



Фонд
интеллект

Анализ транскриптомных данных

Лекция 11

Использование вариационных автоэнкодеров для процессинга scRNA-seq. scVI-tools

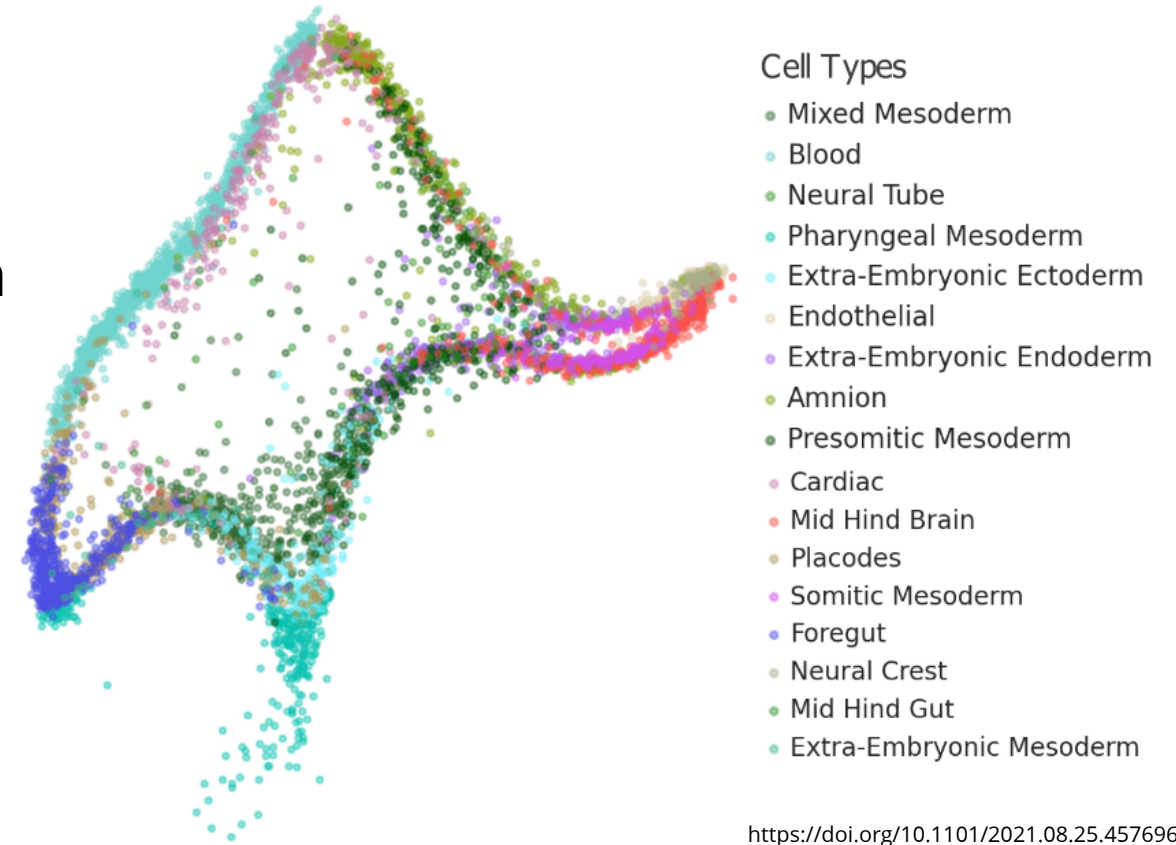
Даниил Бобровский

магистрант EPFL

сотрудник лаборатории системной
биологии нейроразвития, EPFL

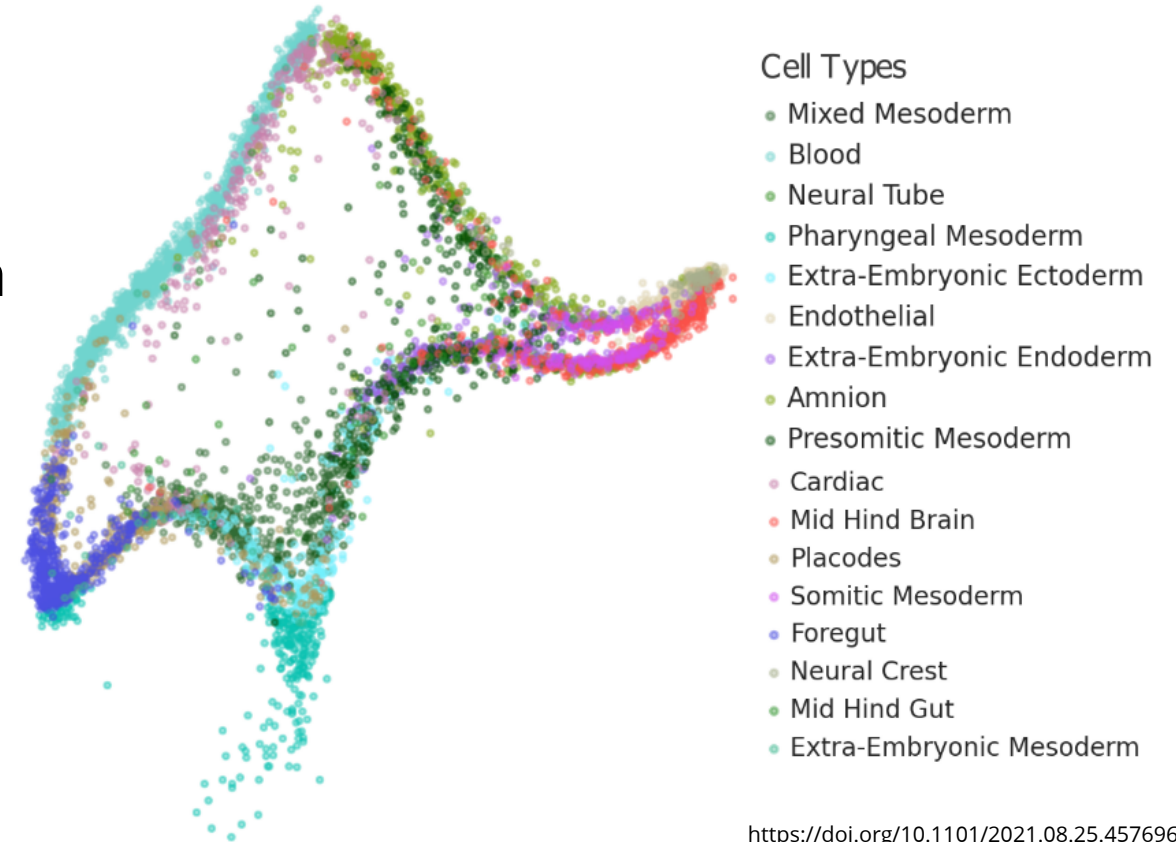
Методы снижения размерности

- PCA, ICA
- t-SNE, UMAP
- Force-directed graphs (ForceAtlas2)
- Hierarchical Poisson matrix factorization
- Diffusion maps



