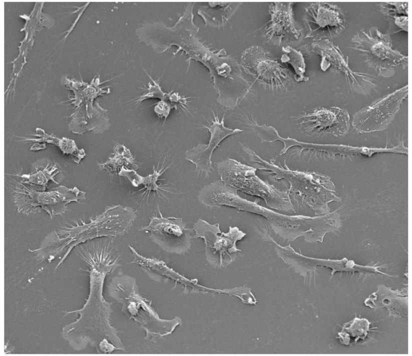
Recitation 6

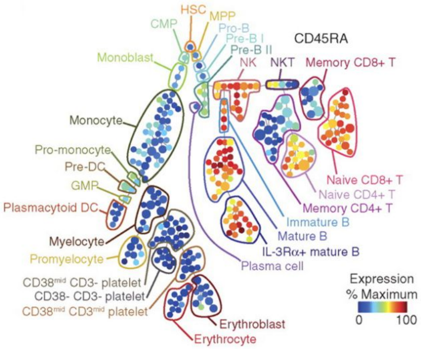
Single cells and their latent representations

Why single cells

Cellular heterogeneity

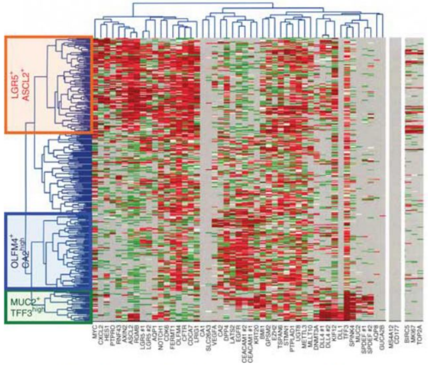


Differentiation trajectories



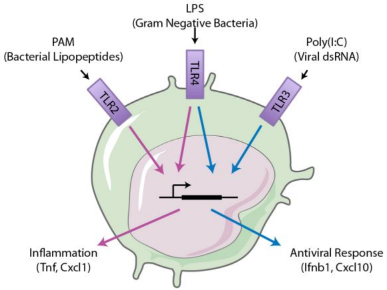
Bendall et al. (2011), Science

Within-cell-type differences

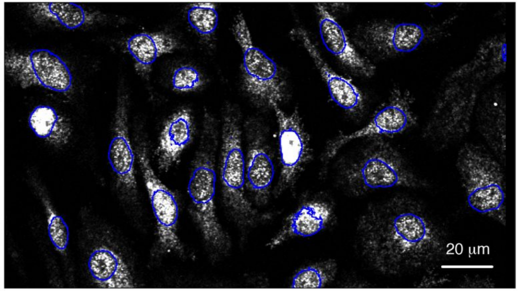


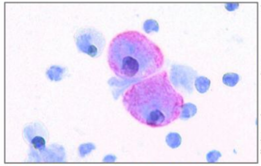
Dalerba et al. (2011), Nature Biotech

TLR Signaling



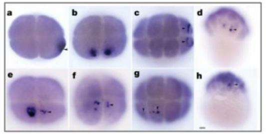
IRF3 Protein Levels - 4h LPS





fightcolorectalcancer.org/

Circulating Tumor Cells



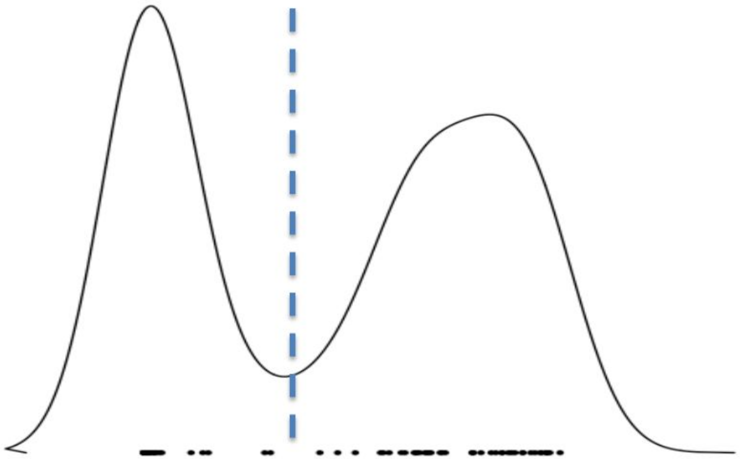
Gore et al. (2005), Nature

Zebrafish early embryo

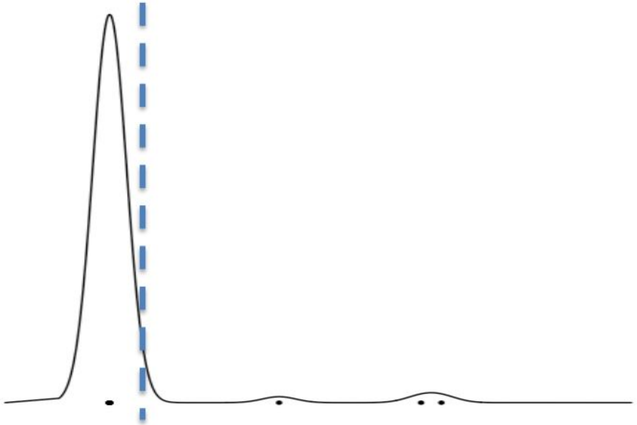
Cellular responses can vary substantially between "identical" cells.

Overcome low input

Whole-sample analysis can lead to misleading views



The average may not represent the population



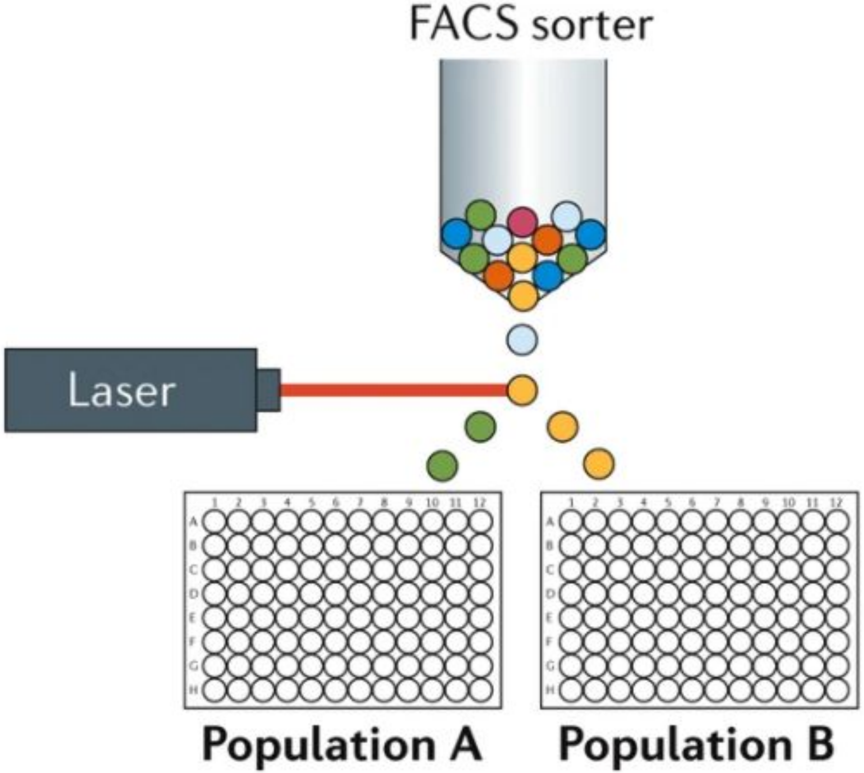
Rare events can be lost …

Common pipeline

1. Do something to make cells distinguishable
2. Amplify RNA and sequence

Cells in wells (SMART-seq)

