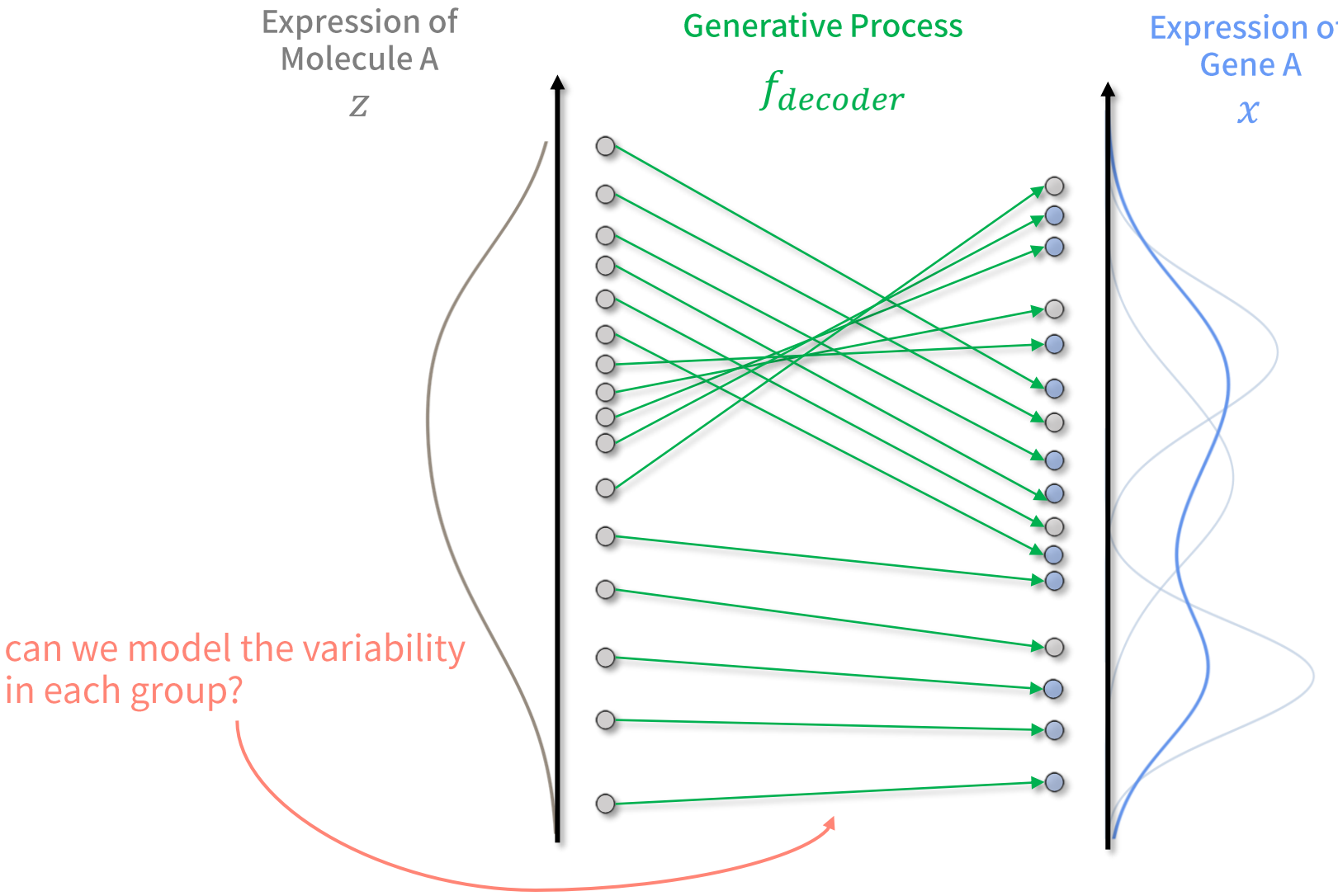
- Limitations:
  - Gaussian mixture model can be used to infer the generative process for a group of samples, but not for each individual sample.
  - Additional analysis for each clusters need to be done to reconstruct the generative process.
  - Not effective on high-dimensional data (curse of dimensionality).
  - Latent variable is a discrete variable.
  - Data distribution needs to be a mixture of Gaussian distribution.

Virtual Universe #04

- Finally:
  - Depend on the expression of Molecule A, the expression of Gene A will change in a non-linear way (even the variability within each group is generated using a specific rule).
  - There are only 16 cells in this universe, each has only Molecule A and Gene A and you can only observe the expression of Gene A in part of these cells.
  - You do not know which molecule regulates Gene A, you can not observe expression of Molecule A neither.
  - The observed expression of Gene A is uncertain.
  - Gene expression follows Negative Binomial distribution.