Методы снижения размерности

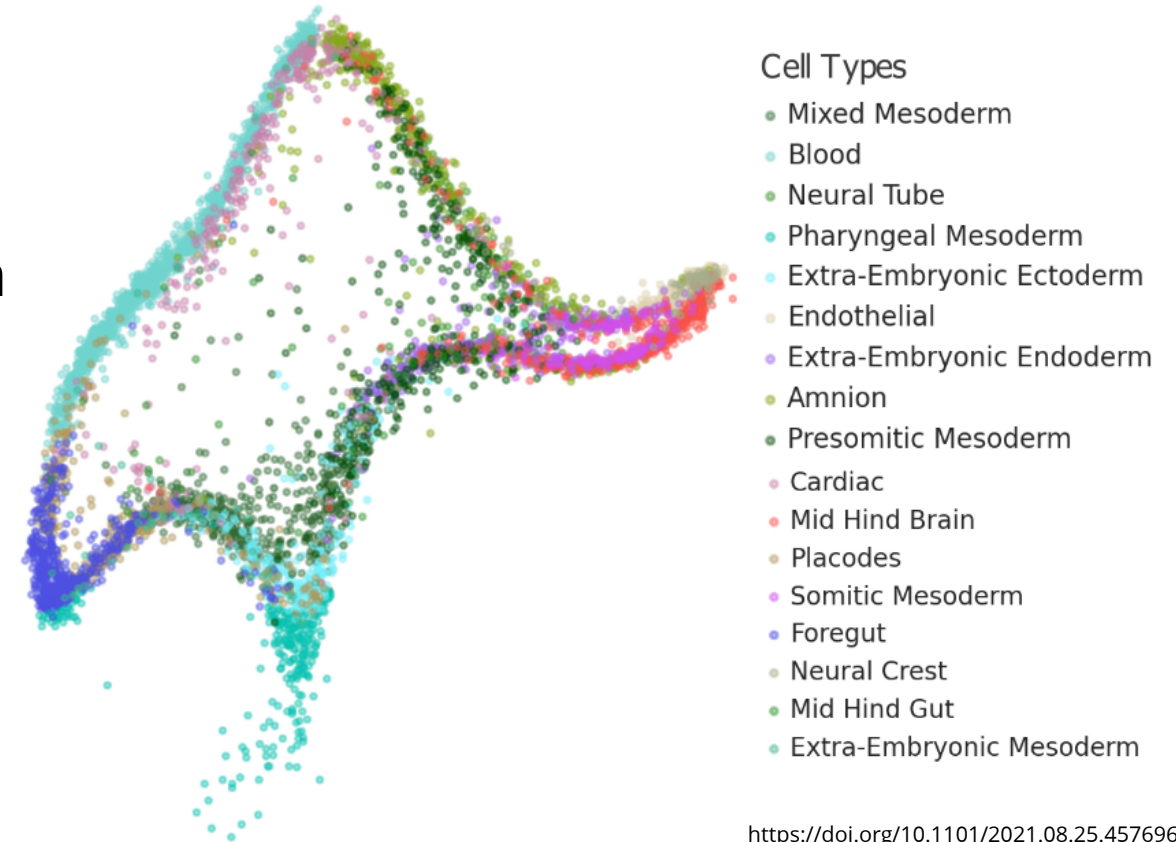
- PCA, ICA
- t-SNE, UMAP
- Force-directed graphs (ForceAtlas2)
- Hierarchical Poisson matrix factorization
- Diffusion maps



Какие с этим бывают проблемы?

Методы снижения размерности

- PCA, ICA
- t-SNE, UMAP
- Force-directed graphs (ForceAtlas2)
- Hierarchical Poisson matrix factorization
- Diffusion maps



Какие с этим бывают проблемы?

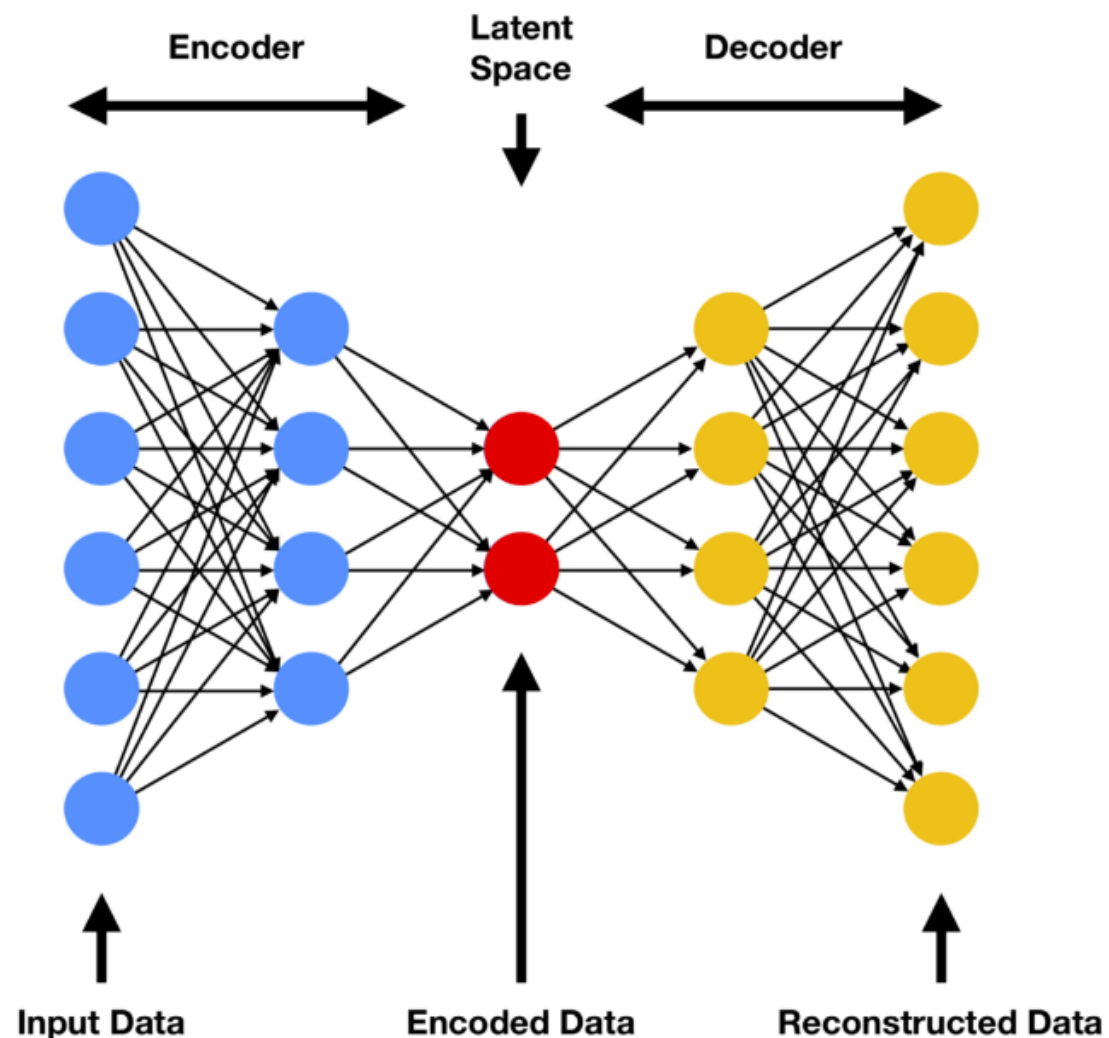
Недостаточно хорошо разделяются клетки, результат зависит от выбранных параметров, ...

