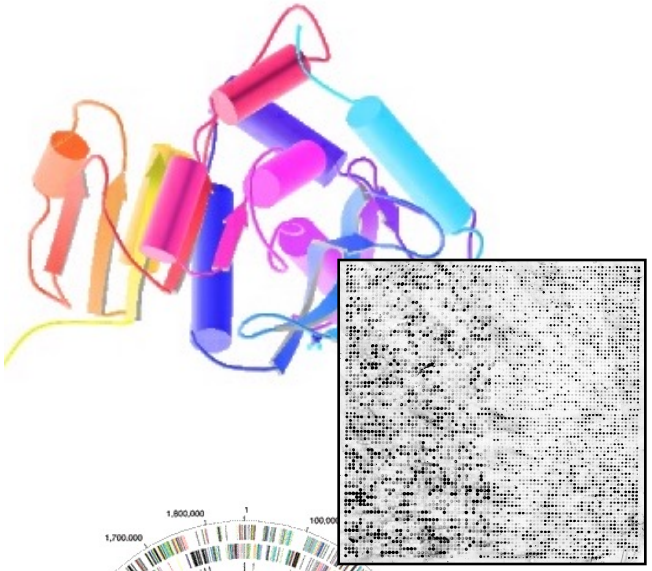
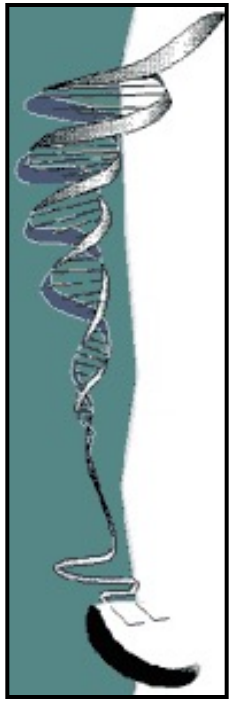
Biomed. Data Science:

Unsupervised Datamining E: LDA & tSNE





Mark Gerstein, Yale University

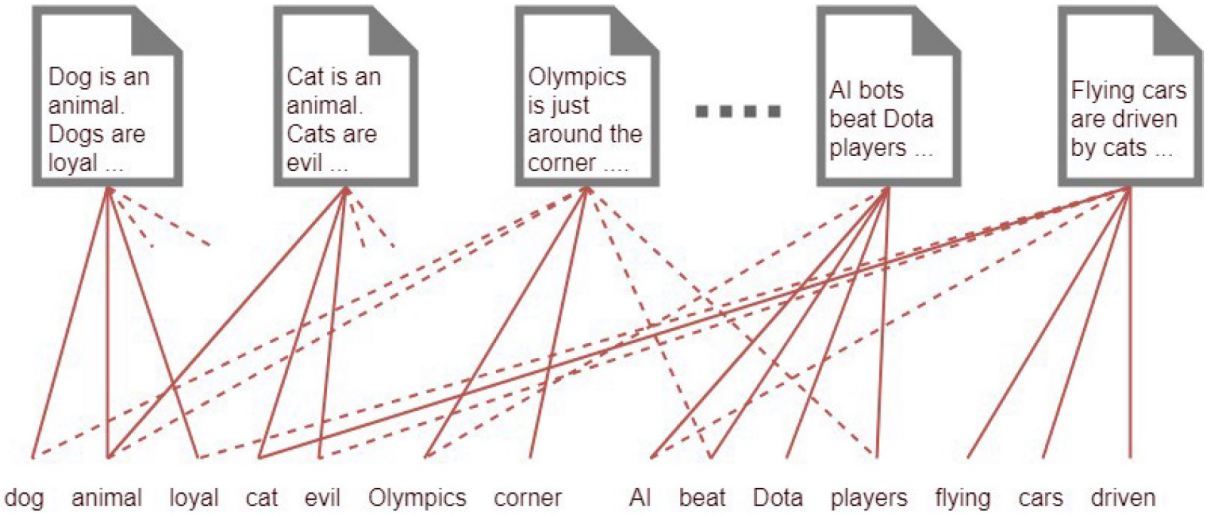
gersteinlab.org/courses/452

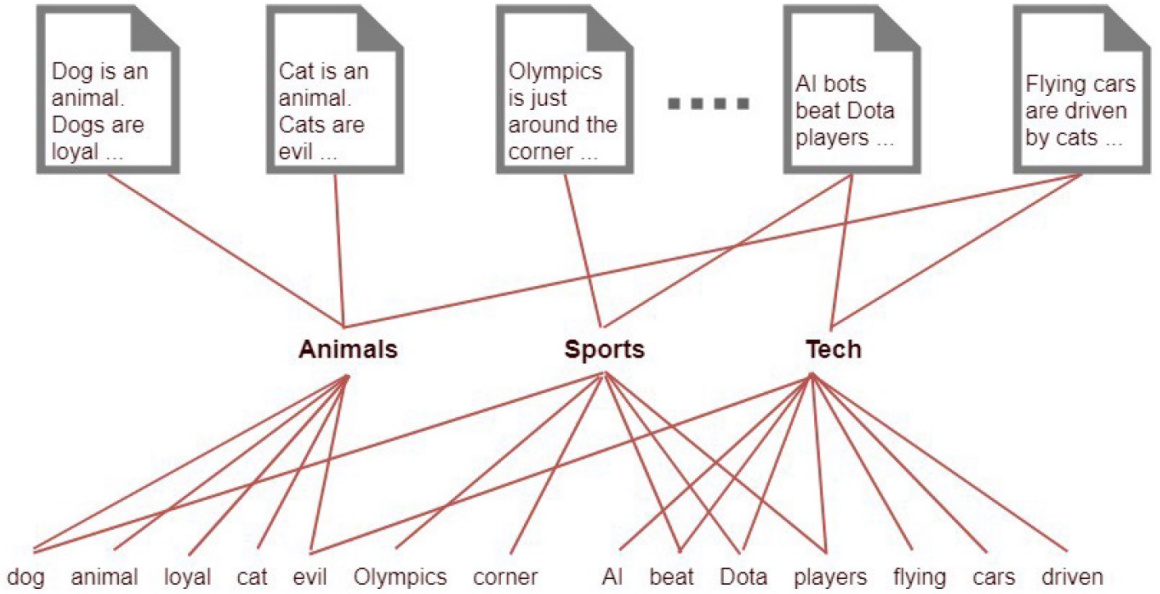
(last edit in spring ’21, pack #9e, final)