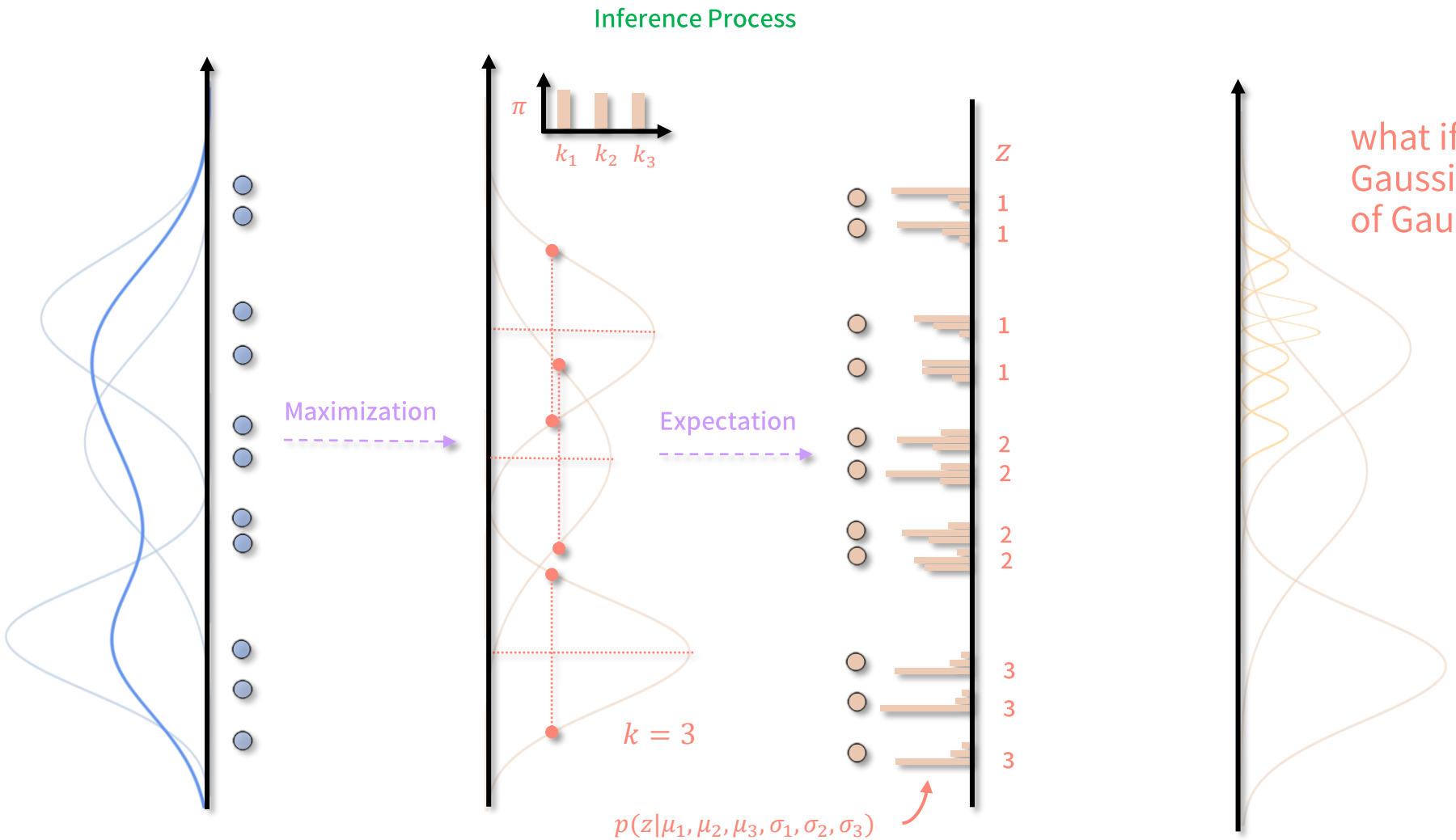
Virtual Universe #04

Improve Gaussian Mixture Model (1)



Virtual Universe #04

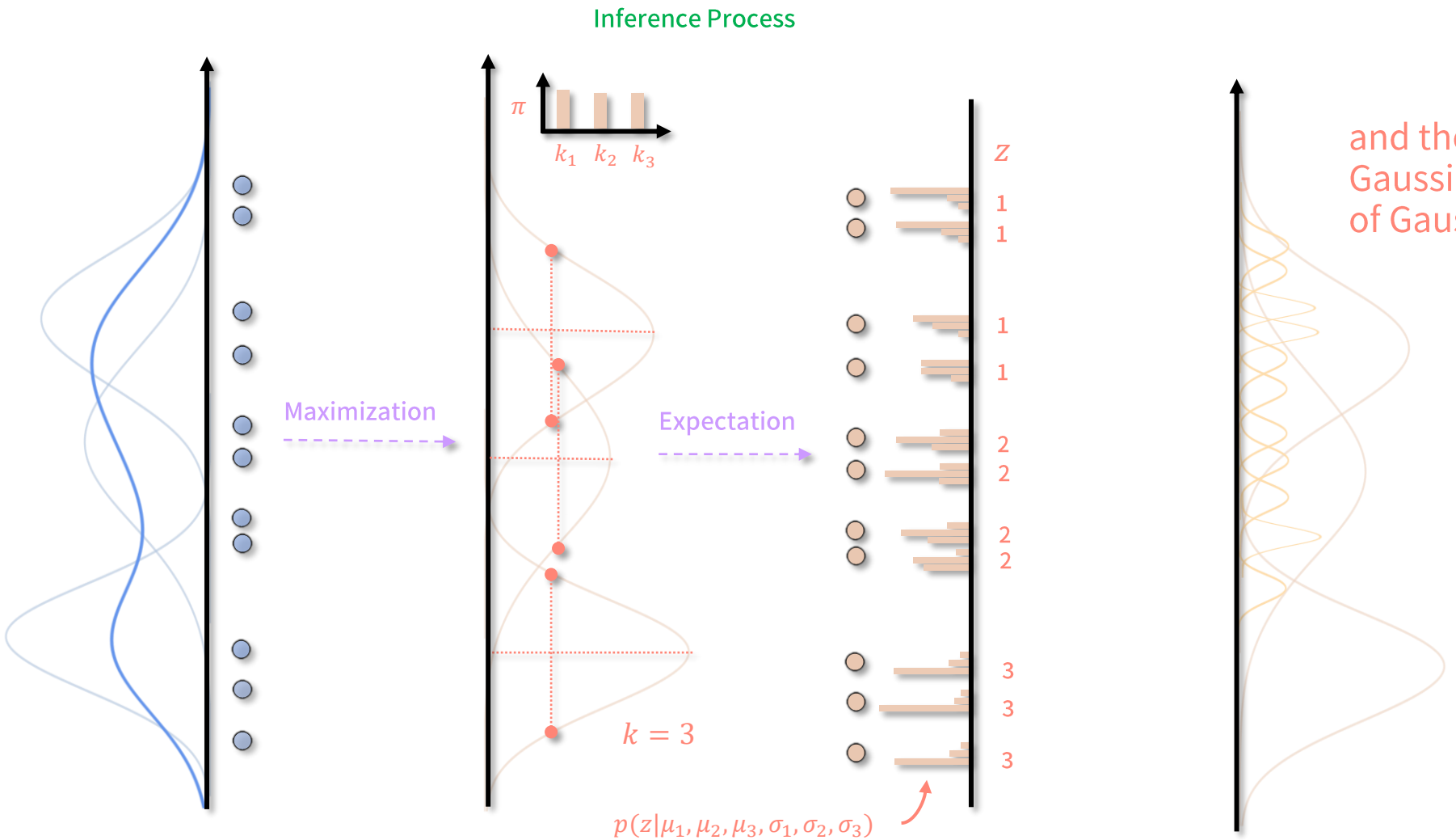
Improve Gaussian Mixture Model (2)