Автоэнкодеры

Нейронные сети, обладающие особой архитектурой:

- Пространство входных данных и латентное пространство (скрытое представление)
- Два "сегмента": энкодер и декодер

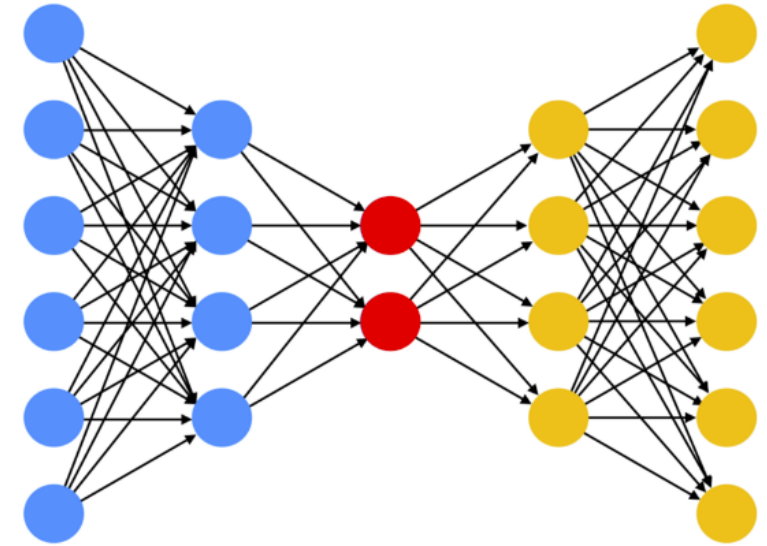
Энкодер снижает размерность входных данных, декодер восстанавливает до исходной (**undercomplete autoencoder**)



Обучение автоэнкодеров

$$MSE = \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2$$

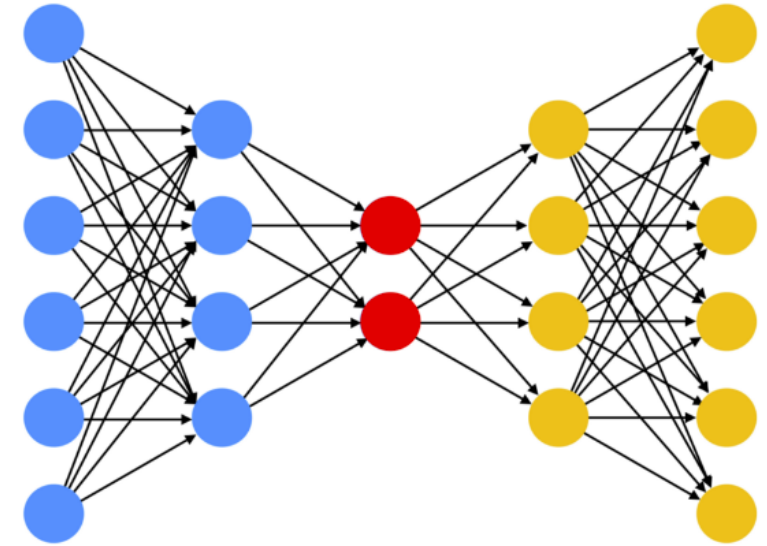
- В качестве loss function для обычного автоэнкодера можно использовать среднеквадратичную ошибку



Обучение автоэнкодеров

$$MSE = \frac{1}{N} \sum_{i=1}^N ||x_i - \hat{x}_i||^2$$

- В качестве loss function для обычного автоэнкодера можно использовать среднеквадратичную ошибку
- В результате мы получаем такое скрытое представление данных, которое сохраняет больше всего информации об объектах, убирая шум
- Автоэнкодеры также можно использовать как генеративные модели



Представим автоэнкодер, декодер которого - линейная нейросеть, а loss function - MSE

Какое латентное пространство он выучит?

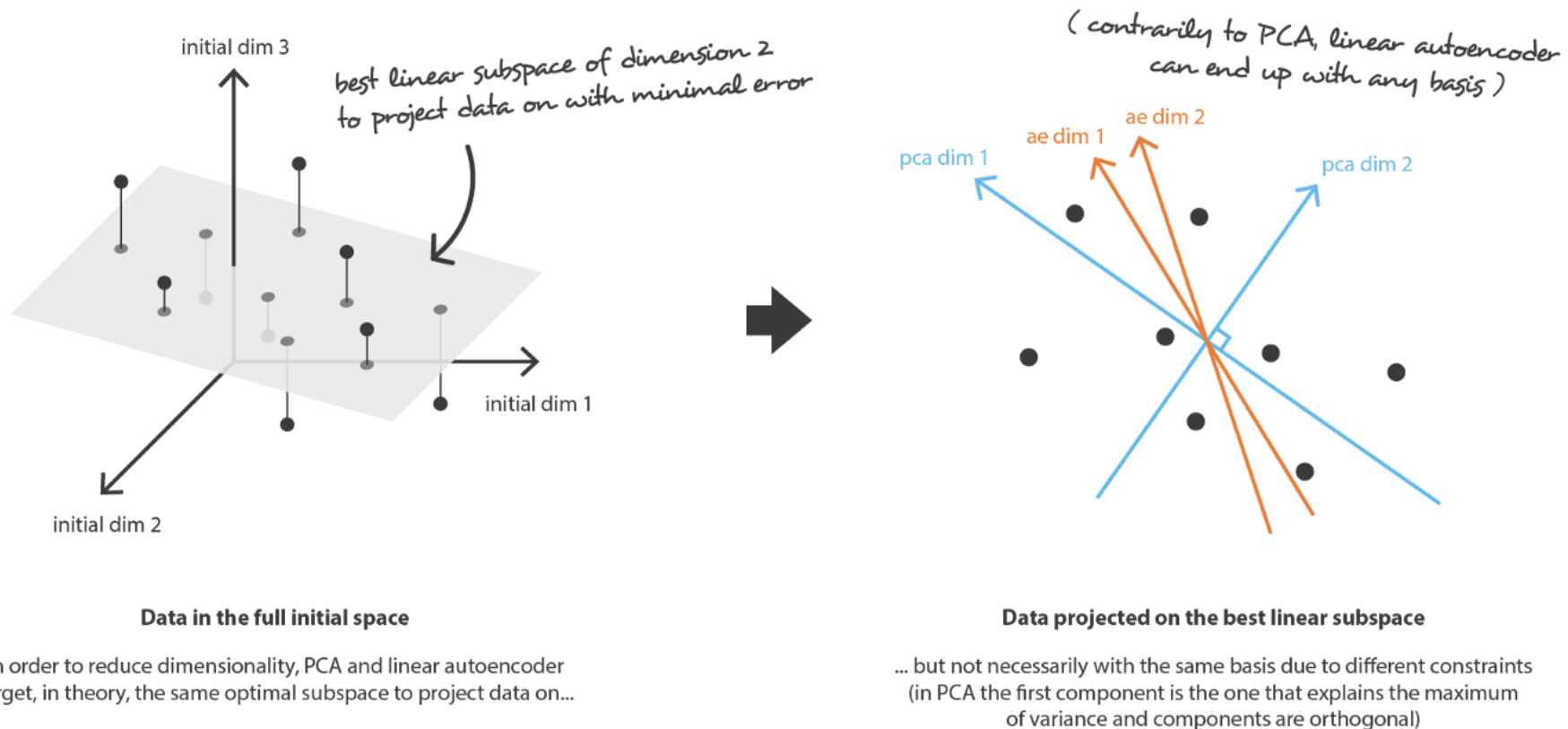
Представим автоэнкодер, декодер которого - линейная нейросеть, а loss function - MSE

Какое латентное пространство он выучит?

Подпространство главных компонент!