Unsupervised Mining

LDA

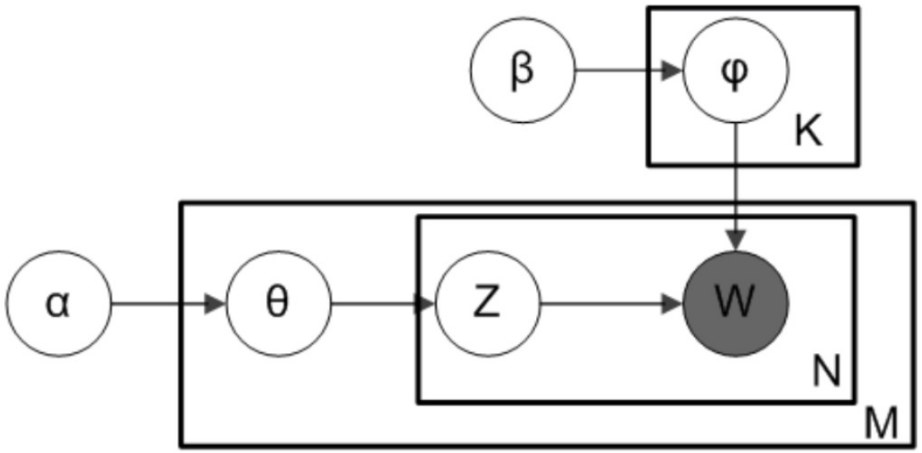
From dimension reduction view





https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158

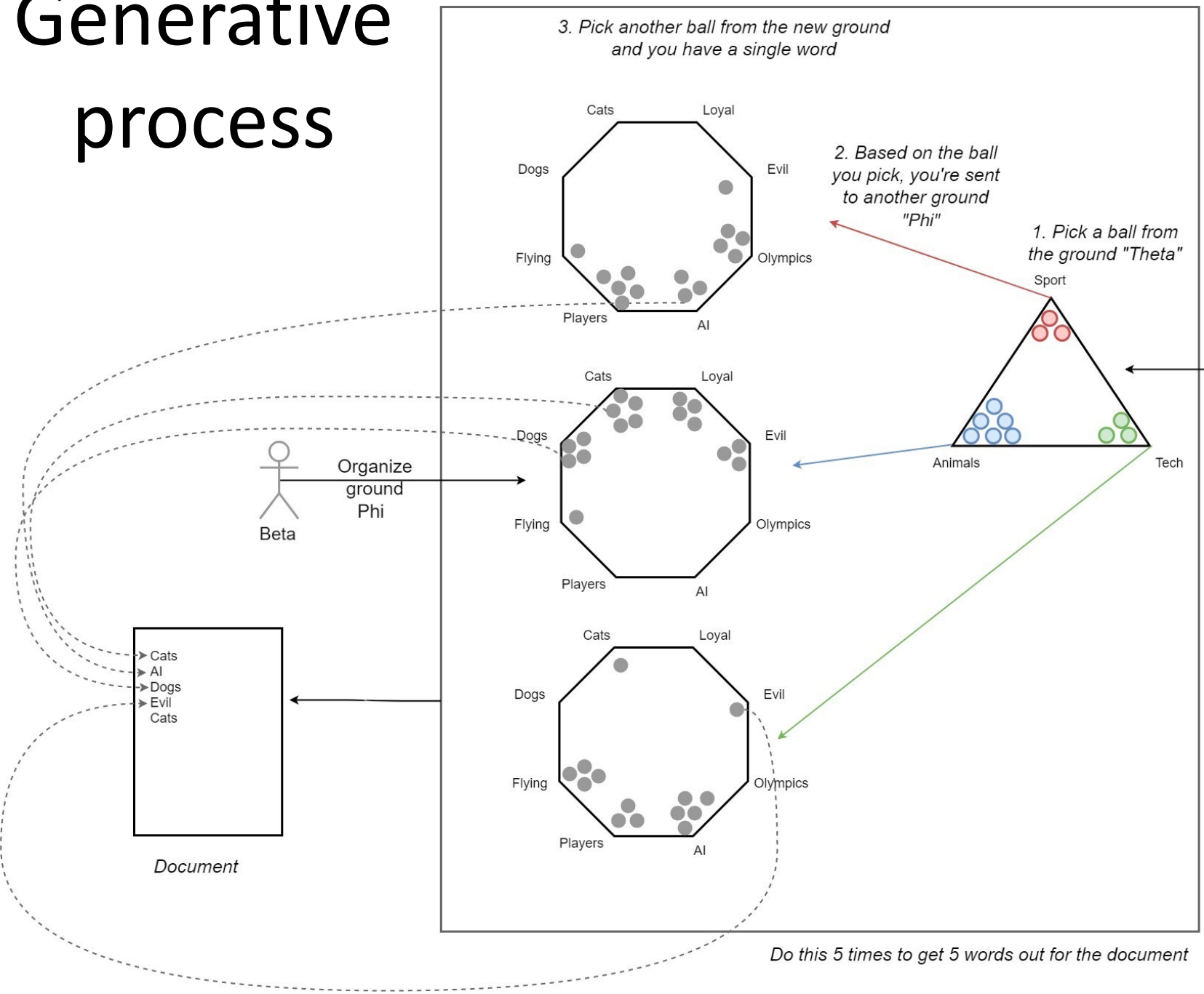
Diagram



$$\varphi_{k=1\ldots K} \sim \text{Dirichlet}_V(\beta)$$
$$\theta_{d=1\ldots M} \sim \text{Dirichlet}_K(\alpha)$$
$$z_{d=1\ldots M, w=1\ldots N_d} \sim \text{Categorical}_K(\theta_d)$$
$$w_{d=1\ldots M, w=1\ldots N_d} \sim \text{Categorical}_V(\varphi_{z_{dw}})$$

α is the parameter of the Dirichlet prior on the per-document topic distributions
β is the parameter of the Dirichlet prior on the per-topic word distribution
θ_i is the topic distribution for document i
φ_k is the word distribution for topic k
$z_{ij}$is the topic for the jth word in document i
w_{ij} is the specific word, and
K is the number of topics, N is the number of word in a document, M is the number of Documents.