what if we partition the first Gaussian into another mixture of Gaussian?

Virtual Universe #04

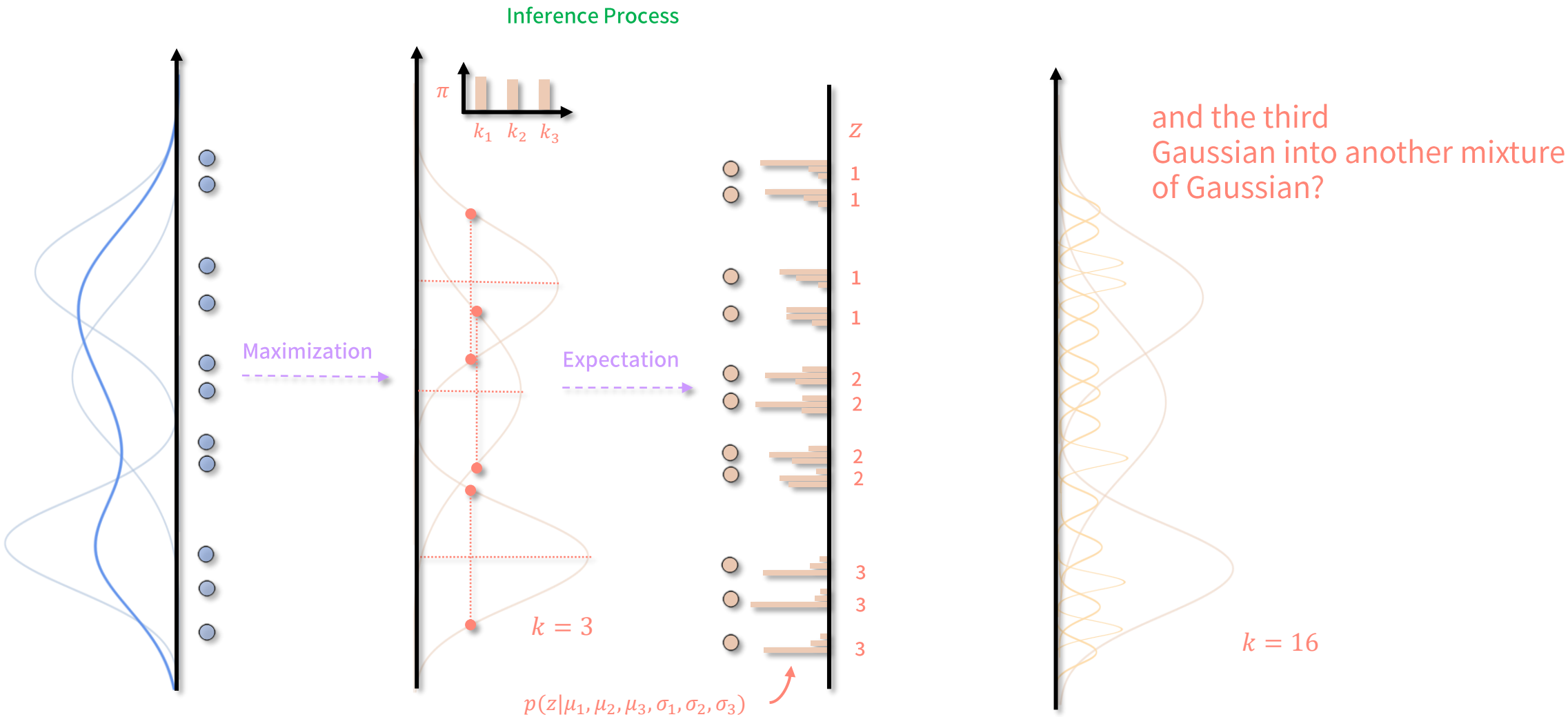
Improve Gaussian Mixture Model (3)



and the second
Gaussian into another mixture
of Gaussian?