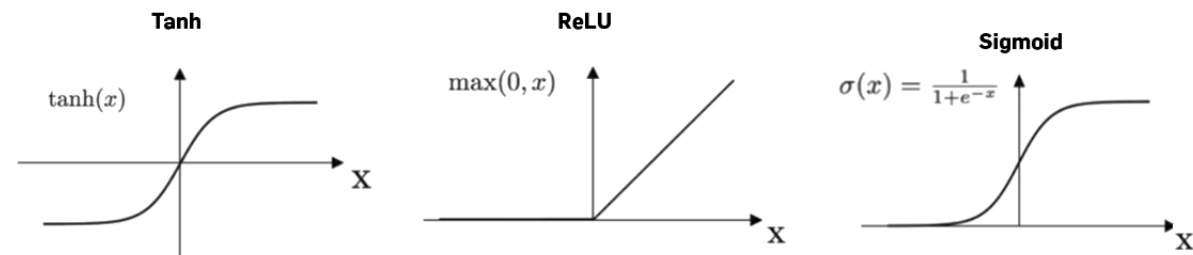
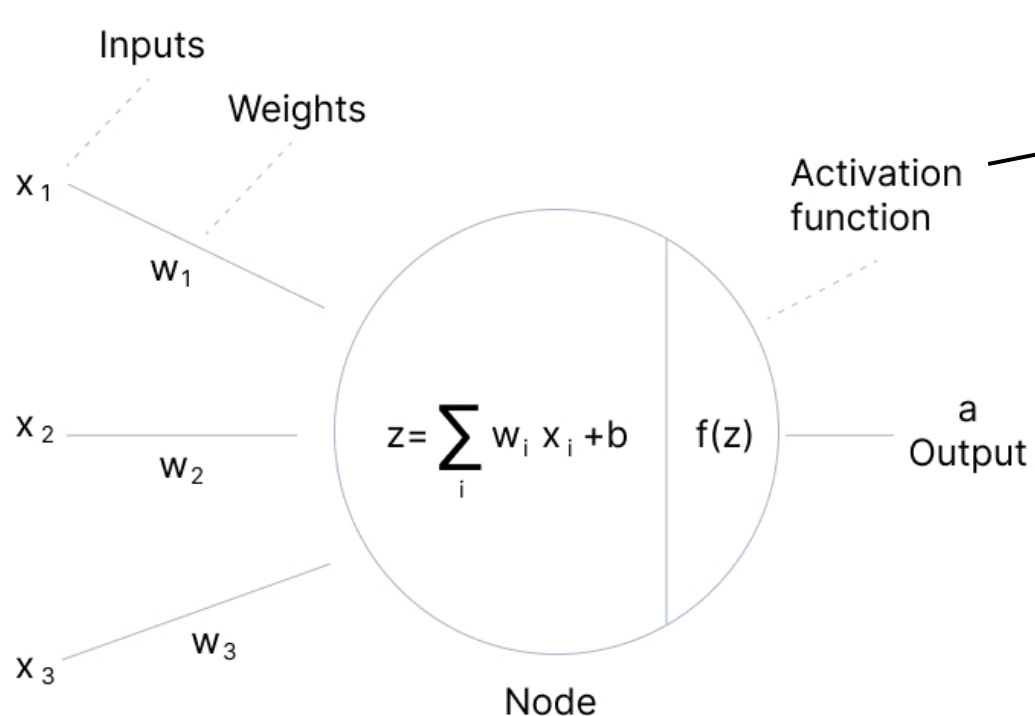
PCA vs linear autoencoder

Впрочем, в случае автоэнкодера базис не обязательно будет ортогональным



Нелинейные автоэнкодеры

Нам же автоэнкодеры интересны как раз тем, что могут выучивать нелинейные комбинации признаков



Самым стандартным выбором для энкодера и декодера будут нейронные сети с ReLU в качестве функции активации

Свойства латентного пространства

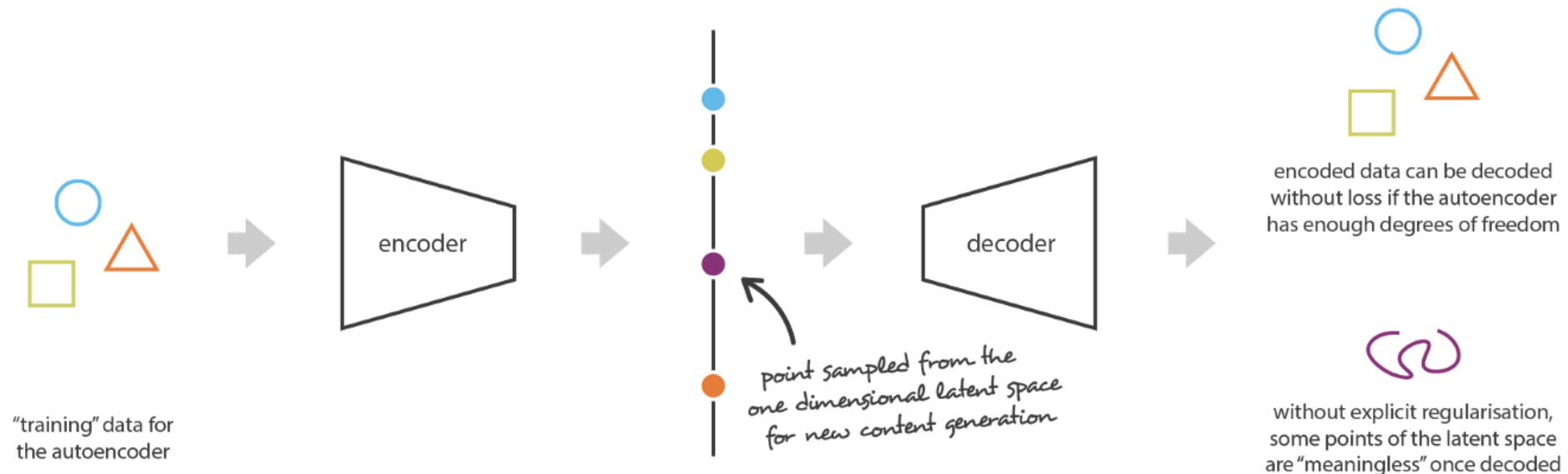
Проблема: такая архитектура ничего не гарантирует нам с точки зрения организации латентного пространства

Свойства латентного пространства

Проблема: такая архитектура ничего не гарантирует нам с точки зрения организации латентного пространства

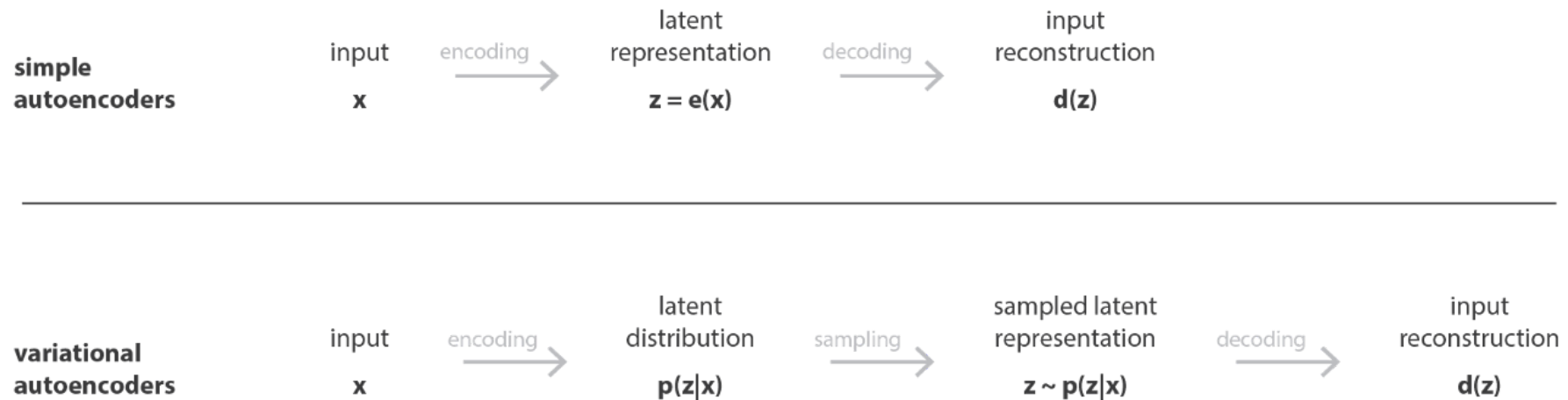
Пусть наша модель смогла закодировать все объекты в латентном пространстве размерности 1

Чему соответствуют области между объектами?



Вариационные автоэнкодеры

Нам бы хотелось, чтобы латентное пространство Z было непрерывным и описывало распределение признаков наших объектов из пространства X . Так давайте и будем кодировать входные данные не как точки в этом пространстве, а как распределение $p(z|x)$.