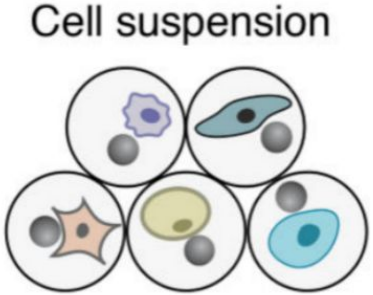
- Use fluorescence-activated cell sorting to get cells into individual wells
- Lyse cells, and carry out individual sequencing reactions for each well
- Analyze 50-500 cells

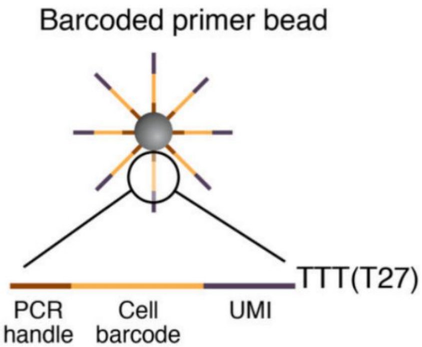


Paplexi et al., 2017

Droplets (Drop-seq)

- Isolate single cells into droplets that contain beads
- Bead is coated with barcoded primers
- Cell is lysed, RNA hybridizes to the barcoded primers
- Droplets are pooled and amplification+sequencing is done on the whole population
- Widely used, adoption by 10x Genomics

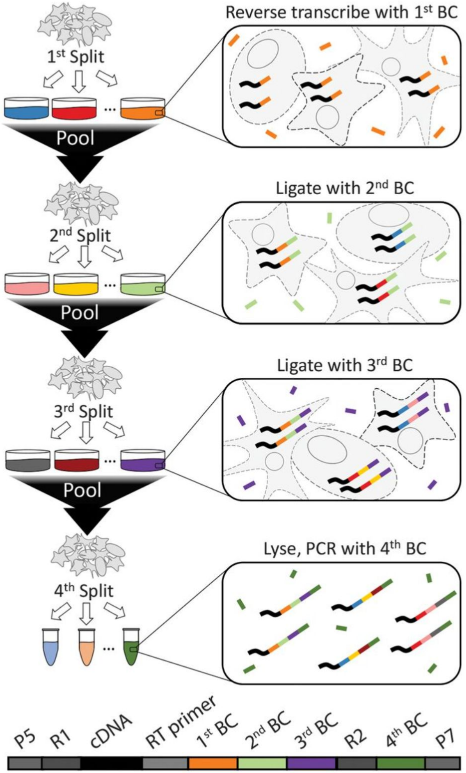

Cell suspension


Barcoded primer bead

Macosko et al., 2015

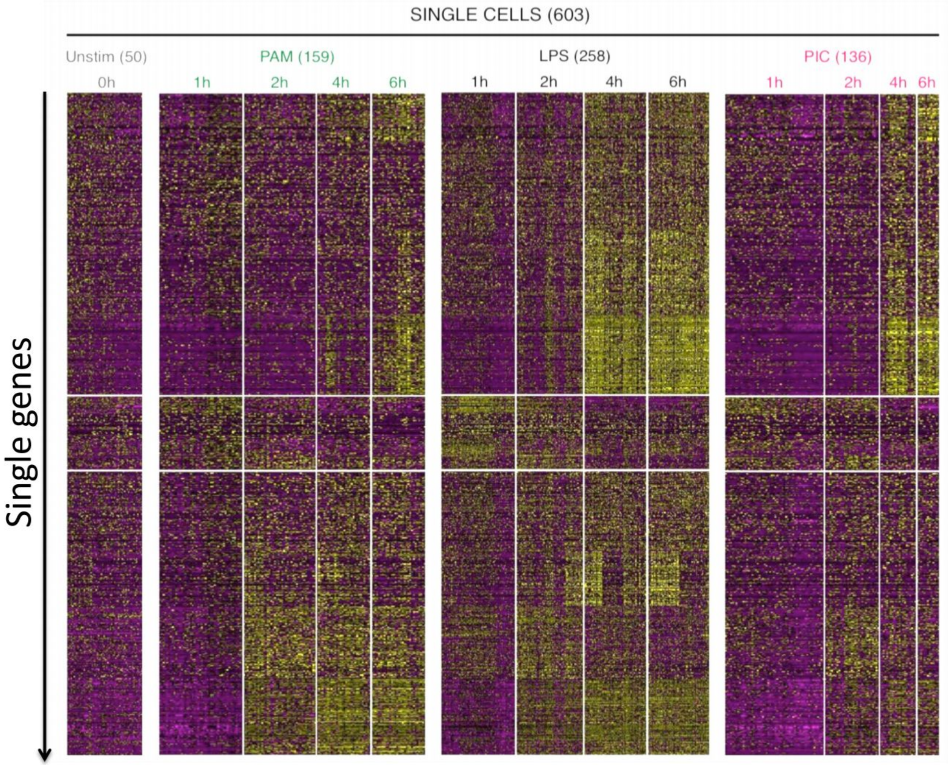
Combinatorial indexing (SPLiT-seq)

- Cells are put into a fixed state
- Cells are repeatedly pooled and randomly split into separate wells
- Cells in each well get a well specific barcode done through a reaction carried out in the cell
- The combinatorics of repeated barcoding ensure the unique labelling of each cell with high probability