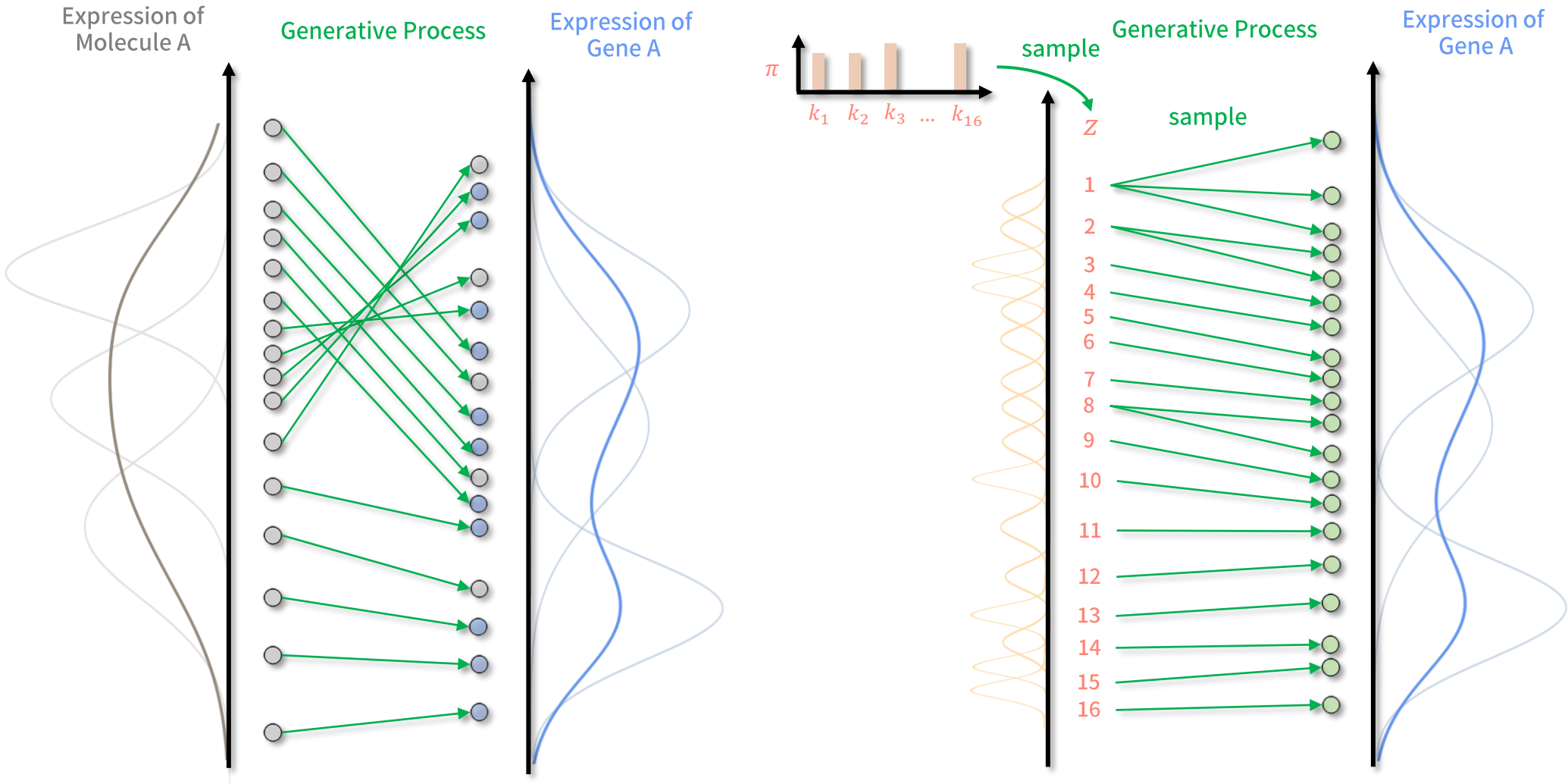
Virtual Universe #04

Improve Gaussian Mixture Model (4)



Virtual Universe #04

Improve Gaussian Mixture Model (5)



Virtual Universe #04

Improve Gaussian Mixture Model (5)

- The expectation-maximization becomes unstable when the number of component increases.

