

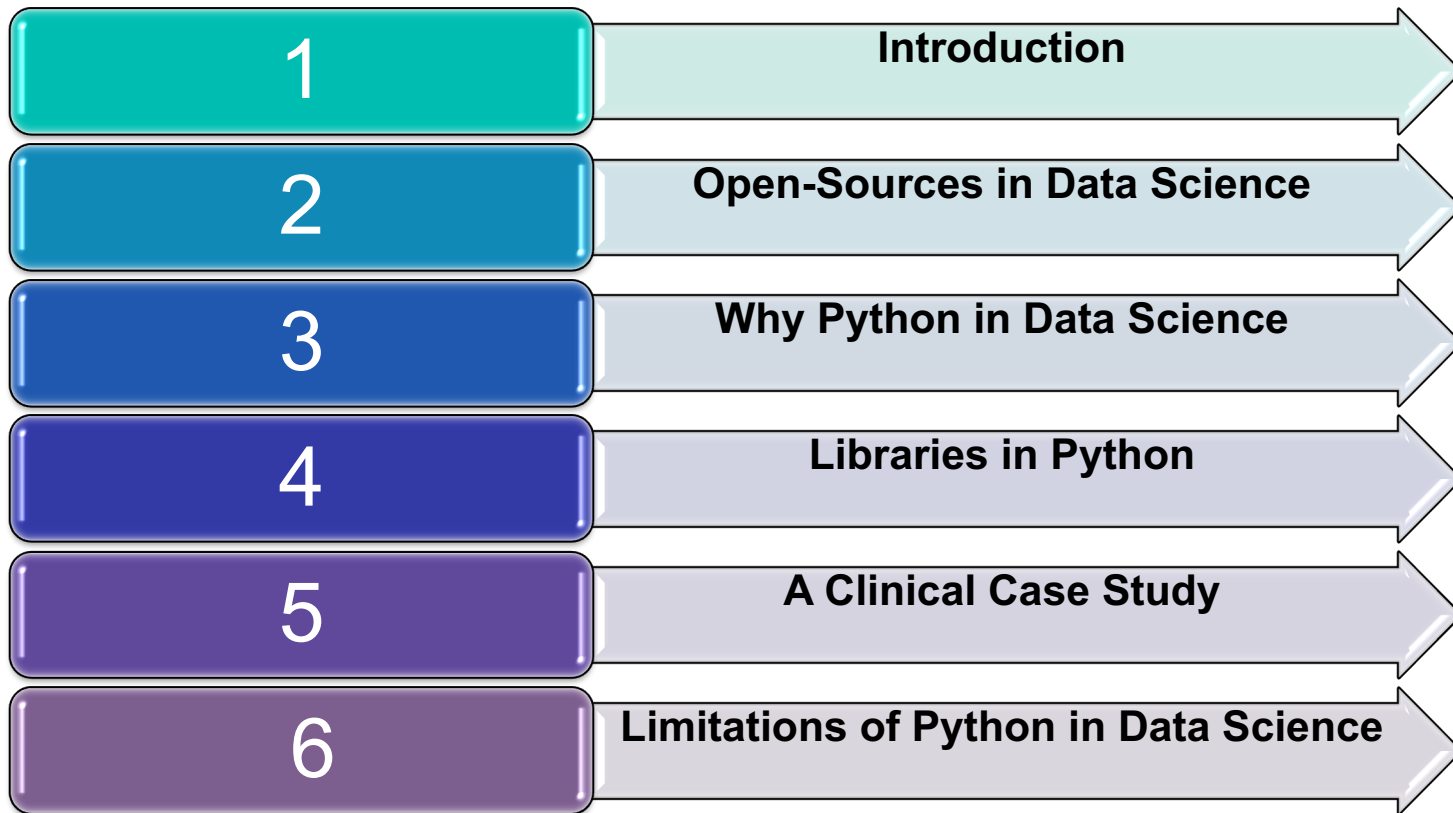


Single Day Event

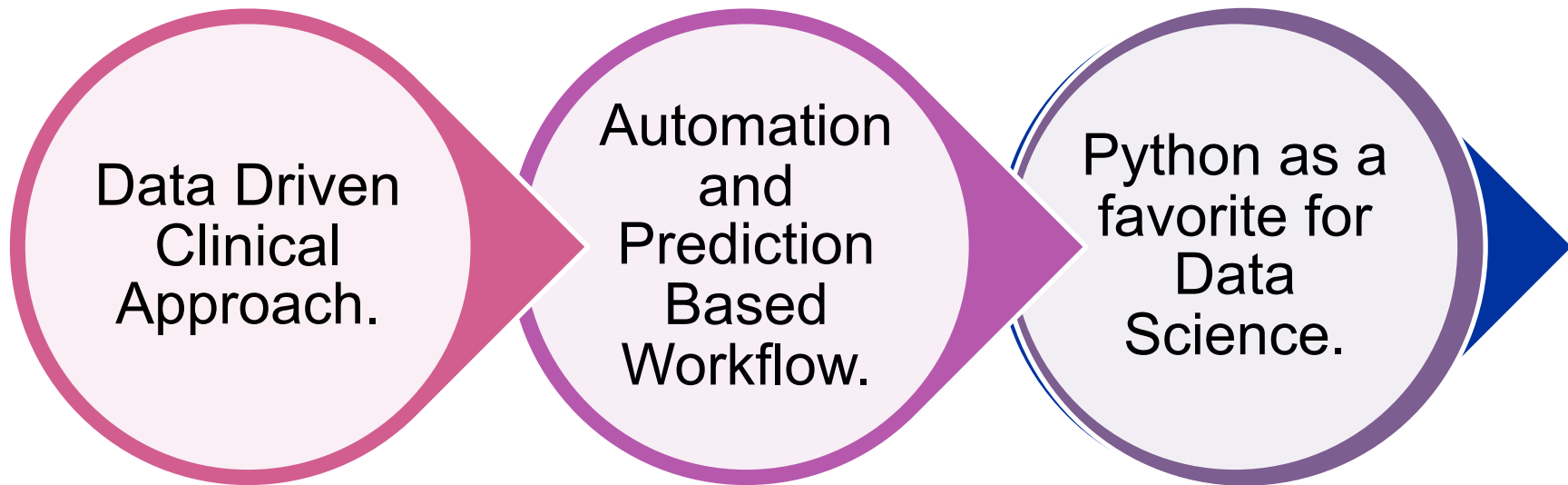
*Soumita Chel, Bayer
Pharmaceuticals, India*

Let's Explore Python In The World of
Clinical Data Science

Agenda



Introduction



Open-Sources in Data Science

Expansive Licensing



Development Environment



Source Code Availability



Data Mining and Big Data



Free Redistribution



Data Analysis and ML



Extensive Support



Data Visualization



kubernetes

Deployment



Open-Sources Languages in Data Science

'Write Once, Run Anywhere'