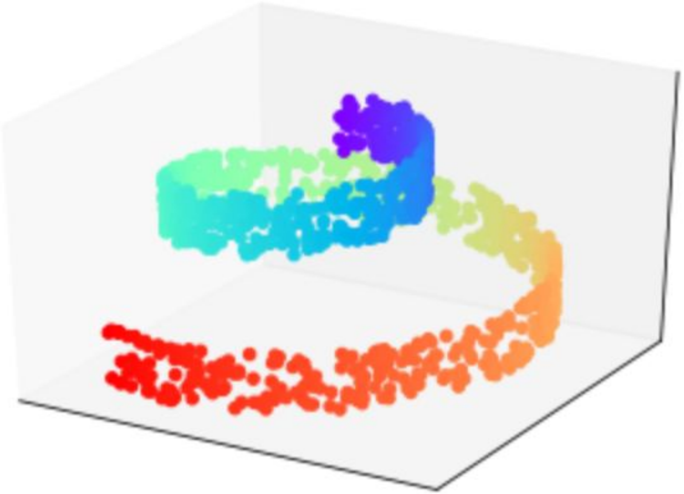
Rosenberg et al., 2018

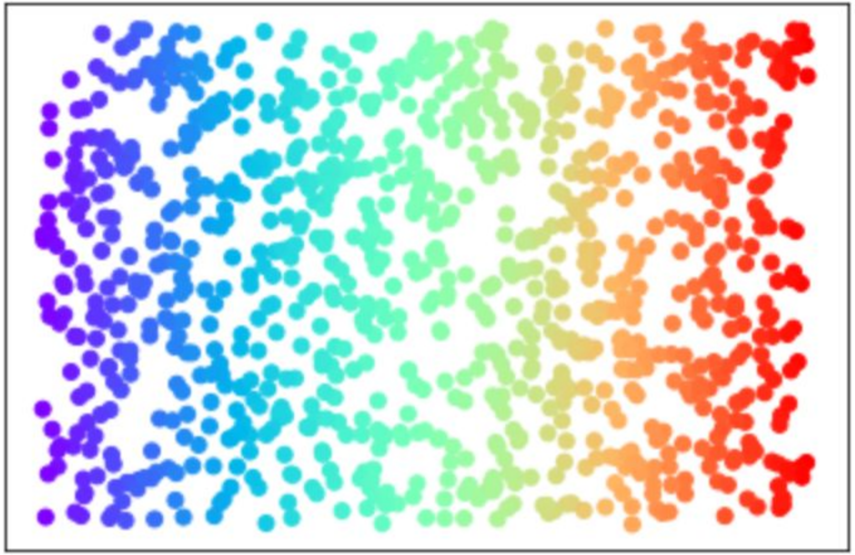
What can we do with this data?



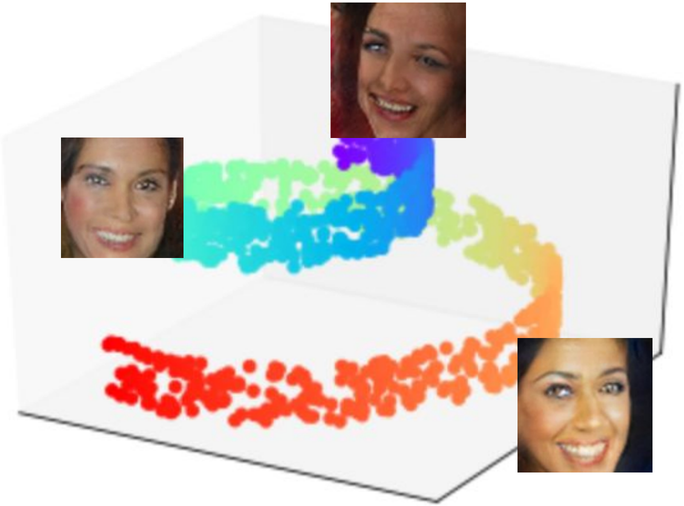
Shalek et al., 2014

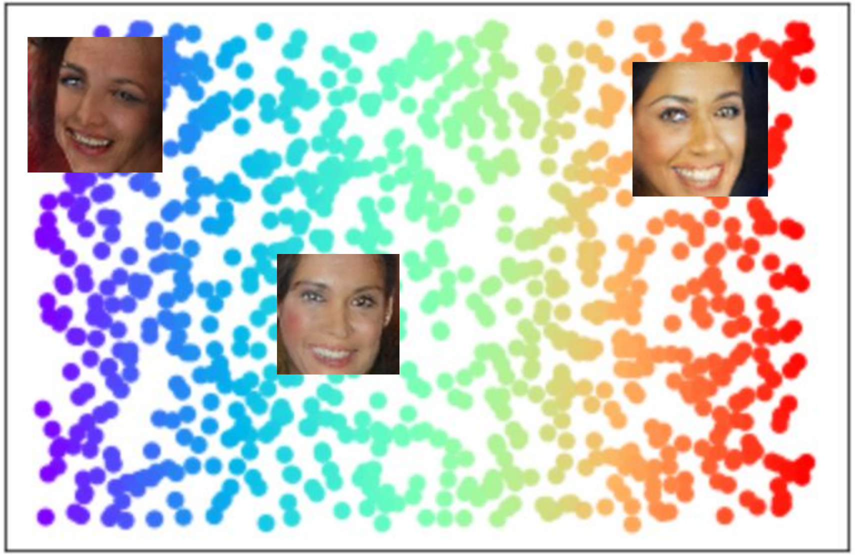
Samples in high dimensional spaces tend to live on lower dimensional surfaces





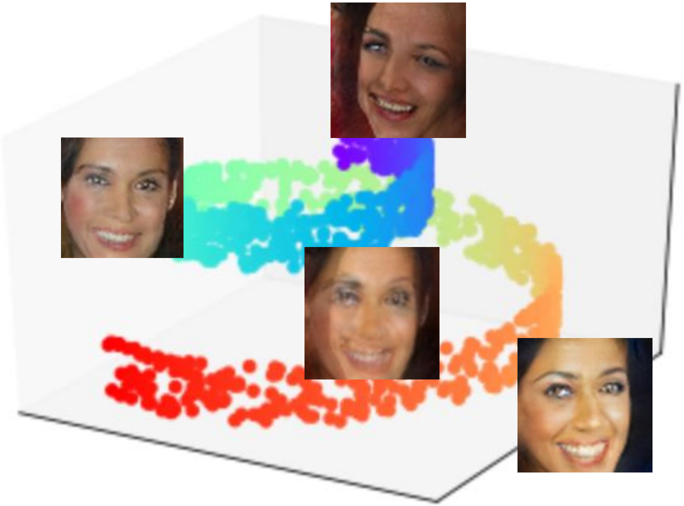
Data in high dimensional spaces tend to live on lower dimensional surfaces