- How to address this issue?
  - Assume $K = N$.
  - For each point we observe, train a function and map them into a Gaussian distribution.
  - Since it is a function, now we can map any input (even if they are not in this universe) to a Gaussian distribution.

$$\mu = \frac{1}{N}(x_1 + x_2 + \cdots + x_N)$$

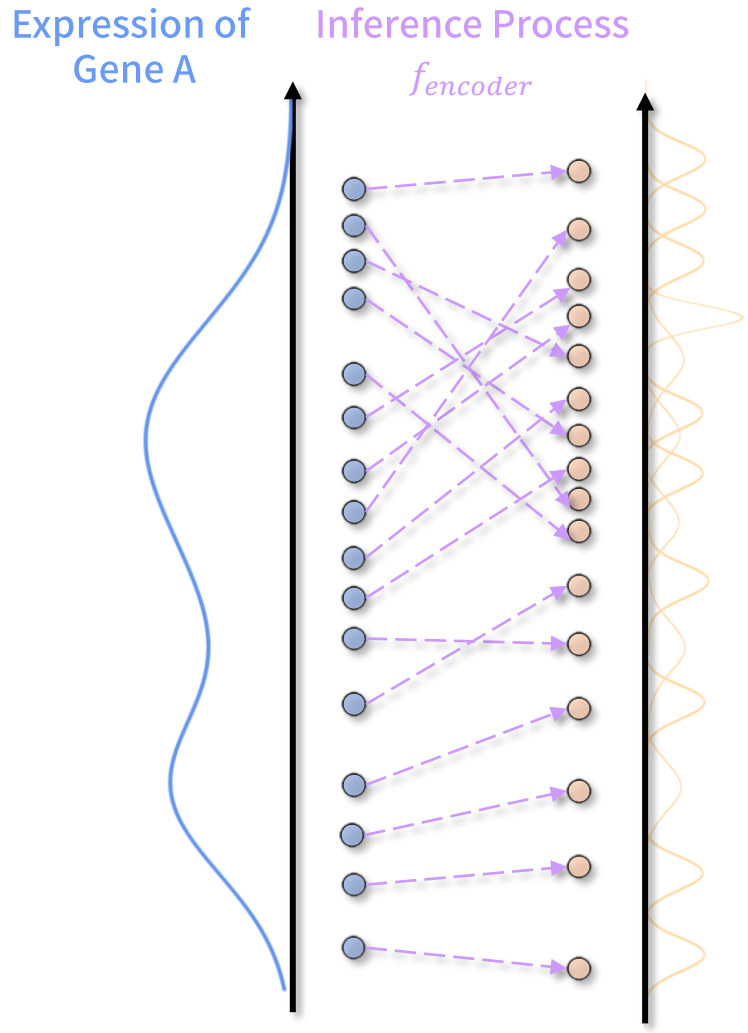
$$\sigma = \sqrt{\frac{1}{N}[(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2]}$$

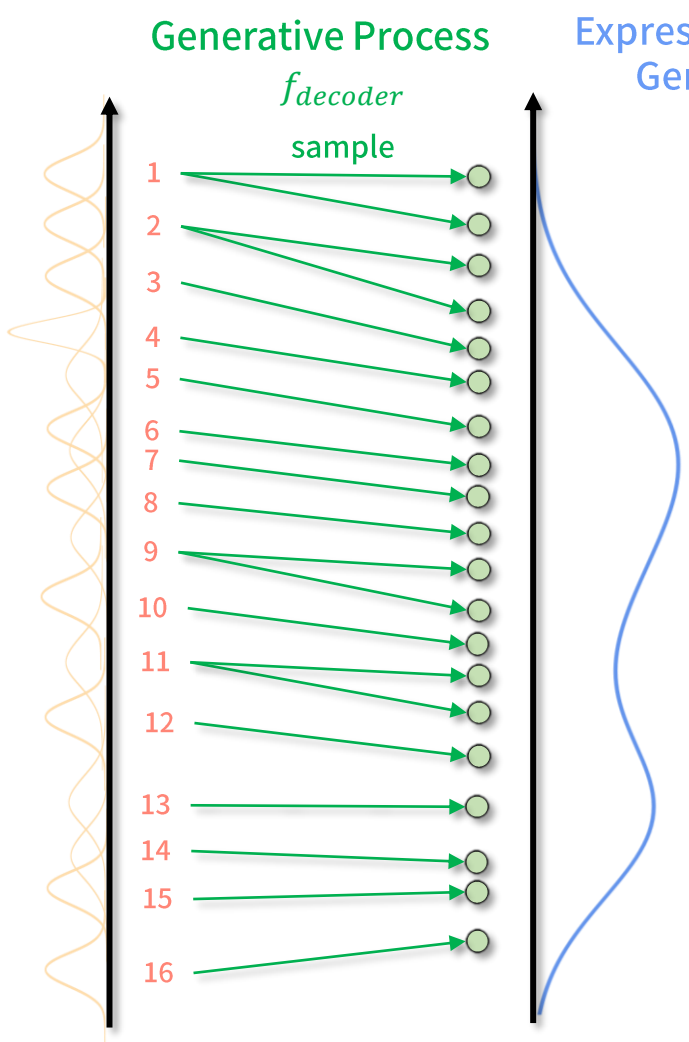
$$\longrightarrow \quad \mu = f_{encoder,\mu}(x)$$

$$\sigma = f_{encoder,\sigma}(x)$$

Virtual Universe #04

Improve Gaussian Mixture Model (6)