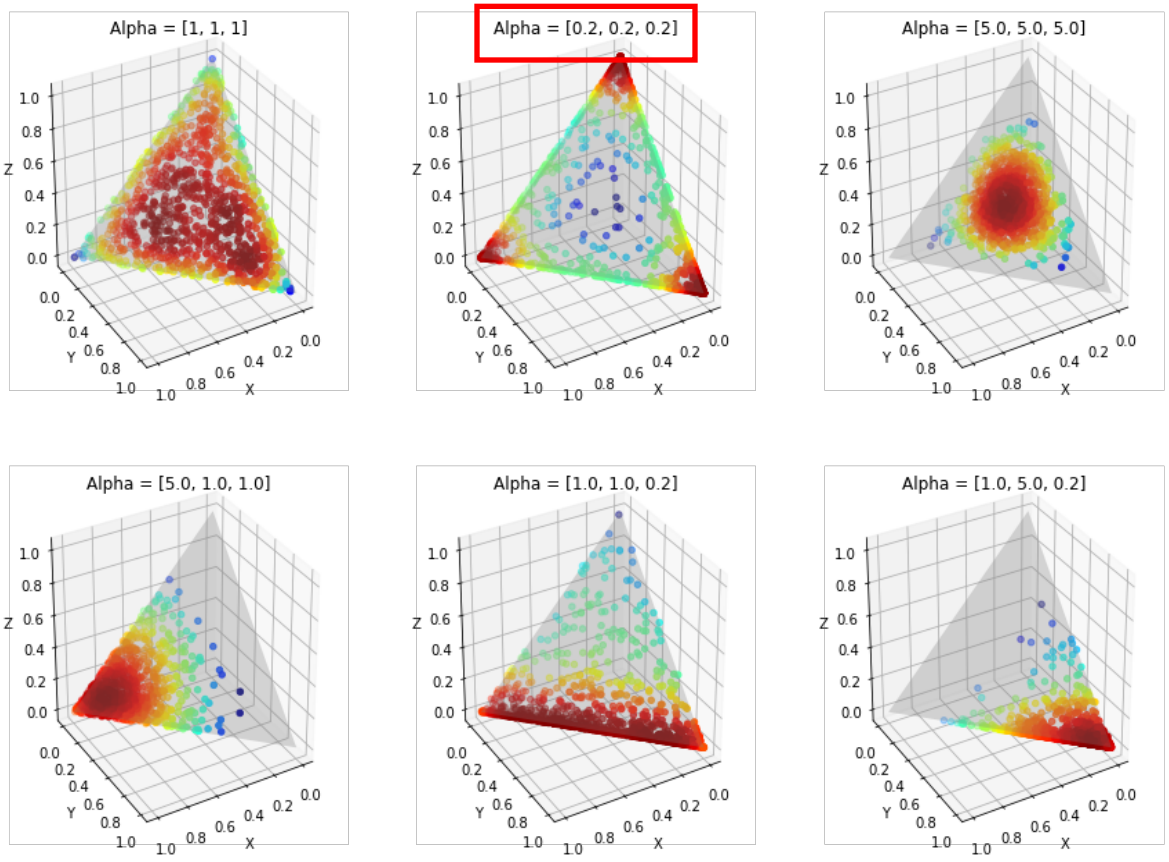
Generative process



https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158

The sparsity is important

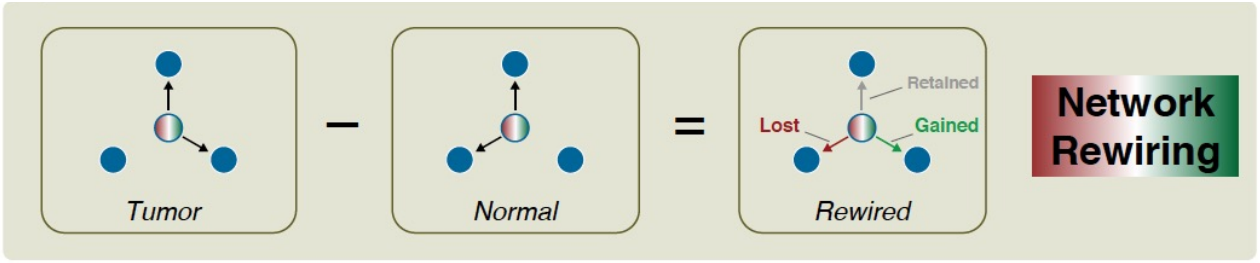
hyperparameter of Dirichlet distribution enable the