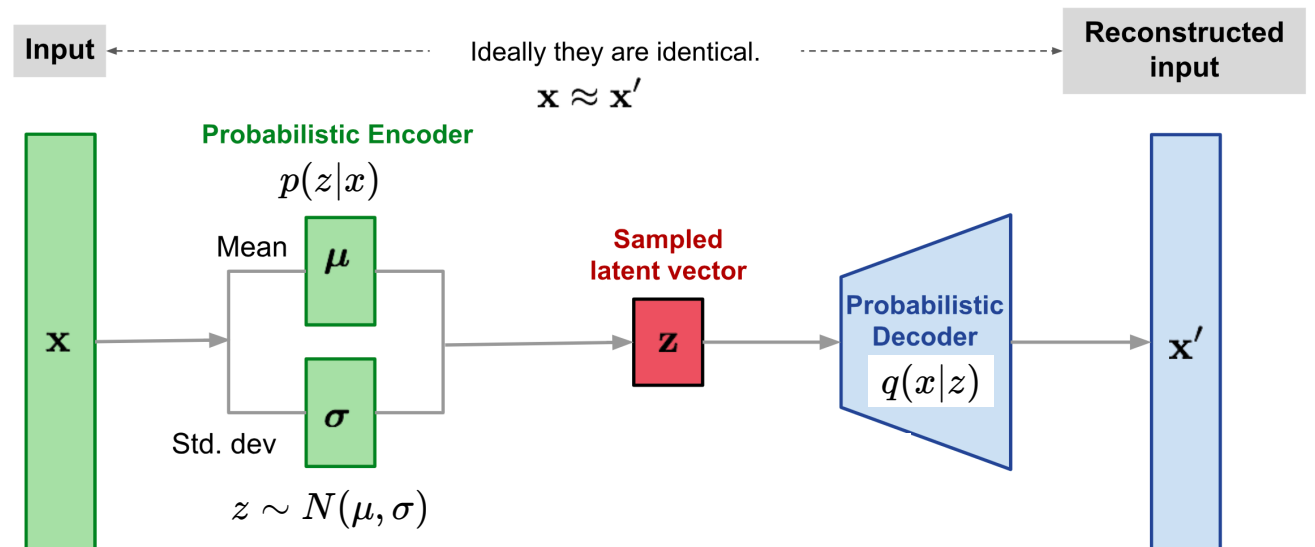
Вариационные автоэнкодеры

Тогда нам нужно выучить это отображение $p(z|x)$, а также обратное отображение $q(x|z)$ - эти функции и будут для нас энкодером и декодером

Мы все еще хотим использовать нейросети, и по некоторым математическим причинам (*reparameterization trick*) лучше сделать распределение в латентном пространстве многомерным нормальным

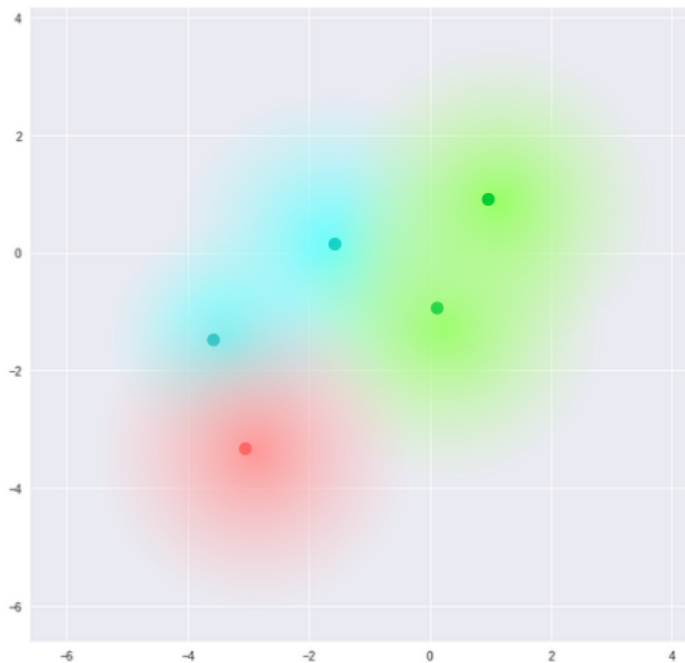
Энкодер будет генерировать для этого распределения:

- вектор средних
- вектор стандартных отклонений

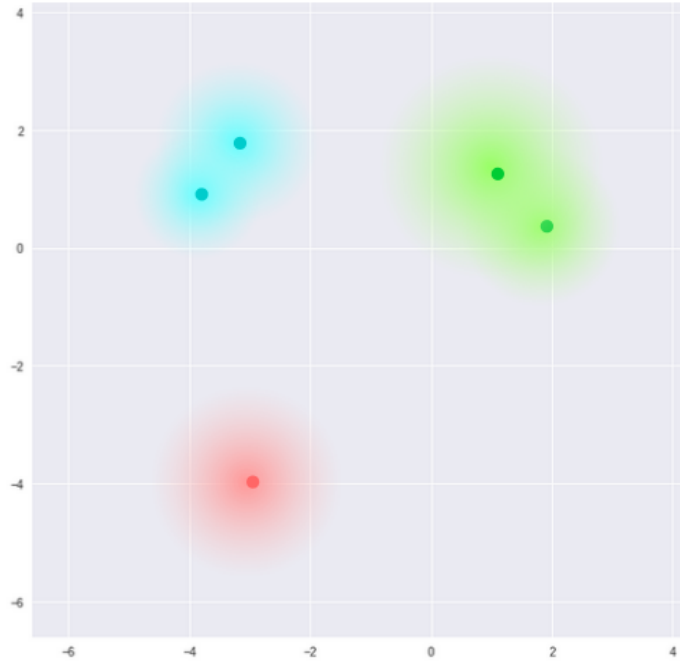


Решает ли такая модель с MSE-лоссом задачу нахождения непрерывного латентного пространства без ничему не соответствующих областей?

Решает ли такая модель с MSE-лоссом задачу нахождения непрерывного латентного пространства без ничему не соответствующих областей?