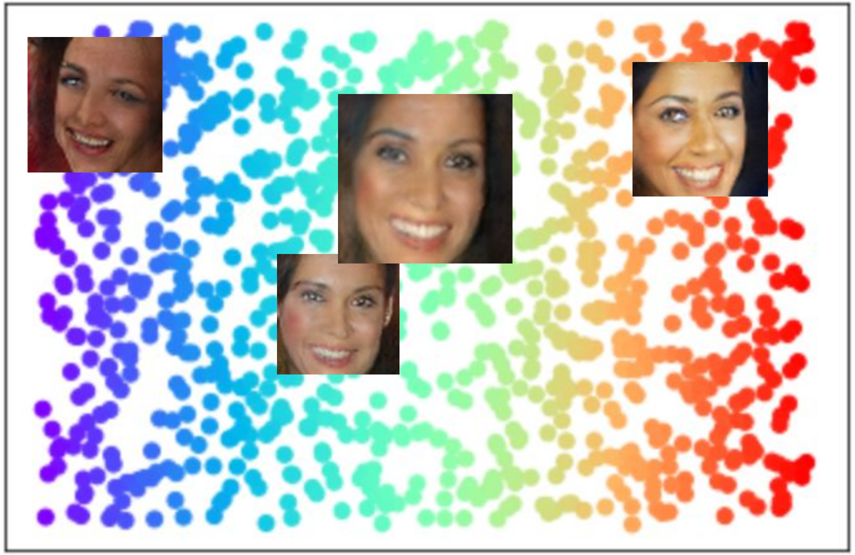




Radford et al., 2016

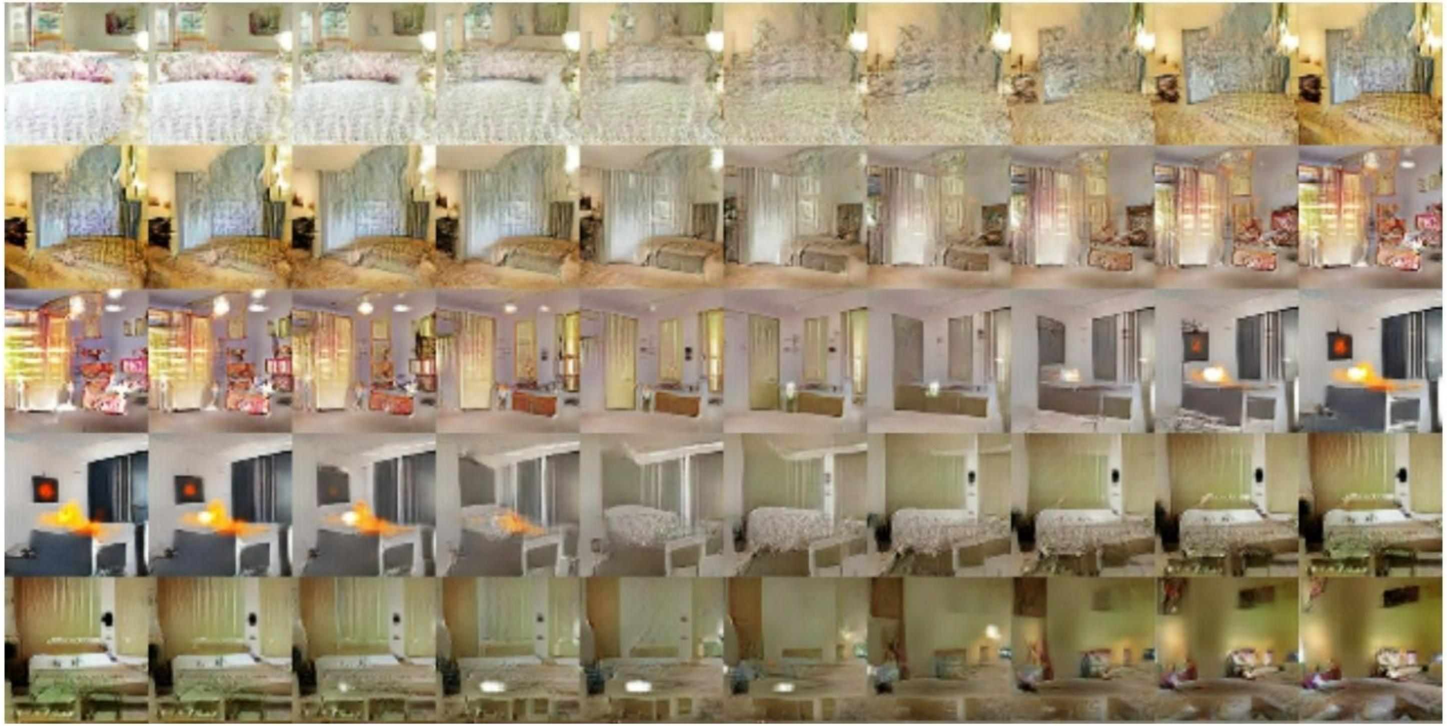
Averaging points in latent space





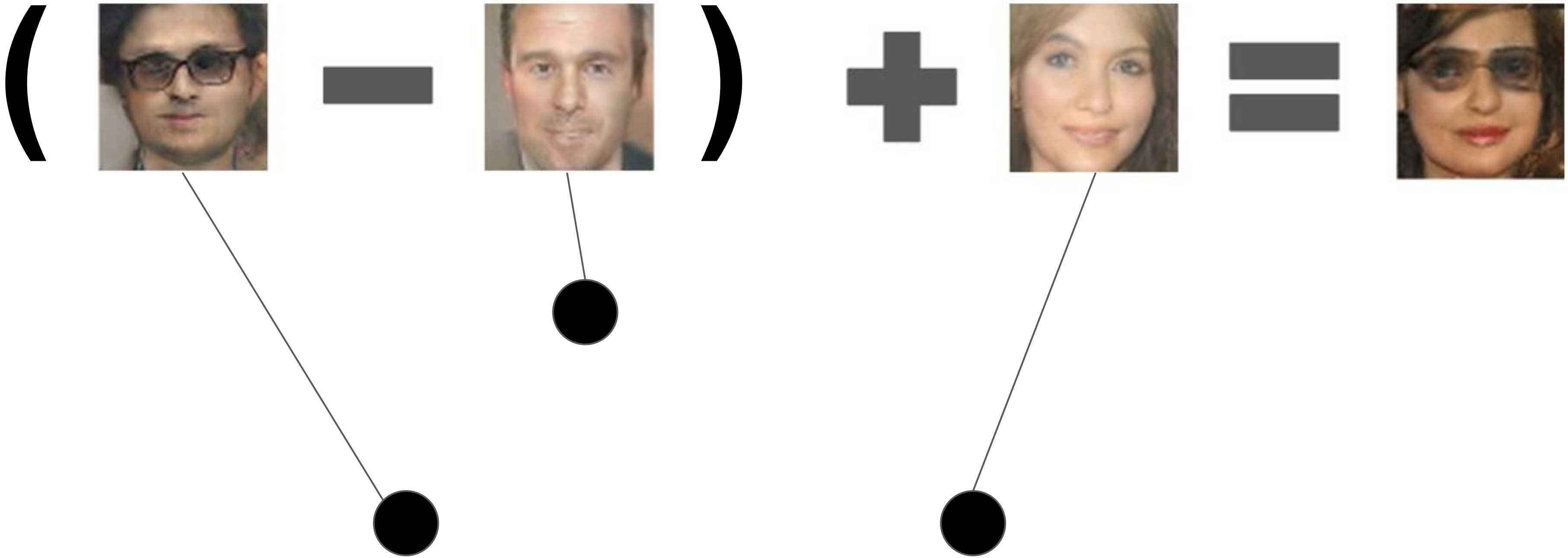
Radford et al., 2016

Interpolating in latent space



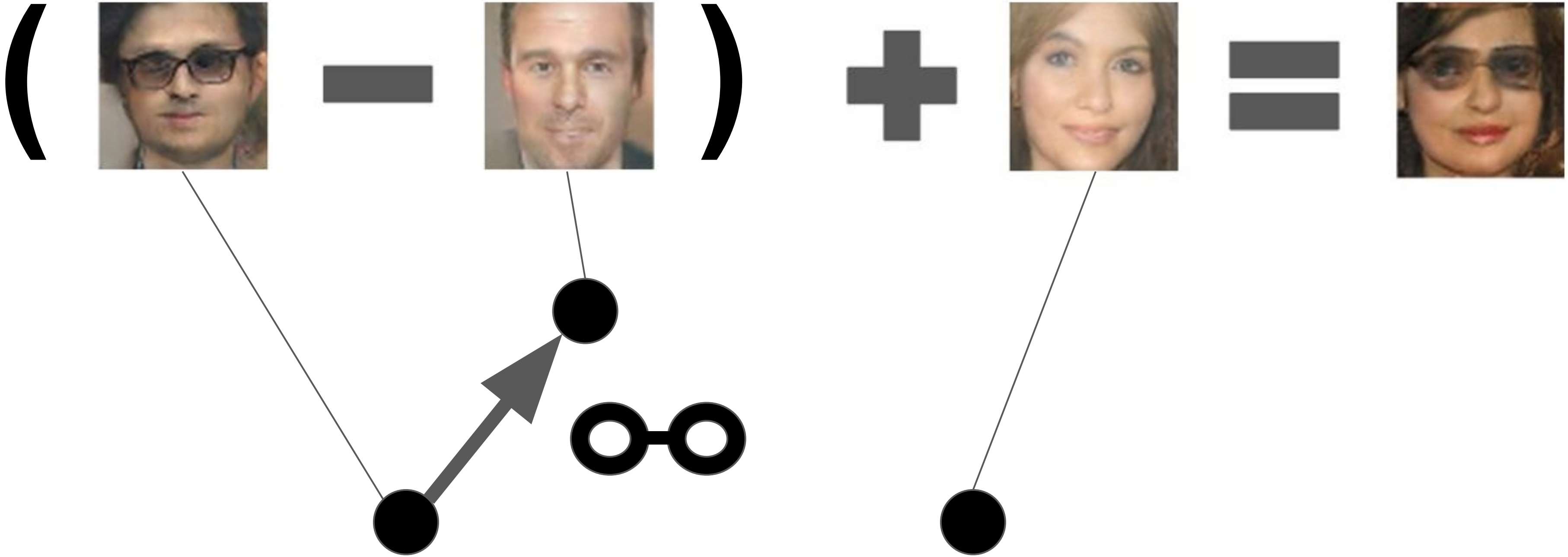
Radford et al., 2016

Algebra in latent space



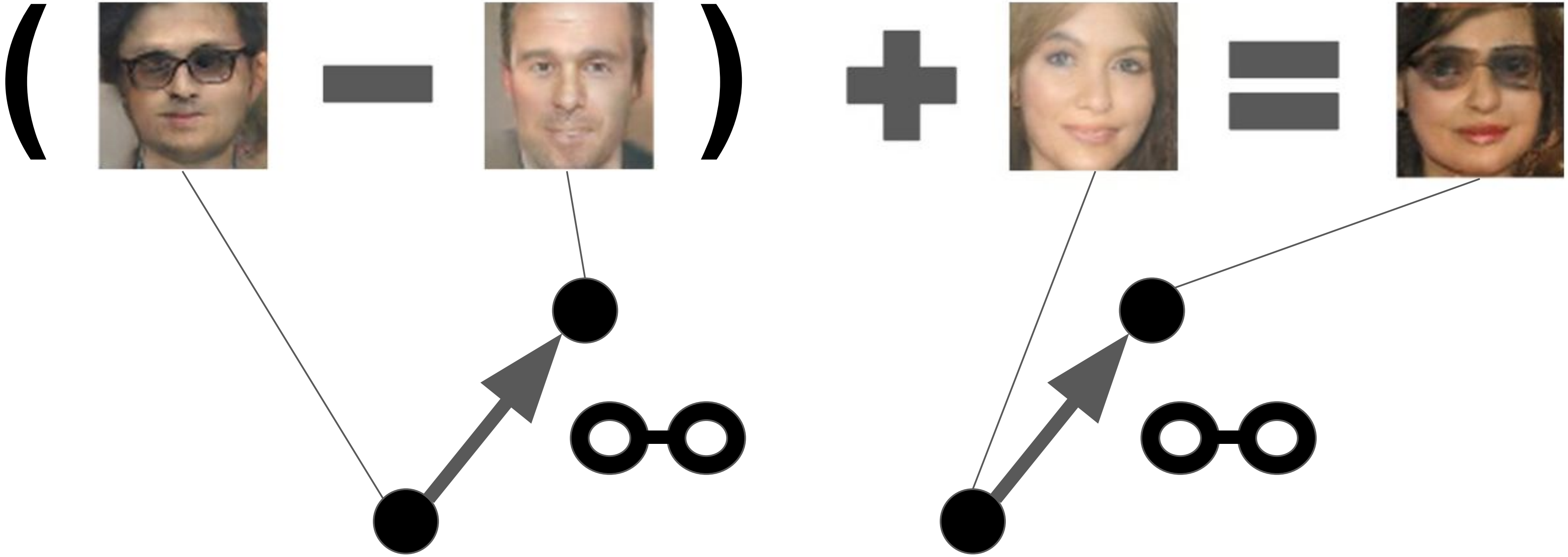
Radford et al., 2016

Algebra in latent space