sparsity of document to topic (**θ**) and word to topic (***φ***) distribution, make LDA works better than others similar methods most of time.



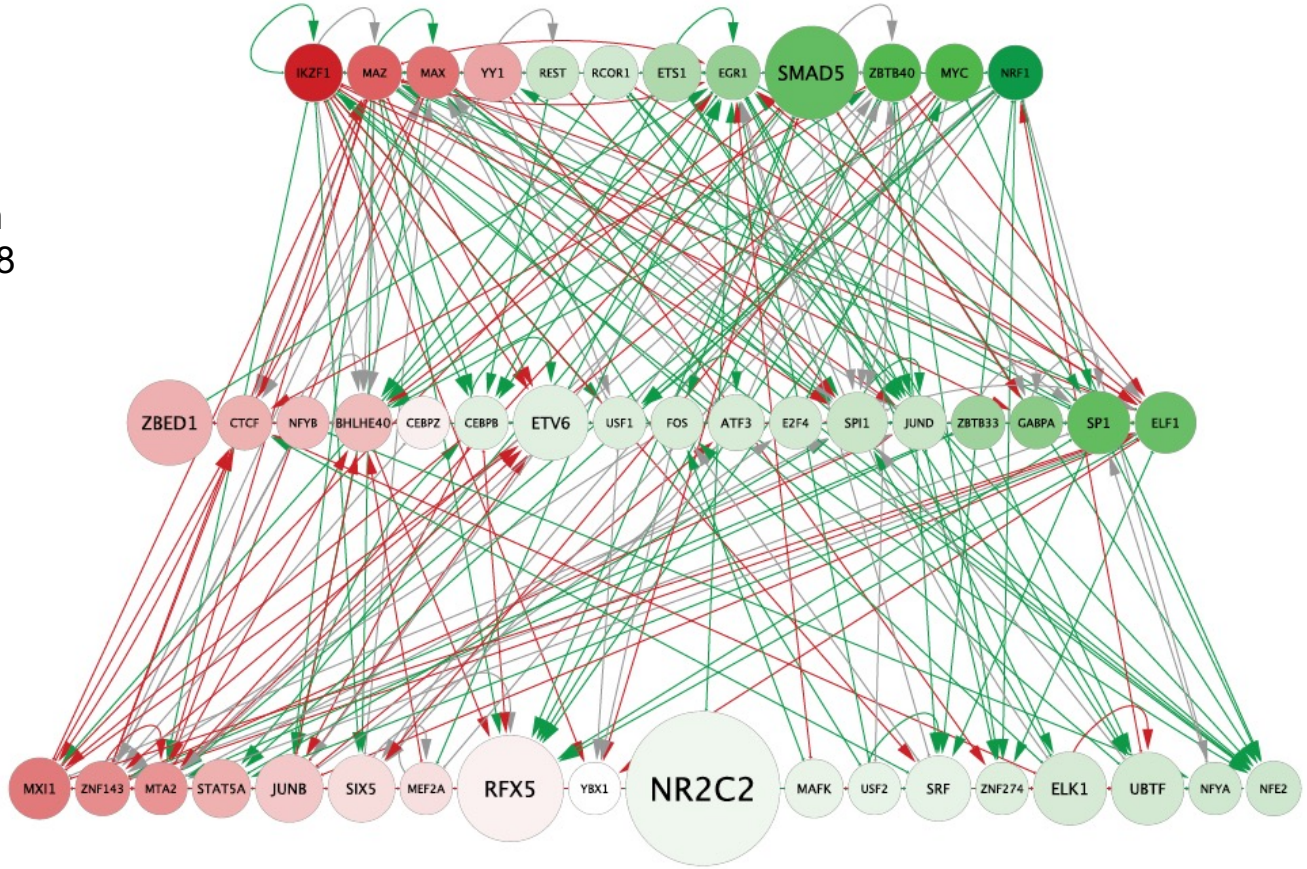
How the distribution of θ changes with different α values