Virtual Universe #04

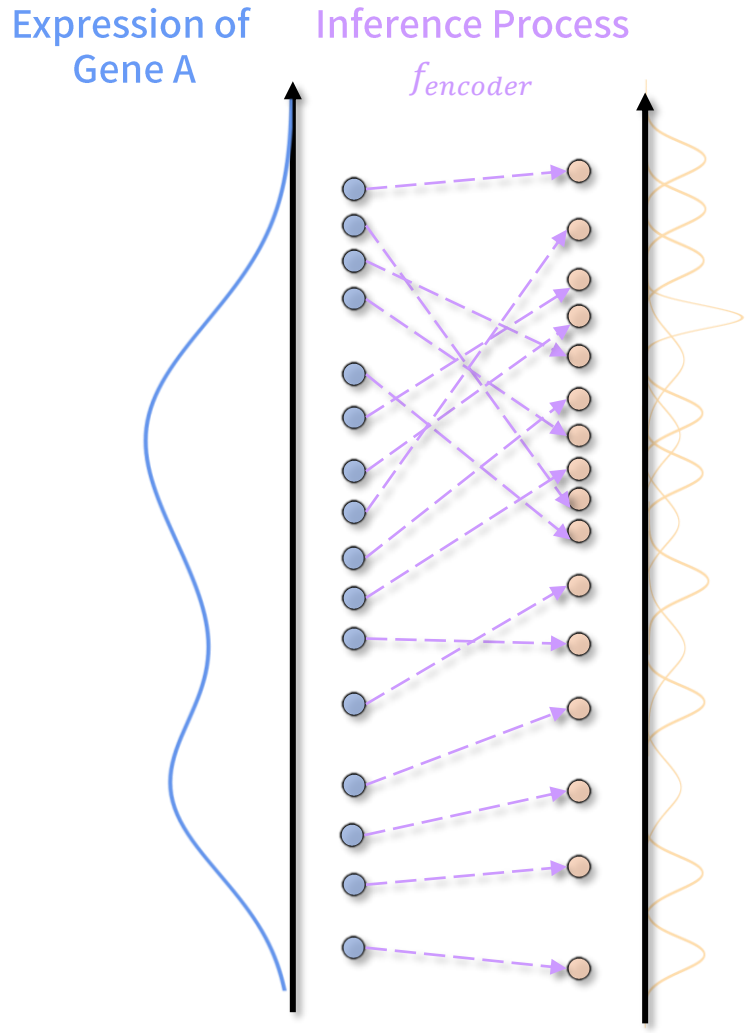
Improve Gaussian Mixture Model (7)

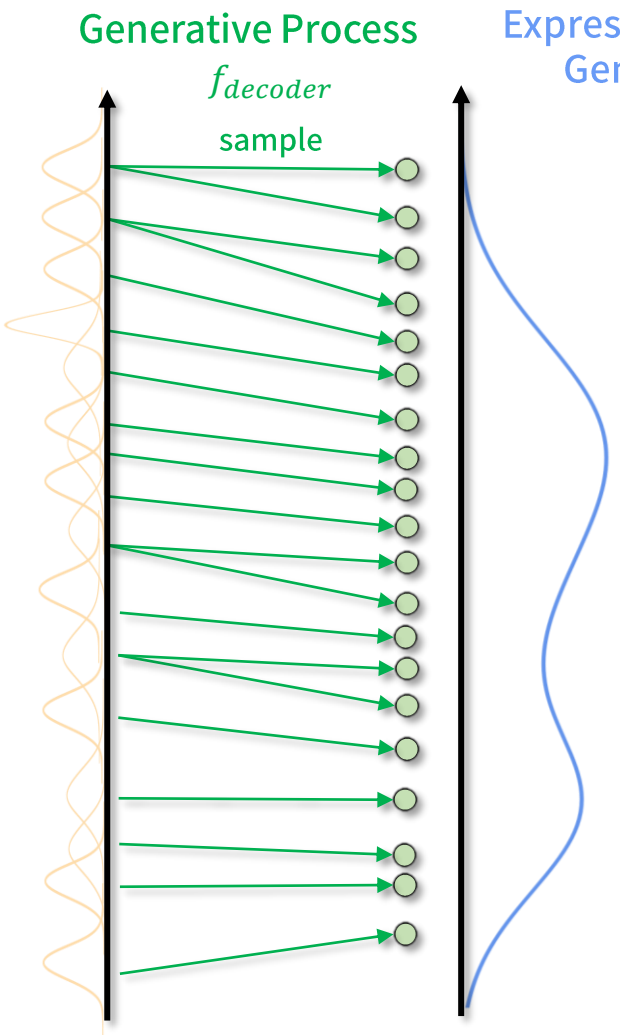
- Assumptions:
  - There are different clusters in the samples.
  - Samples from the same cluster undergo similar generative process.

- Limitations:
  - Gaussian mixture model can be used to infer the generative process for a group of samples, but not for each individual sample.
  - Additional analysis for each clusters need to be done to reconstruct the generative process.
  - Not effective on high-dimensional data (curse of dimensionality).
  - Latent variable is a discrete variable.
  - Data distribution needs to be a mixture of Gaussian distribution.