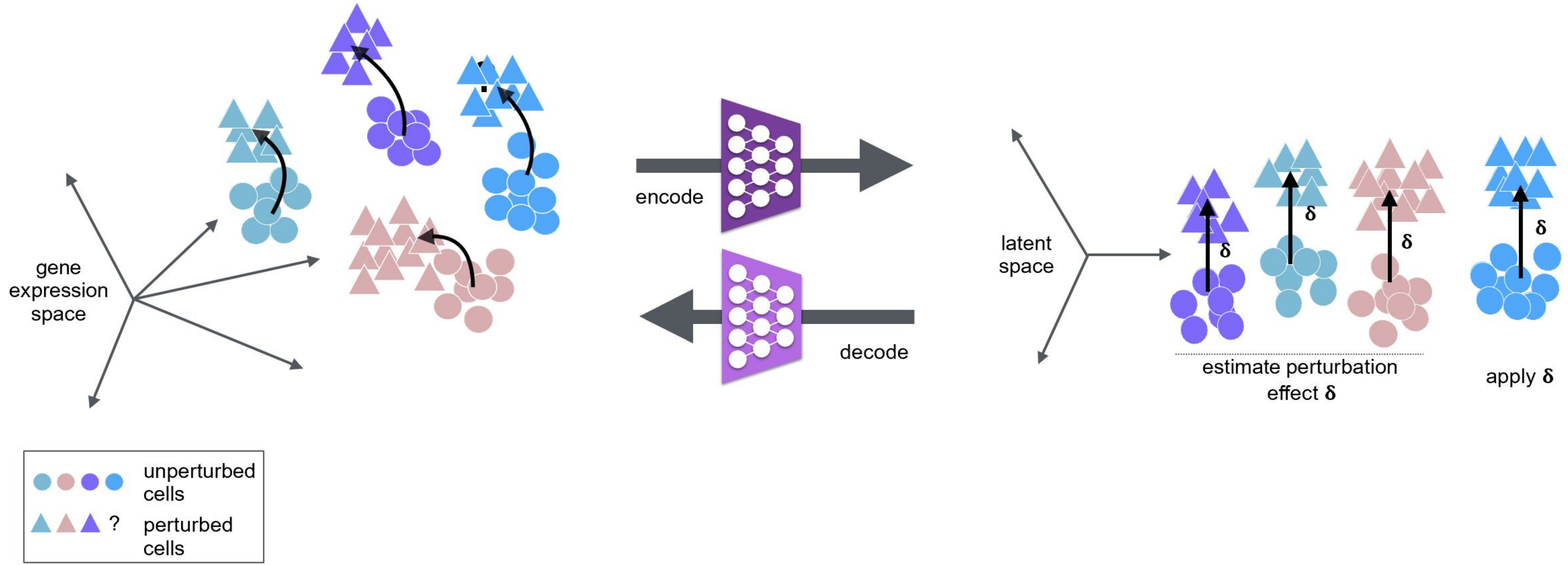
Radford et al., 2016

Algebra in latent space



Radford et al., 2016

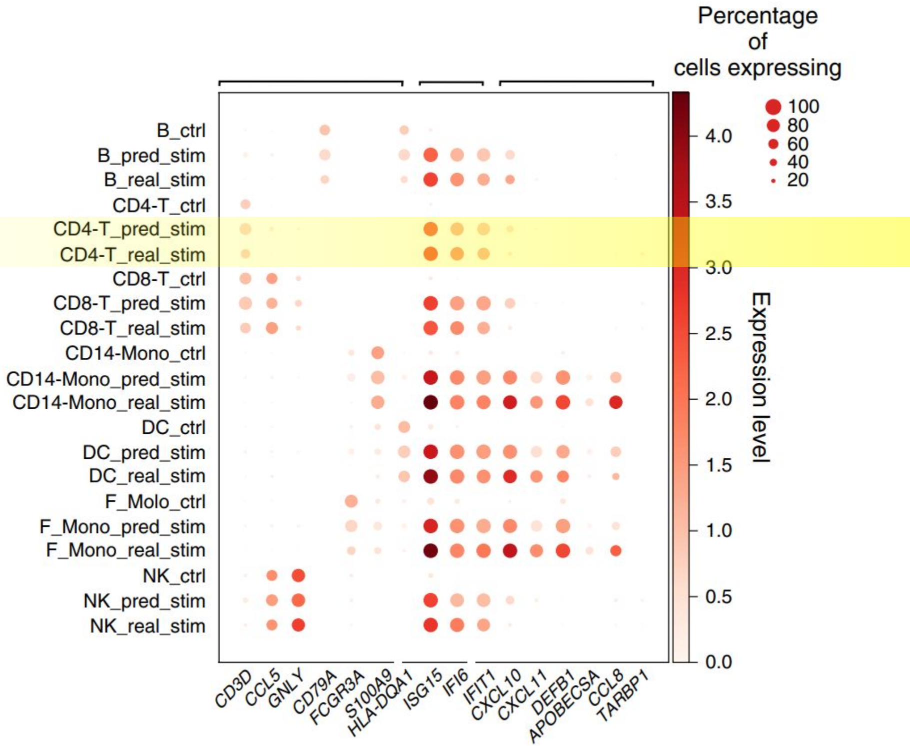
scGen: Apply this idea to scRNA data!



Lotfallahi et al., 2019

Perturbations can generalize to unseen data

- The highlighted cells was held out from the training data
- Model is still able to predict expression despite not being trained on it