What we require

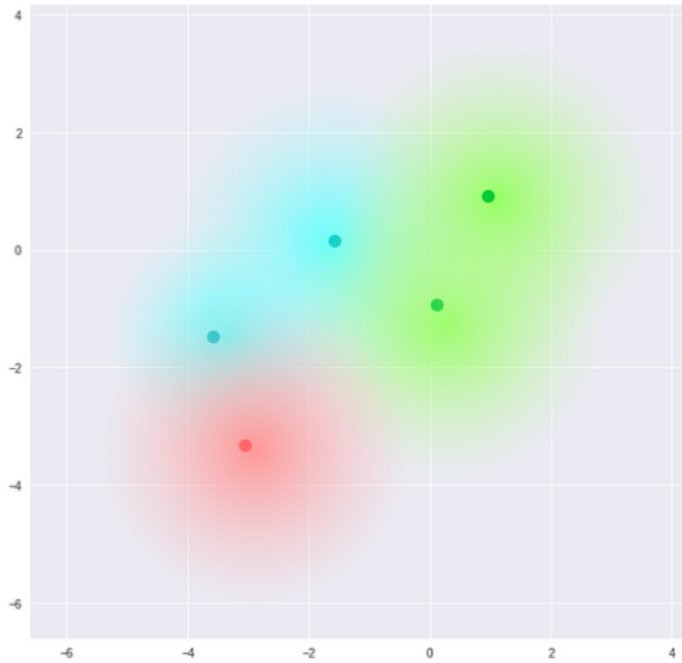


What we may inadvertently end up with

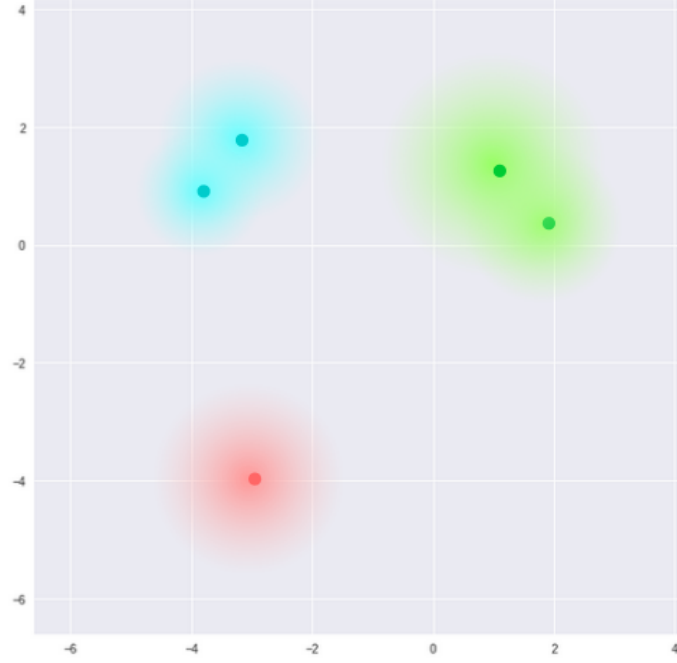
Не совсем: модель сможет занулить стандартные отклонения и работать только со средними

<https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf>

Решает ли такая модель с MSE-лоссом задачу нахождения непрерывного латентного пространства без ничему не соответствующих областей?



What we require



What we may inadvertently end up with

Не совсем: модель сможет занулить стандартные отклонения и работать только со средними

<https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf>

Чтобы решить эту проблему, давайте на время забудем про автоэнкодеры и поговорим про вероятностные модели

Вероятностные модели

У нас есть переменные x и параметры модели θ - это всё случайные величины

- $p(\theta)$ - априорное распределение
- $p(x|\theta)$ - функция правдоподобия модели
- $p(\theta|x)$ - апостериорное распределение

Из теоремы Байеса:

$$p(\theta|x) \propto p(x|\theta) * p(\theta)$$

Для удобства обозначим $p(x|\theta) \equiv p_\theta(x)$

Латентные переменные

Теперь добавим к нашей модели еще один набор случайных величин - латентные переменные z , чтобы получить представление наших данных в латентном пространстве

Функция правдоподобия:

$$p_{\theta}(x) = \int p_{\theta}(x, z) dz$$

Как правило, это неберущийся интеграл, а значит мы не умеем максимизировать правдоподобие по θ

Латентные переменные

Теперь добавим к нашей модели еще один набор случайных величин - латентные переменные z , чтобы получить представление наших данных в латентном пространстве

Функция правдоподобия:

$$p_{\theta}(x) = \int p_{\theta}(x, z) dz$$

Как правило, это неберущийся интеграл, а значит мы не умеем максимизировать правдоподобие по θ

Более того, не забываем про латентные переменные: мы хотели бы найти $p_{\theta}(z|x)$ в явном виде, но и тогда мы столкнемся со сложными интегралами:

$$p_{\theta}(x, z) = p_{\theta}(z|x) * p_{\theta}(x) \quad \Rightarrow \quad p_{\theta}(z|x) = \frac{p_{\theta}(x, z)}{\int p_{\theta}(x, z) dz}$$

ELBO и KL-divergence

Сделаем хитрость и введем $q_\phi(z) \in Q$ - некоторое выбранное нами семейство распределений с параметрами ϕ

Используя $p_\theta(x, z) = p_\theta(z|x) * p_\theta(x)$, получаем:

$$\int q_\phi(z) \ln \frac{p_\theta(x, z)}{q_\phi(z)} dz = \int q_\phi(z) \ln \frac{p_\theta(z|x)}{q_\phi(z)} dz + \int q_\phi(z) \ln p_\theta(x) dz$$

ELBO и KL-divergence

Сделаем хитрость и введем $q_\phi(z) \in Q$ - некоторое выбранное нами семейство распределений с параметрами ϕ

Используя $p_\theta(x, z) = p_\theta(z|x) * p_\theta(x)$, получаем:

$$\int q_\phi(z) \ln \frac{p_\theta(x, z)}{q_\phi(z)} dz = \int q_\phi(z) \ln \frac{p_\theta(z|x)}{q_\phi(z)} dz + \int q_\phi(z) \ln p_\theta(x) dz$$

$$\ln p_\theta(x) = \underbrace{\int q_\phi(z) \ln \frac{p_\theta(x, z)}{q_\phi(z)} dz}_{\text{ELBO}} + \underbrace{\left(- \int q_\phi(z) \ln \frac{p_\theta(z|x)}{q_\phi(z)} dz \right)}_{KL(q(z)||p(z|x))}$$

ELBO и KL-divergence

$$\ln p_{\theta}(x) = \underbrace{\int q_{\phi}(z) \ln \frac{p_{\theta}(x, z)}{q_{\phi}(z)} dz}_{\text{ELBO}} + \underbrace{\left(- \int q_{\phi}(z) \ln \frac{p_{\theta}(z|x)}{q_{\phi}(z)} dz\right)}_{KL(q(z)||p(z|x))}$$

Kullback–Leibler divergence (KL) - мера удаленности двух распределений, можно доказать, что она неотрицательна

ELBO и KL-divergence

$$\ln p_{\theta}(x) = \underbrace{\int q_{\phi}(z) \ln \frac{p_{\theta}(x, z)}{q_{\phi}(z)} dz}_{\text{ELBO}} + \underbrace{\left(- \int q_{\phi}(z) \ln \frac{p_{\theta}(z|x)}{q_{\phi}(z)} dz\right)}_{KL(q(z)||p(z|x))}$$

Kullback–Leibler divergence (KL) - мера удаленности двух распределений, можно доказать, что она неотрицательна

По определению $\text{ELBO} = \mathbb{E}[\ln p_{\theta}(x, z) - \ln q_{\phi}(z)]$, и его мы умеем оценивать, а значит, максимизируя его по θ , мы сможем максимизировать правдоподобие

ELBO и KL-divergence

$$\ln p_{\theta}(x) = \underbrace{\int q_{\phi}(z) \ln \frac{p_{\theta}(x, z)}{q_{\phi}(z)} dz}_{\text{ELBO}} + \underbrace{\left(- \int q_{\phi}(z) \ln \frac{p_{\theta}(z|x)}{q_{\phi}(z)} dz\right)}_{KL(q(z)||p(z|x))}$$

Kullback–Leibler divergence (KL) - мера удаленности двух распределений, можно доказать, что она неотрицательна