Virtual Universe #04

Improve Gaussian Mixture Model (6)





Virtual Universe #04

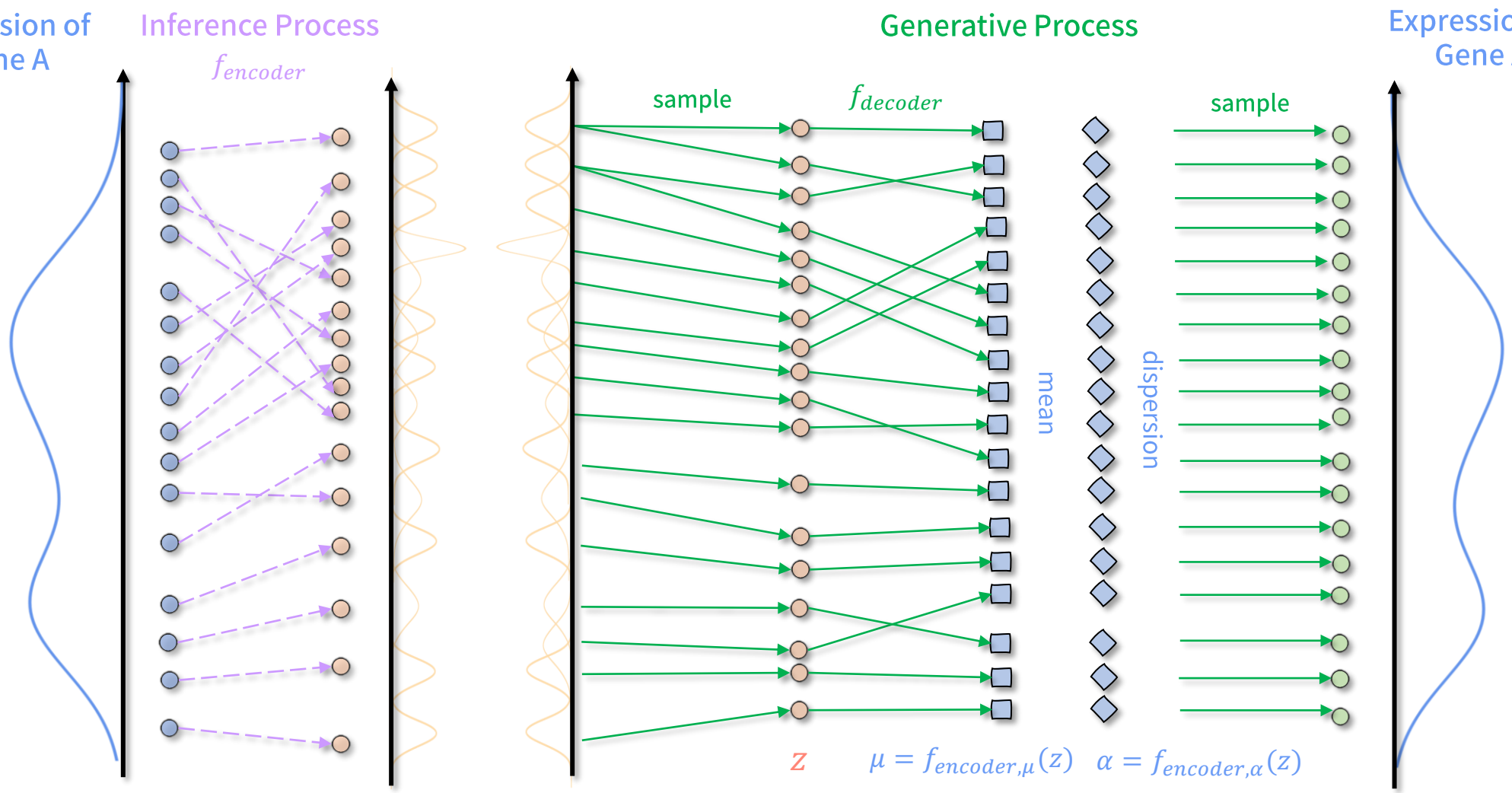
Improve Gaussian Mixture Model (7)

- Assumptions:
  - There are different clusters in the samples.
  - Samples from the same cluster undergo similar generative process.

- Limitations:
  - Gaussian mixture model can be used to infer the generative process for a group of samples, but not for each individual sample.
  - Additional analysis for each clusters need to be done to reconstruct the generative process.
  - Not effective on high-dimensional data (curse of dimensionality).
  - Latent variable is a discrete variable.
  - Data distribution needs to be a mixture of Gaussian distribution.