In LDA analysis, alpha should be tuned for topic distribution(**θ**)

https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158



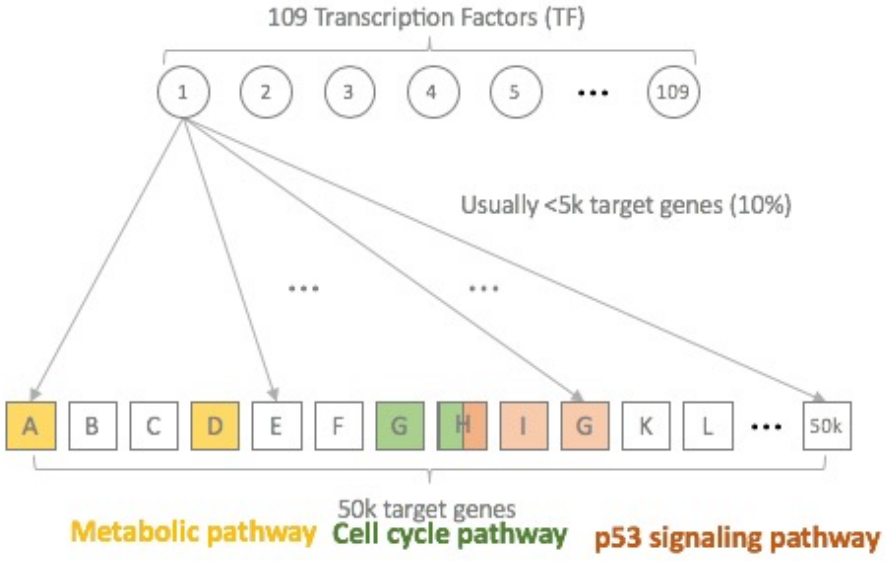
Rewired edges in comparison of GM12878 to K562 109 node TF-TF network (approx. CML)

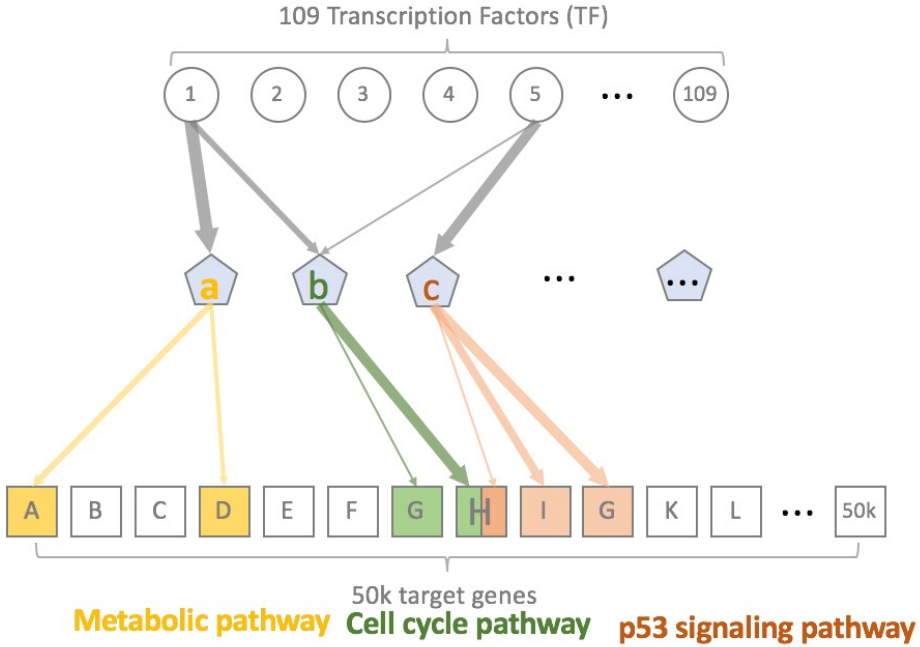


[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

From $TF \rightarrow gene$ (109×50,000)
to $TF \rightarrow pathway$ (109×50)

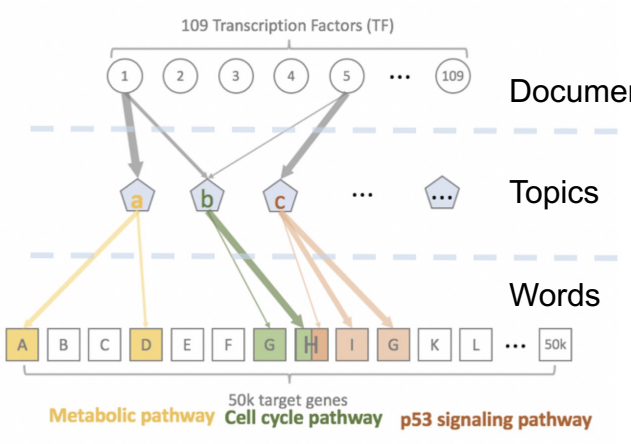
Hidden Layer (50 biological pathways?)