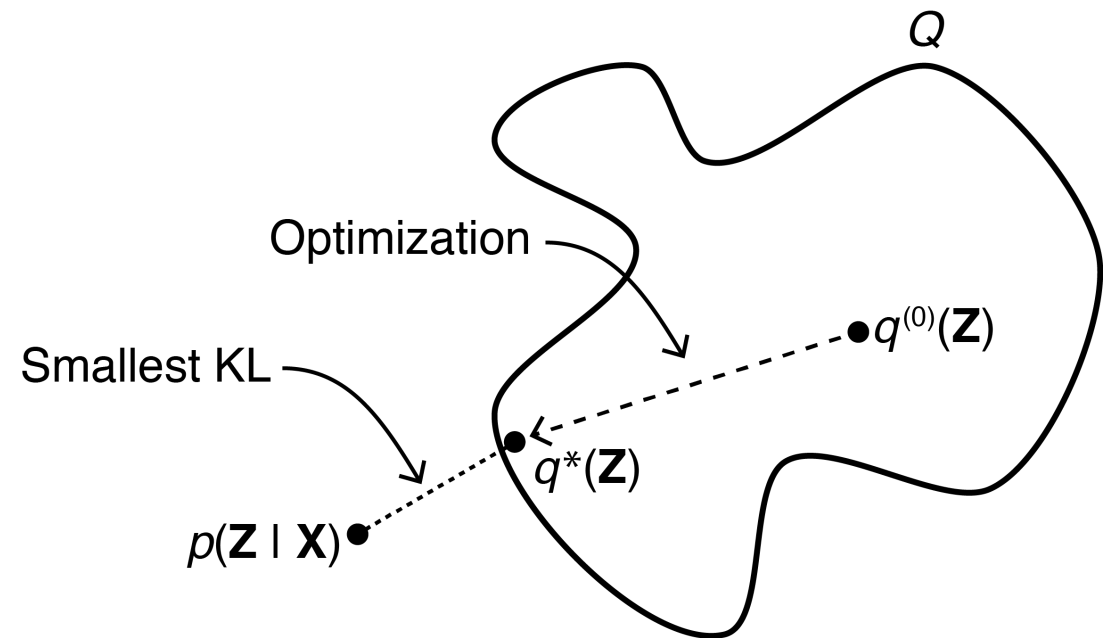
По определению $\text{ELBO} = \mathbb{E}[\ln p_{\theta}(x, z) - \ln q_{\phi}(z)]$, и его мы умеем оценивать, а значит, максимизируя его по θ , мы сможем максимизировать правдоподобие

Более того, зафиксировав θ и при этом максимизируя ELBO по ϕ , мы автоматически минимизируем KL-дивергенцию, а значит приближаем введенное нами распределение $q(z)$ к апостериорному $p(z|x)$

Variational inference

Мы только что ввели все основные понятия variational inference, цель которого - взять распределение $q_{\theta}(z)$ из семейства Q , с которым нам было бы удобно работать, и приблизить его к апостериорному распределению вместо прямого вычисления $p(z|x)$

А заодно, как мы убедились, и правдоподобие максимизировать



Обратно к VAE

Полученные функции вам ничего не напоминают?

В нашем случае $p_\theta(z|x)$ - энкодер, $q_\phi(x|z)$ - декодер, θ и ϕ - их параметры (веса нейросетей)

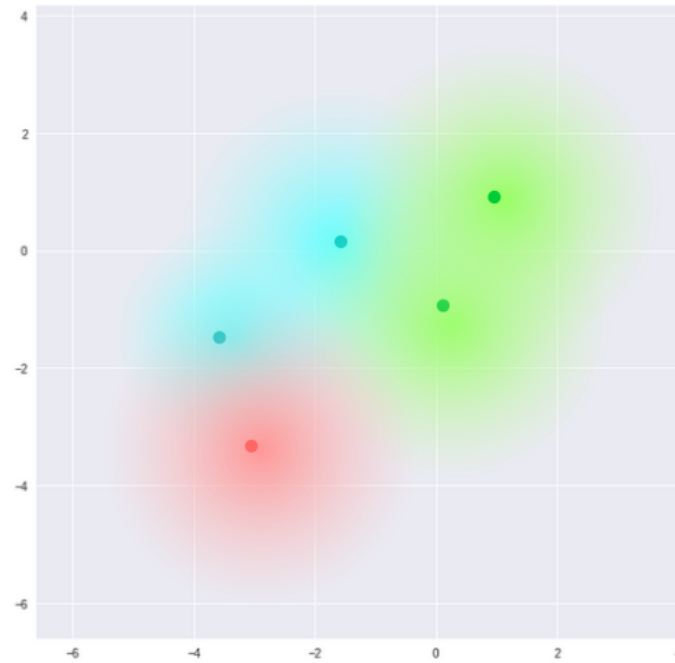
Обратно к VAE

Полученные функции вам ничего не напоминают?

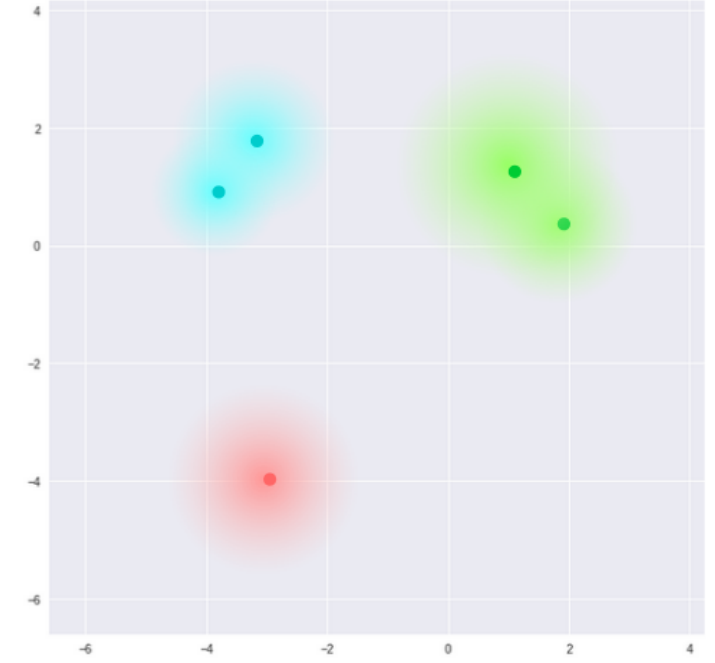
В нашем случае $p_{\theta}(z|x)$ - энкодер, $q_{\phi}(x|z)$ - декодер, θ и ϕ - их параметры (веса нейросетей)

Выбор ELBO в качестве loss function позволит нам:

- обучать модель кодировать объекты
- контролировать распределения латентных переменных, не допуская ситуаций, как на картинке справа



What we require



What we may inadvertently end up with

VAE для транскриптомики

Article | [Open Access](#) | [Published: 21 May 2018](#)

Interpretable dimensionality reduction of single cell transcriptome data with deep generative models

[Jiarui Ding](#) , [Anne Condon](#) & [Sohrab P. Shah](#) 

[Nature Communications](#) **9**, Article number: 2002 (2018) | [Cite this article](#)


21k Accesses | **139** Citations | **62** Altmetric | [Metrics](#)

Method

VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder

[Dongfang Wang](#) ^a, [Jin Gu](#) ^{a,b} 

scVAE: variational auto-encoders for single-cell gene expression data

[Christopher Heje Grønbech](#) , [Maximillian Fornitz Vording](#), [Pascal N Timshel](#), [Casper Kaae Sønderby](#), [Tune H Pers](#), [Ole Winther](#)

Bioinformatics, Volume 36, Issue 16, 15 August 2020, Pages 4415–4422, <https://doi.org/10.1093/bioinformatics/btaa293>

Published: 16 May 2020 [Article history](#) ▾

Article | [Published: 30 November 2018](#)

Deep generative modeling for single-cell transcriptomics

[Romain Lopez](#), [Jeffrey Regier](#), [Michael B. Cole](#), [Michael I. Jordan](#) & [Nir Yosef](#) 

[Nature Methods](#) **15**, 1053–1058 (2018) | [Cite this article](#)

51k Accesses | **407** Citations | **174** Altmetric | [Metrics](#)

Article | [Open Access](#) | [Published: 23 January 2019](#)

Single-cell RNA-seq denoising using a deep count autoencoder

[Gökçen Eraslan](#), [Lukas M. Simon](#), [Maria Mircea](#), [Nikola S. Mueller](#) & [Fabian J. Theis](#) 

[Nature Communications](#) **10**, Article number: 390 (2019) | [Cite this article](#)

47k Accesses | **288** Citations | **186** Altmetric | [Metrics](#)

Article | [Open Access](#) | [Published: 05 November 2018](#)

AutoImpute: Autoencoder based imputation of single-cell RNA-seq data

[Divyanshu Talwar](#), [Aanchal Mongia](#), [Debarka Sengupta](#)  & [Angshul Majumdar](#)

[Scientific Reports](#) **8**, Article number: 16329 (2018) | [Cite this article](#)

10k Accesses | **65** Citations | **21** Altmetric | [Metrics](#)

Interpretable factor models of single-cell RNA-seq via variational autoencoders

[Valentine Svensson](#) , [Adam Gayoso](#), [Nir Yosef](#), [Lior Pachter](#)

Bioinformatics, Volume 36, Issue 11, June 2020, Pages 3418–3421, <https://doi.org/10.1093/bioinformatics/btaa169>

Published: 16 March 2020 [Article history](#) ▾

Особенности транскриптомных данных

- Какое взять распределение в качестве $q(z)$?