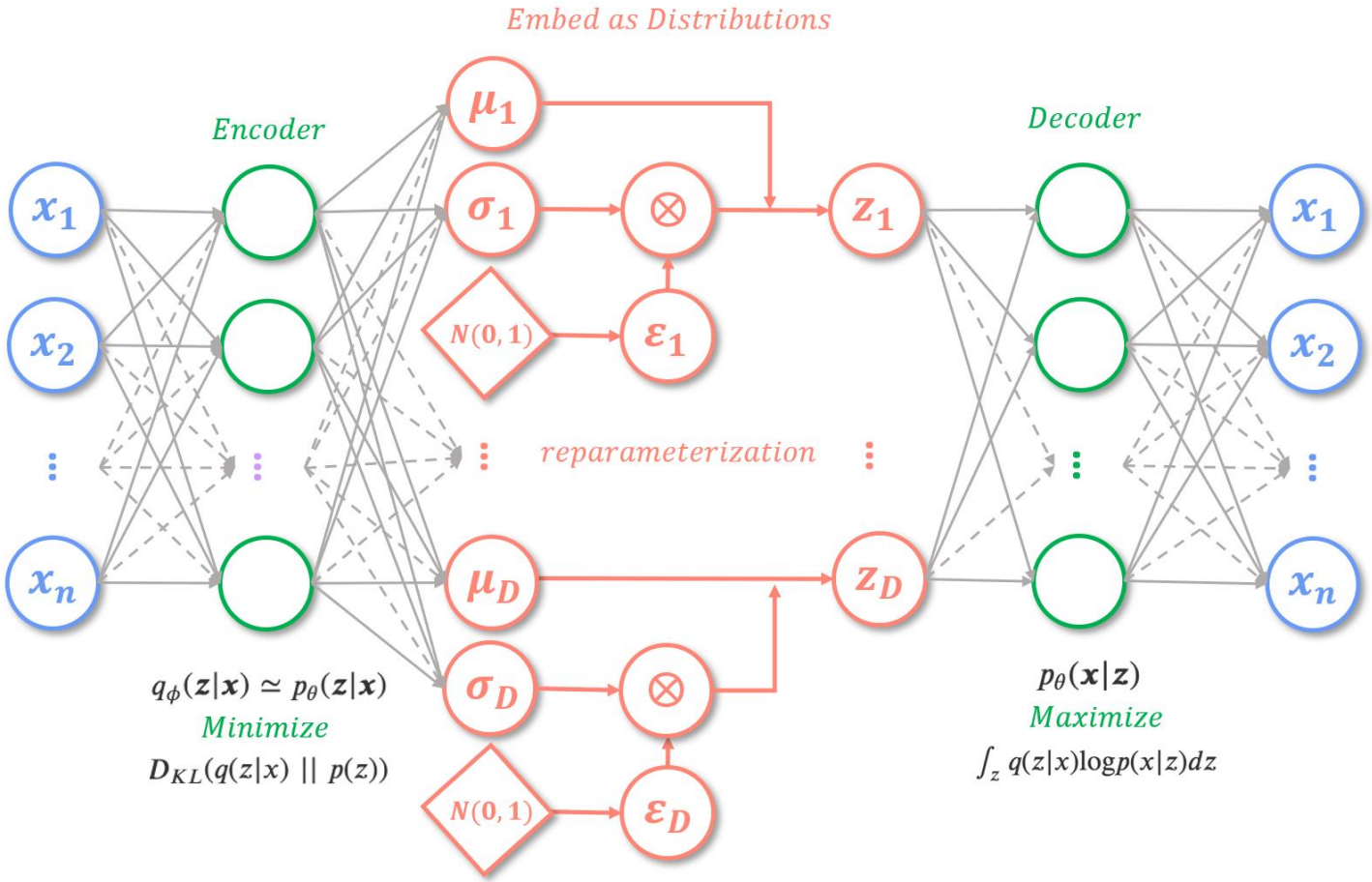
Virtual Universe #04

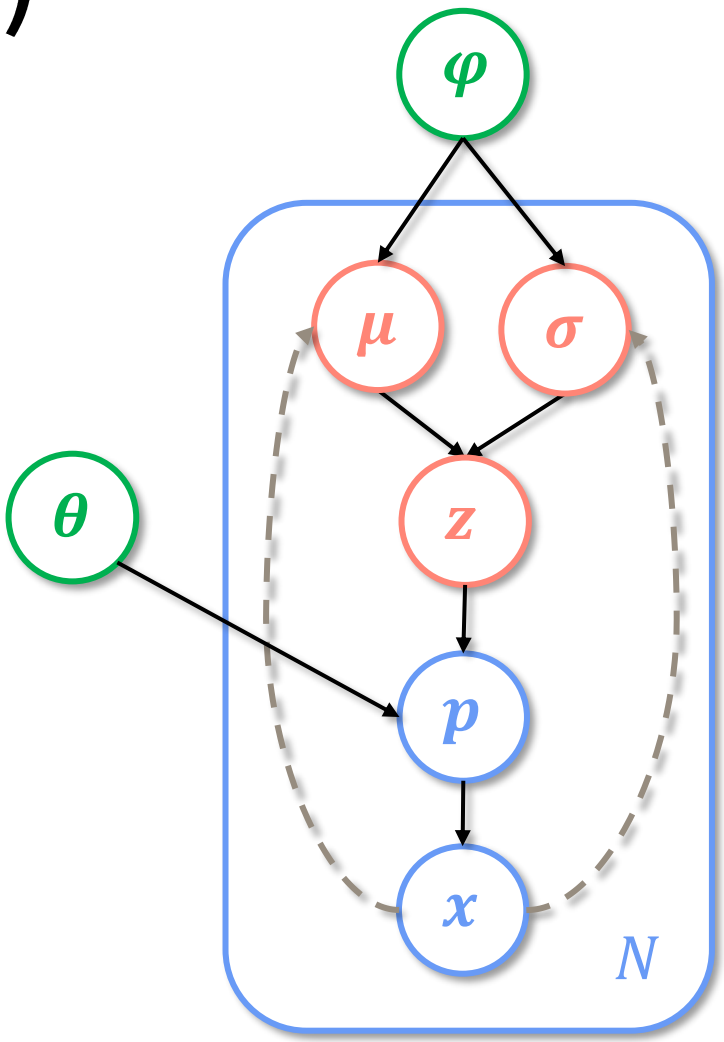
Variational Autoencoder (1)



Virtual Universe #04

Variational Autoencoder (2)





Conclusion

- Generative Models
  - Allow us to reconstruct the generative process of the system we observe.
  - Allow us to identify the potential causal factors that generative the data we observe.
  - Allow us to understand the cluster of samples
  - Allow us to generative new data based on the system we observe.

- Variational Autoencoder
  - A flexible framework that allow us to model the generative process of the system we observe if the data distribution is continuous (and differentiable).

Thanks