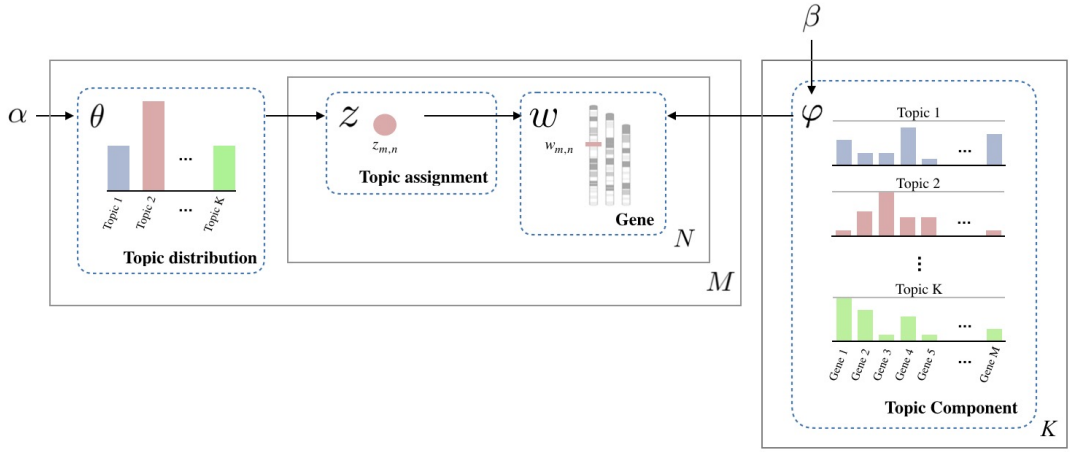




[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

TopicNet: Measuring transcriptional regulatory network change using LDA

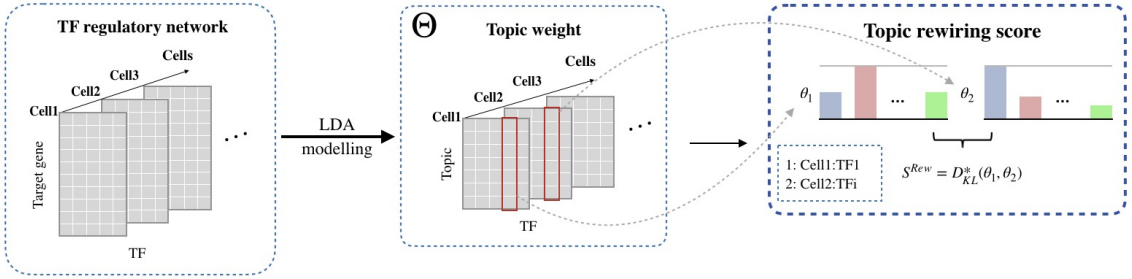




α Prior info β Prior info

θ: topic distribution per document

φ: word distribution per topic



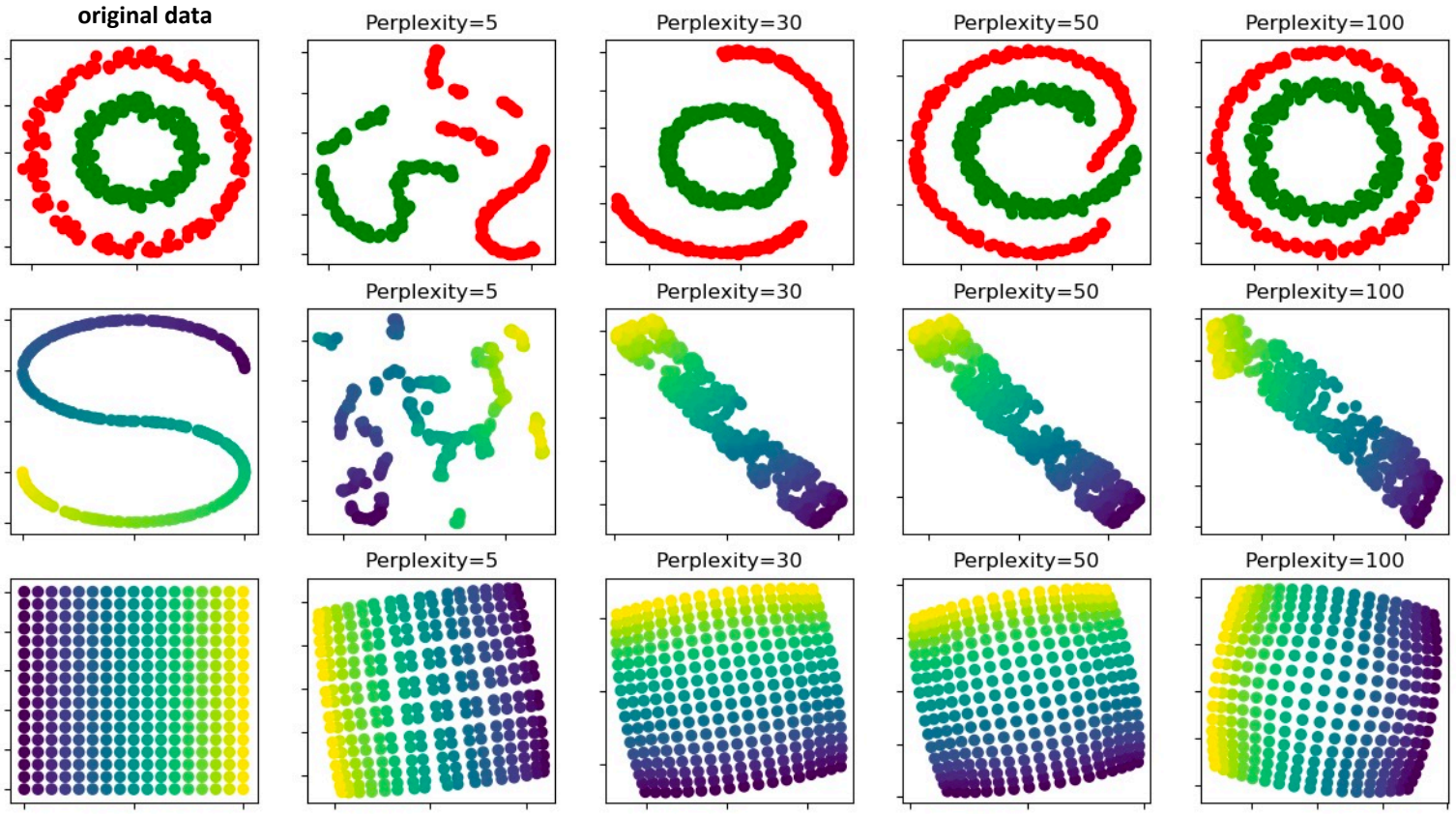
[Lou et al. bioxriv + Bioinformatics ('20)]