Scalable Application Development

Most Popular Data Science Language

Python

Programming Languages

Java

R

Interpreted Language

Huge statistical packages

High Computational Speed

Easy Code Conversion

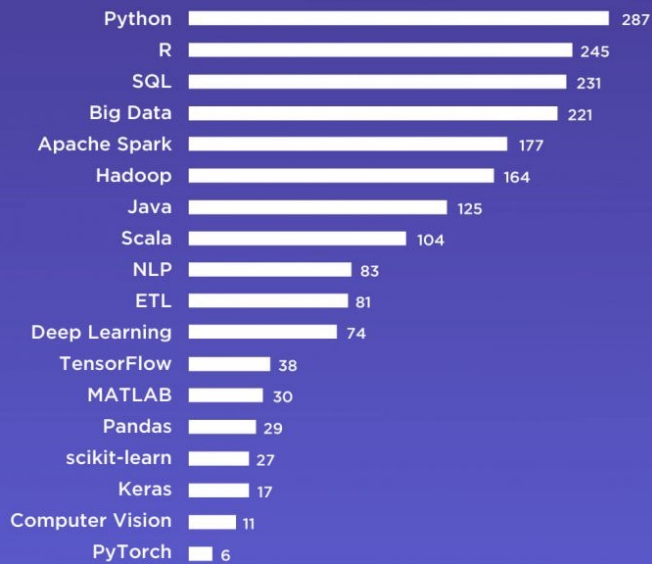
Julia



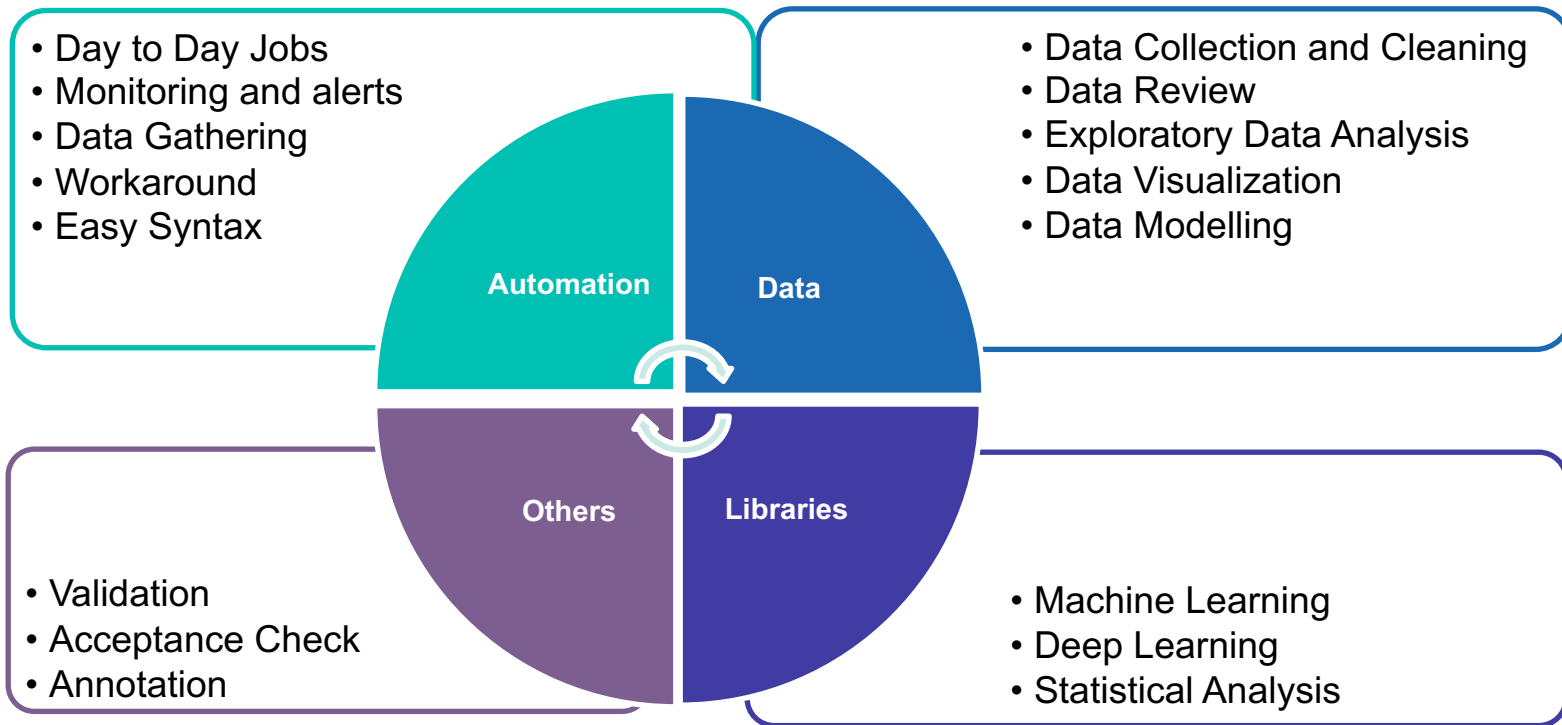
Popularity in Data Science : Where is Python?

The skills Data Scientists need today

(based on 300 job listings from tech companies in June 2019)



Clinical Context: Why Python?