Lotfallahi et al., 2019

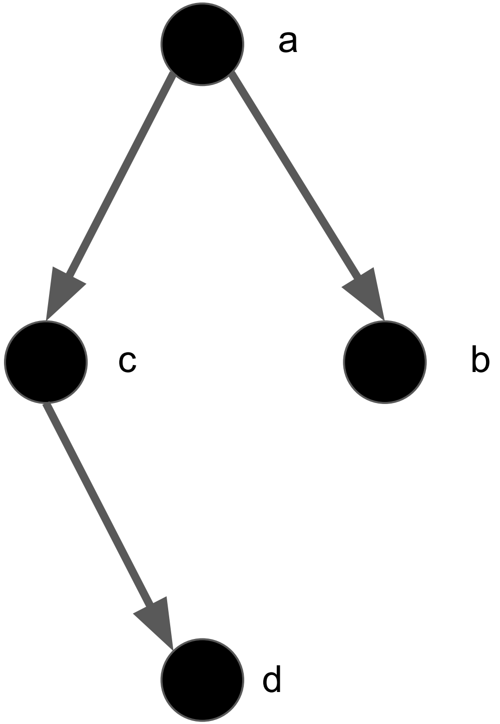
Correcting for artifacts with variational inference

- Many interesting downstream analyses like differential expression analysis is hindered by the presence of things like noise, artifacts, batch effects, etc
- We can use variational inference to correct for these

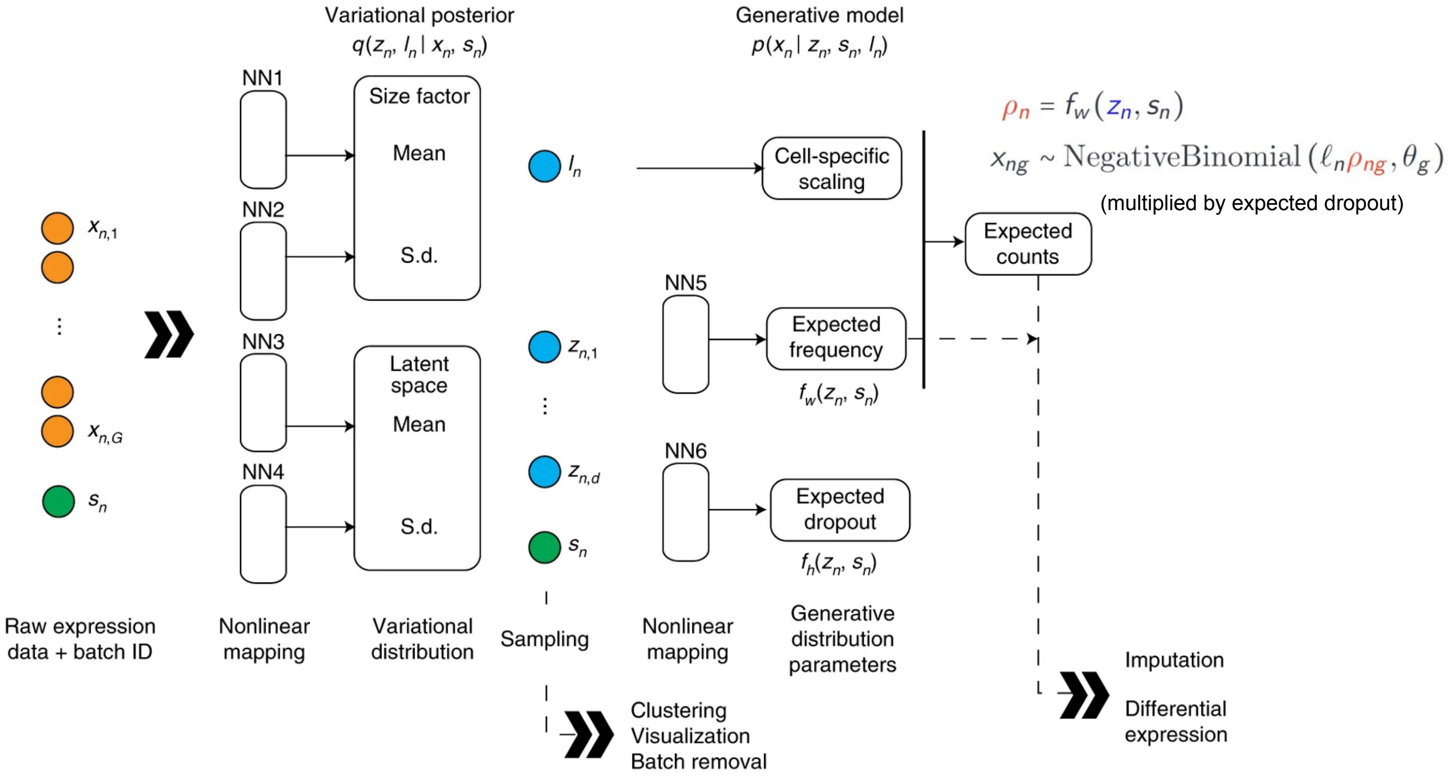
Graphical models factorize distributions