Unsupervised Mining

tSNE

tSNE

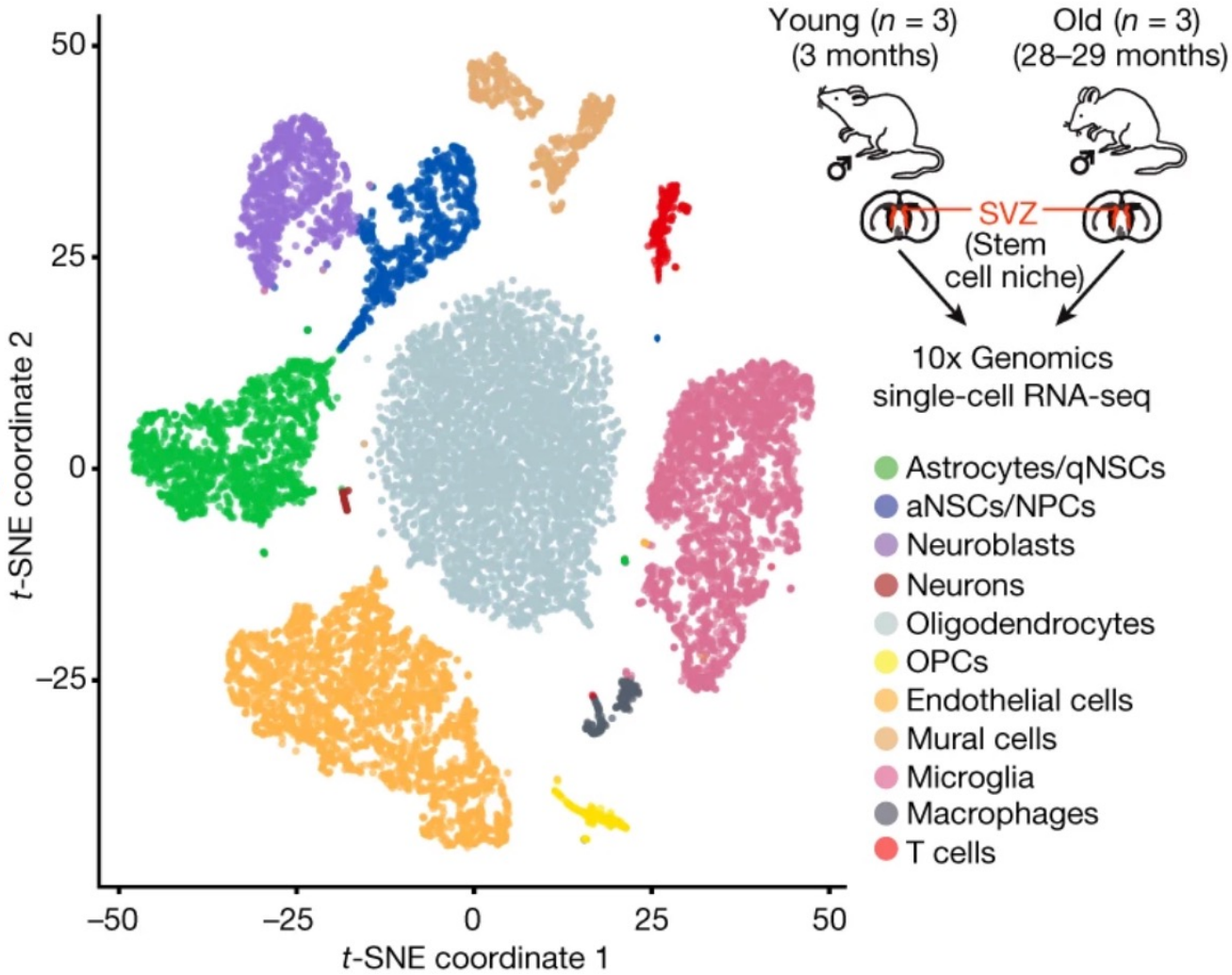
a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets



hyperparameters 'perplexity' really matter; Cluster sizes in a t-SNE plot mean nothing; Distances between clusters might not mean anything

https://distill.pub/2016/misread-tsne/
https://scikit-learn.org/stable/auto_examples/manifold/plot_t_sne_perplexity.html#sphx-glr-auto-examples-manifold-plot-t-sne-perplexity-py