Особенности транскриптомных данных

- Какое взять распределение в качестве $q(z)$?

Уже известное вам отрицательное биномиальное!

Ну или zero-inflated negative binomial (ZINB) распределение, которое в scVI почему-то используется по умолчанию :

$$y_i = \begin{cases} 0 \text{ with probability } \pi_i \\ NB(y_i | \dots) \text{ with probability } 1 - \pi_i \end{cases}$$

Особенности транскриптомных данных

- Какое взять распределение в качестве $q(z)$?

Уже известное вам отрицательное биномиальное!

Ну или zero-inflated negative binomial (ZINB) распределение, которое в scVI почему-то используется по умолчанию :

$$y_i = \begin{cases} 0 \text{ with probability } \pi_i \\ NB(y_i | \dots) \text{ with probability } 1 - \pi_i \end{cases}$$

- Дополнительные факторы:
 - Размер библиотеки разный для разных клеток
 - Батч-эффект

Архитектура scVI

N - число клеток, G - число генов

n, g - индексы клеток/генов

s_n - батч клетки

Два энкодера генерируют средние и стандартные отклонения для:

- $z_n \sim \text{Normal}(\mu_z, \Sigma_z)$ - закодированная экспрессия генов отдельно для каждой клетки, многомерное нормальное распределение с диагональной ковариационной матрицей
- $l_n \sim \text{LogNormal}(\mu_l, \sigma_l^2)$ - фактор размера библиотеки

Архитектура scVI

- Отрицательное биномиальное распределение - это распределение Пуассона, в котором λ - случайная величина, имеющая Гамма-распределение
- Декодер $f_w(z_n, s_n)$ позволяет найти оценку доли транскриптов гена в клетке: $w_{ng} \sim \text{Gamma}(f_w(z_n, s_n), \theta)$
- Находим оценку числа транскриптов: $y_{ng} \sim \text{Poisson}(l_n * w_{ng})$
- Декодер $f_h(z_n, s_n)$ позволяет найти вероятность дропаута гена в клетке
- В результате получаем:

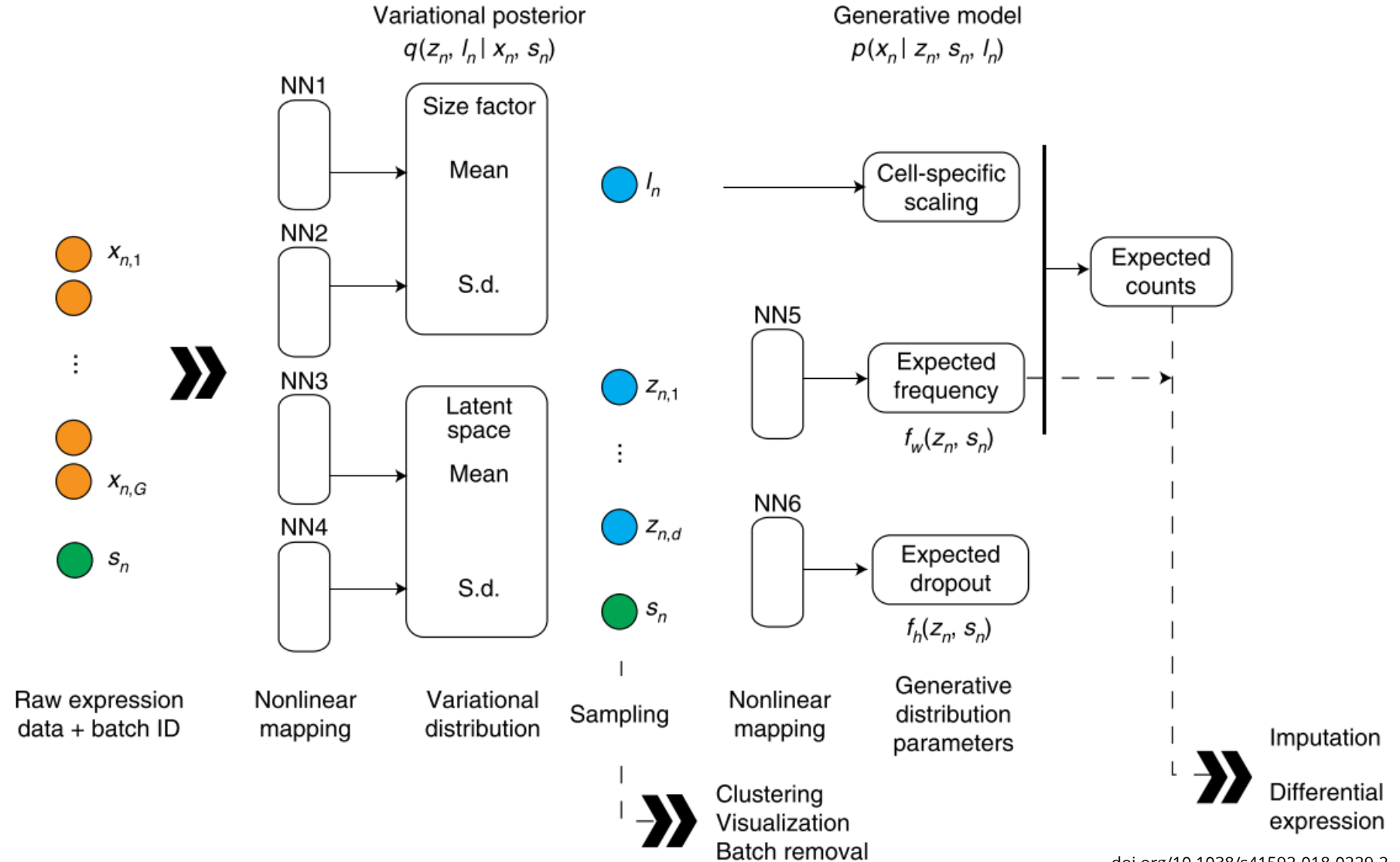
$$x_{ng} = \begin{cases} 0 & \text{with probability } f_h(z_n, s_n) \\ y_{ng} & \text{with probability } 1 - f_h(z_n, s_n) \end{cases}$$

Архитектура scVI

Обратите
внимание, здесь:

- $q(z, \dots | x, \dots)$ - энкодер
- $p(x | z, \dots)$ - декодер

Мы пользовались
обратными
обозначениями



*The variational autoencoder approach
is elegant, theoretically pleasing, and
simple to implement*

"Deep Learning", Goodfellow et al.

Variational models & hypothesis testing

- Если у нас есть апостериорное распределение, мы можем получать байесовские доверительные интервалы для проверки любых гипотез
- Например, делать дифф. экспрессию: с помощью декодера сэмплировать матрицу экспрессии и смотреть, как часто экспрессия гена X выше в одних клетках, чем в других

Интеграция датасетов

Вариационные автоэнкодеры можно использовать для интеграции атласов или мультимодальных датасетов