$P(a,b,c,d) = P_d(d|c)P_c(c|a)P_b(b|a)P_a(a)$

- Variables can be visible or hidden
- Illustrates conditional independence
    - Given a, b and c are independent
        - $P(b,c,d|a=1) = P_b(b|1)P_c(c|1)P_d(d|c)P_a(1)/P_a(1)$



scVI removes nuisance factors by factoring them out



$$\rho_n = f_w(z_n, s_n)$$

$$x_{ng} \sim \mathrm{NegativeBinomial}(\ell_n \rho_{ng}, \theta_g)$$

(multiplied by expected dropout)

Lopez et al., 2018

Refresher on VAE loss functions

x = elements of sample space

z = elements of latent space

lower case = deterministic variables

upper case = random variables

We want a generative model p that maximizes p(x) for our samples

Refresher on VAE loss functions

$$
\begin{aligned}
log\big(p(x)\big) &= log\Big(\int_z p(x|z)p(z)dz\Big) \\
&= log\Big(\int_z p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz\Big) \\
&= log\Big(\mathbb{E}_{Z\sim q(Z|x)}[\frac{p(x|Z)p(Z)}{q(Z|x)}]\Big) \\
&\geq \mathbb{E}_{Z\sim q(Z|x)}[log(\frac{p(x|Z)p(Z)}{q(Z|x)})] \\
&= \mathbb{E}_{Z\sim q(Z|x)}[log\big(p(x|Z)\big)] + \mathbb{E}_{Z\sim q(Z|x)}[log(\frac{p(Z)}{q(Z|x)})] \\
&= \mathbb{E}_{Z\sim q(Z|x)}[log\big(p(x|Z)\big)] - D_{KL}(q(Z|x)||p(Z))
\end{aligned}
$$

Step 1: Express p(x) as an expectation

$$log\big(p(x)\big) = log\Big(\int_z p(x|z)p(z)dz\Big)$$

Law of total probability

$$= log\Big(\int_z p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz\Big)$$

Multiply by 1

$$= log\Big(\int_z p(x|z)p(z)\frac{q(z|x)}{q(z|x)}dz\Big)$$

Reformulate as expectation

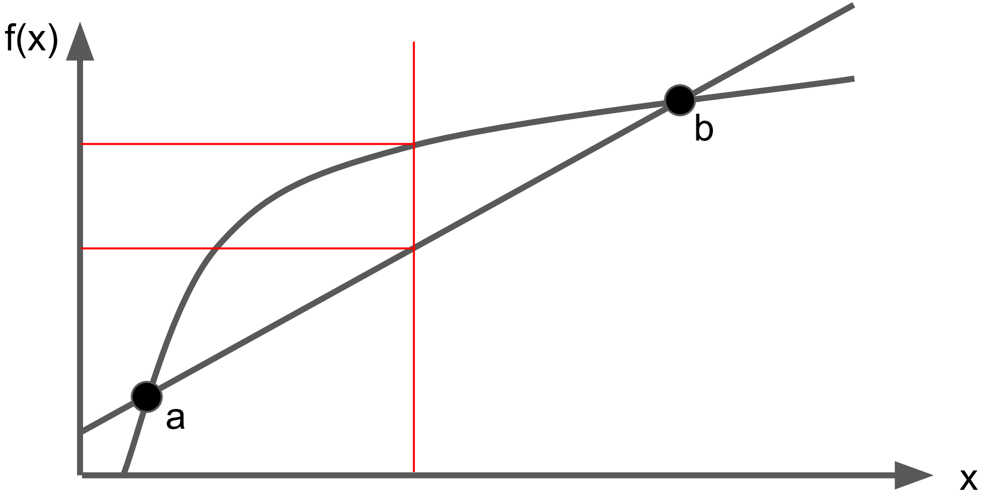
$$= log\Big(\mathbb{E}_{Z\sim q(Z|x)}[\frac{p(x|Z)p(Z)}{q(Z|x)}]\Big)$$

Step 2: Jensen's inequality

$$log(\mathbb{E}_{Z\sim q(Z|x)}[\frac{p(x|Z)p(Z)}{q(Z|x)}]) \;\geq\; \mathbb{E}_{Z\sim q(Z|x)}[log(\frac{p(x|Z)p(Z)}{q(Z|x)})]$$

For a concave function f:

f of the average
≥
average of f



Step 3: Break up the logarithms

$$\mathbb{E}_{Z\sim q(Z|x)}[log(\frac{p(x|Z)p(Z)}{q(Z|x)})]$$

$$= \mathbb{E}_{Z\sim q(Z|x)}[log\Big(p(x|Z)\Big)] + \mathbb{E}_{Z\sim q(Z|x)}[log(\frac{p(Z)}{q(Z|x)})]$$

$$= \mathbb{E}_{Z\sim q(Z|x)}[log\Big(p(x|Z)\Big)] - D_{KL}(q(Z|x)||p(Z))$$

Intuition

Want to maximize:

$$\mathbb{E}_{Z\sim q(Z|x)}[log\big(p(x|Z)\big)] - D_{KL}(q(Z|x)||p(Z))$$

$$\mathbb{E}_{Z\sim q(Z|x)}[log\big(p(x|Z)\big)]$$

Want the probability of reconstructing the original input x to be high.

$$- D_{KL}(q(Z|x)||p(Z))$$

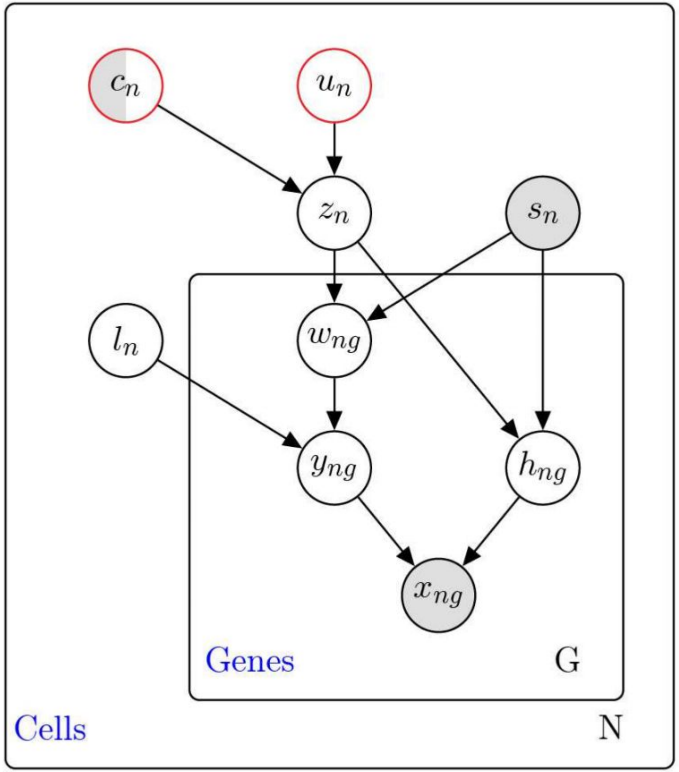
Want to “minimize” the distance between the posterior and the prior of the latent distribution.
Penalize encodings that drift very far.