Example: t-SNE clustering of **14,685** single-cell transcriptomes



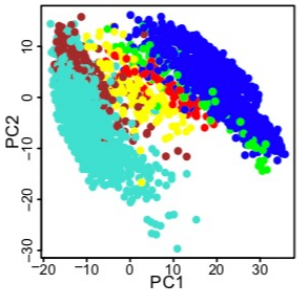
Dulken, B.W. et al. nature 2019

Unsupervised Mining

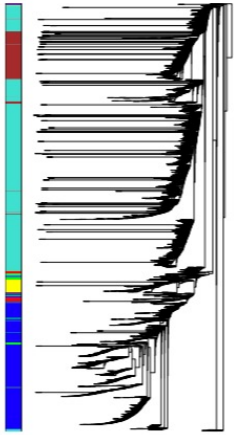
Comparisons

Comparison on real datasets (melanoma scRNA-seq dataset)

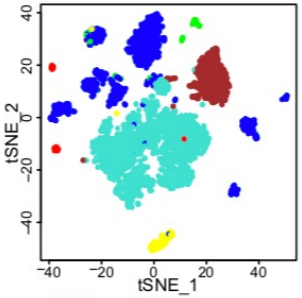
PCA



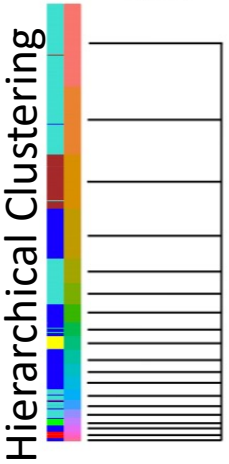
Hierarchical Clustering



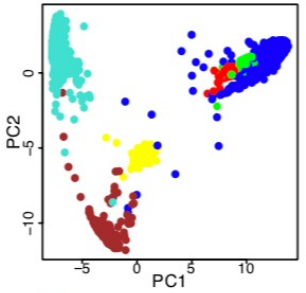
Seurat



Hierarchical Clustering

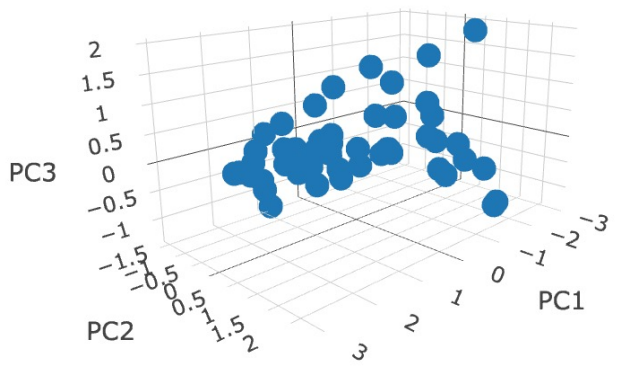


RCA



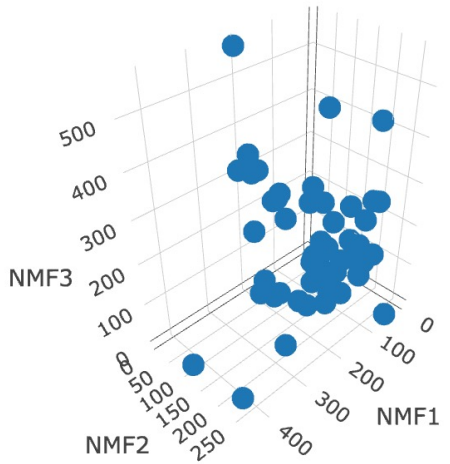
Hierarchical Clustering

Comparison of sparsity using **USArrests** dataset in a three-dimensional space



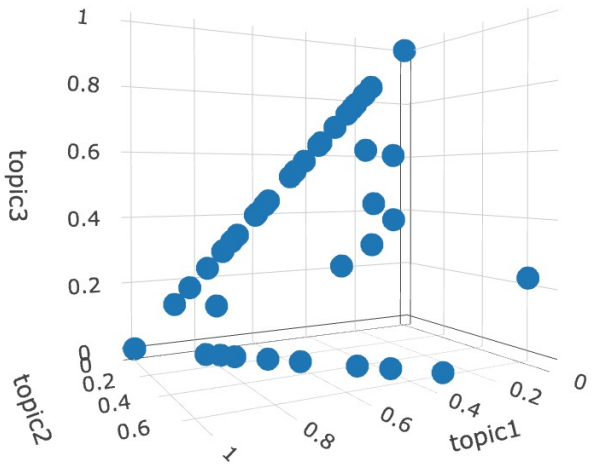
PCA

Rotation=XW(loading)



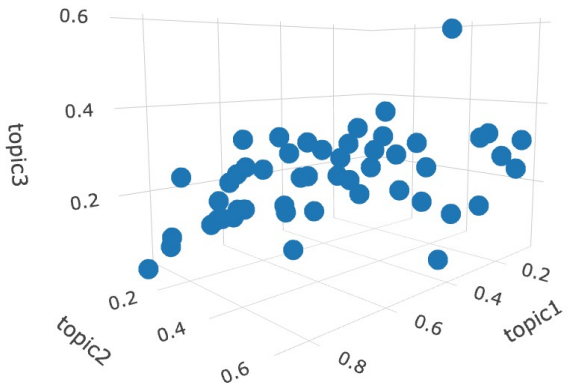
NMF (rank=3) X=W*H

W



LDA K=3, alpha=0.2

θ



LDA K=3, alpha=5

θ