Does the latent prior need to be unimodal Gaussian?

c_n is the cell state annotation

u_n represents additional variability

Model the latent variable (z_n) with as a mixture and treat c_n and u_n as mixture assignments



Xu et al., 2021

Information constraints on AEVB

- We restrain the search space for the variational distribution: in particular, we wish to enforce statements of the form $q(u) \perp\!\!\!\perp q(v)$.
- **Problem**: any measure of mutual information is intractable from the current graphical model and its variational approximation.
- **Solution**: we compute on each mini-batch a non-parametric measure of dependence from kernel embedding of joint distributions :

$$-\lambda \widehat{\mathrm{HSIC}}(q(u, v)),$$

where $\widehat{\mathrm{HSIC}}$ is the empirical estimate of the Hilbert-Schmidt norm of the cross-covariance operator $\mathcal{C}_{q(u,v)}$ that embeds the joint.

We call this modification **HSIC Constrained VAEs (HCV)**.

Lopez et al., Neural Information Processing Systems, (2018)

$$\mathrm{H\hat{S}IC}_n(P) = \frac{1}{n^2}\sum_{i,j}^{n} k(u_i, u_j)l(v_i, v_j) + \frac{1}{n^4}\sum_{i,j,k,l}^{n} k(u_i, u_j)l(v_k, v_l) - \frac{2}{n^3}\sum_{i,j,k}^{n} k(u_i, u_j)l(v_i, v_k).$$

$$\underbrace{\log p_\theta(x)}_{\text{evidence}} = \underbrace{\mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z \mid x)}}_{\text{ELBO}} + \underbrace{\Delta_{\mathrm{KL}}(q_\phi \parallel p_\theta)}_{\text{reverse KL VG}}, \tag{VI}$$

$$\underbrace{\log p_\theta(x)}_{\text{evidence}} = \underbrace{\log \mathbb{E}_{p_\theta(z|x)} \frac{p_\theta(x,z)}{q_\phi(z \mid x)}}_{\text{EUBO}} - \underbrace{\Delta_{\mathrm{KL}}(p_\theta \parallel q_\phi)}_{\text{forward KL VG}}, \tag{RWS}$$

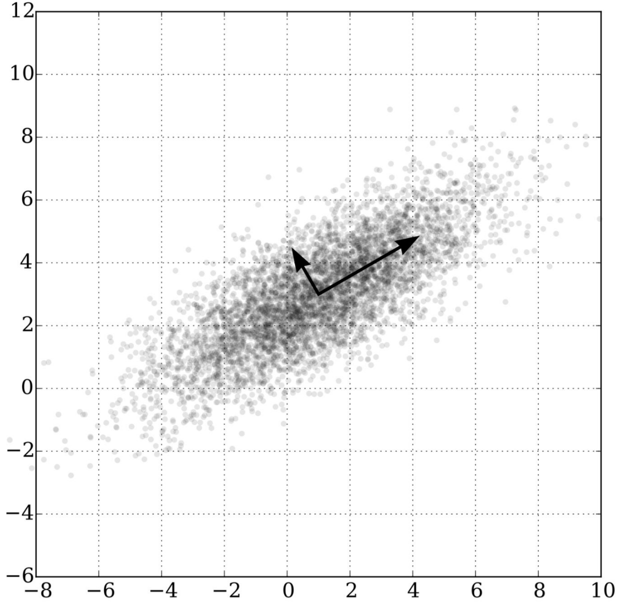
$$\underbrace{\log p_\theta(x)}_{\text{evidence}} = \underbrace{\frac{1}{2} \log \mathbb{E}_{q_\phi(z|x)} \left(\frac{p_\theta(x,z)}{q_\phi(z \mid x)} \right)^2}_{\text{CUBO}} - \underbrace{\frac{1}{2} \log \left(1 + \Delta_{\chi^2}(p_\theta \parallel q_\phi) \right)}_{\chi^2 \text{ VG}}. \tag{CHIVI}$$

Slide from Lopez

PCA - principal component analysis

Idea: we want to capture the axis where the most variability comes from

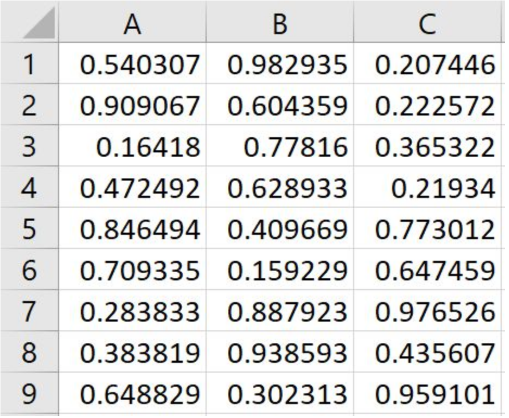
Other dimensions are “unimportant”



https://commons.wikimedia.org/wiki/File:GaussianScatterPCA.svg

Formulation

Consider a data matrix where each row represents a data point



	A	B	C
1	0.540307	0.982935	0.207446
2	0.909067	0.604359	0.222572
3	0.16418	0.77816	0.365322
4	0.472492	0.628933	0.21934
5	0.846494	0.409669	0.773012
6	0.709335	0.159229	0.647459
7	0.283833	0.887923	0.976526
8	0.383819	0.938593	0.435607
9	0.648829	0.302313	0.959101

Since we want to capture variance, the first thing we do is to shift each column such that all features have zero mean

Formulation

Let X be the resulting data matrix

We want to find a direction (unit vector) **w** in the inputs space such that the following expression is maximized

$$||Xw||^2$$

Xw is a vector where each entry is the projection of a sample on **w**

The square sum is then proportional to the variance

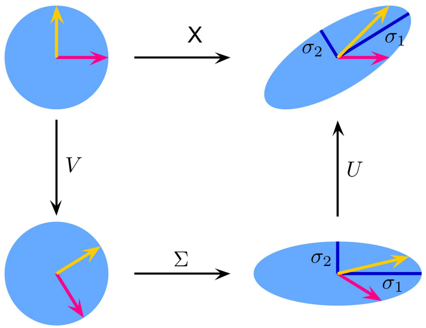
Singular value decomposition

Every matrix X has a singular value decomposition (SVD):

$$X = U\Sigma V$$

Where U and V are orthonormal matrices, and **Σ** is a diagonal matrix

Viewing X as a linear operator, you can think of U and V as rotations and **Σ** as scaling



https://commons.wikimedia.org/wiki/File:Singular-Value-Decomposition.svg

Singular value decomposition

U and V come from the fact that XX^T is symmetric and therefore has an orthogonal set of eigenvectors. **Σ** are equal because X^TX and XX^T share eigenvalues.

$$XX^T = U\Sigma^2U^T$$

$$X^TX = V^T\Sigma^2V$$

We can maximize Xw by picking the largest diagonal entry in ***Σ***

Since U and V are orthonormal:

|w| = |w_1| and |w_2| = |w_3|

$$
\begin{aligned}
Xw &= U\Sigma V w \\
&= U\Sigma w_1 \\
&= U w_2 \\
&= w_3
\end{aligned}
$$

Therefore the only scaling occurs when we multiply by ***Σ***

We can maximize this by selecting w such that w_1 is a “one-hot” vector for the largest eigenvalue coordinate in ***Σ***

Another view of PCA: networks without non-linearities