Data Science Libraries of Python

 pandas

The pandas logo features a stylized vertical bar chart with four bars of increasing height, colored yellow, red, green, and blue from left to right.

spaCy



NumPy



matplotlib



TensorFlow

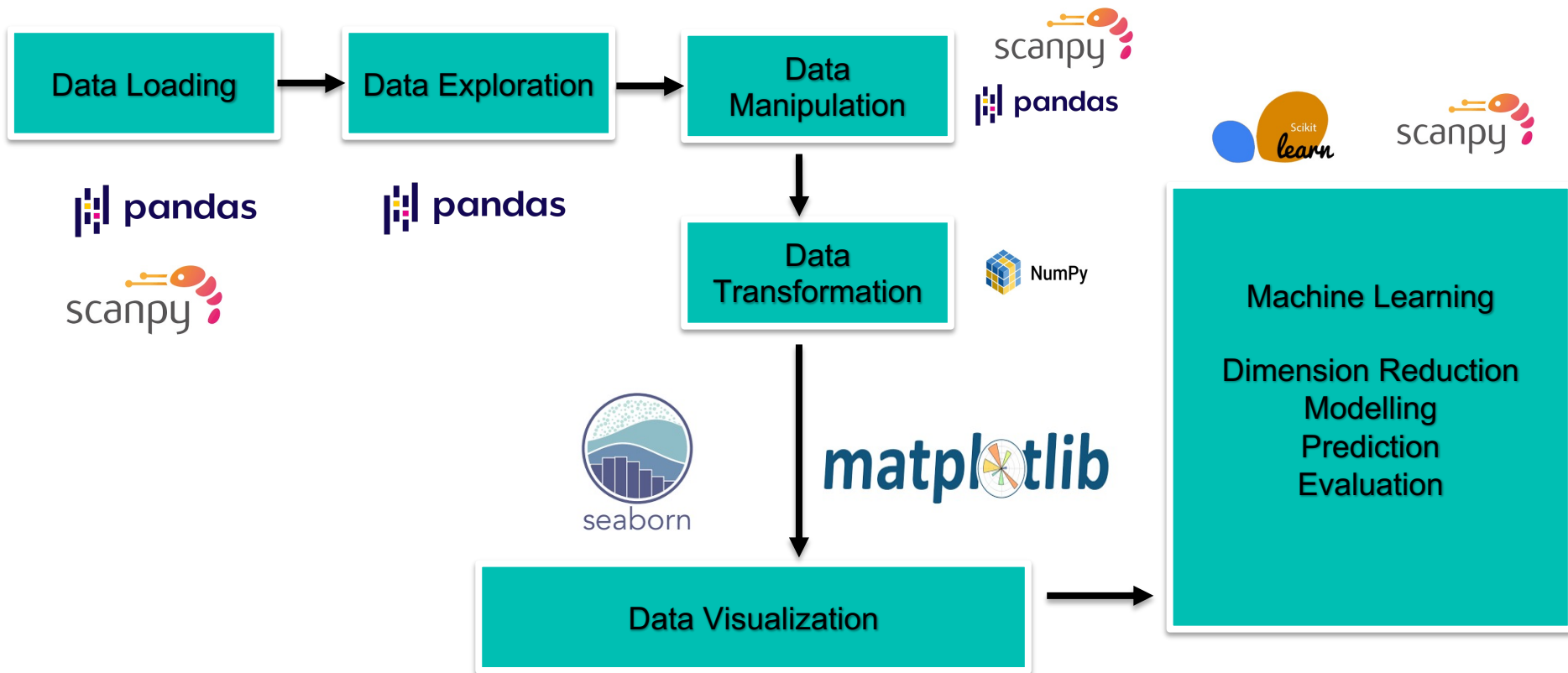


A Clinical Data Science Case Study

About the Use Case

Problem Statement: 'Clustering of Bronchoalveolar Cells of a Severe Covid Patient Using Gene Expression Data'

Workflow





Data Loading

```
1 # import numpy as np
2 import pandas as pd
3 import scanpy as sc
4 import anndata
5 import json
6 import os
7 import requests
```

Import Statements

```
1 adata = sc.read_10x_h5(filename)
2 adata

reading D:\Ongoing\IIT_Hyd\IIT COVID Project\Bronch_Nature_Paper1\Data\GSE145926_R1
x.h5

Variable names are not unique. To make them unique, call `.var_names_make_unique`.

(0:00:01)

Variable names are not unique. To make them unique, call `.var_names_make_unique`.

AnnData object with n_obs × n_vars = 18044 × 33539
var: 'gene_ids', 'feature_types', 'genome'

1 adata.var_names_make_unique()
```

Data Loading and Data Shape

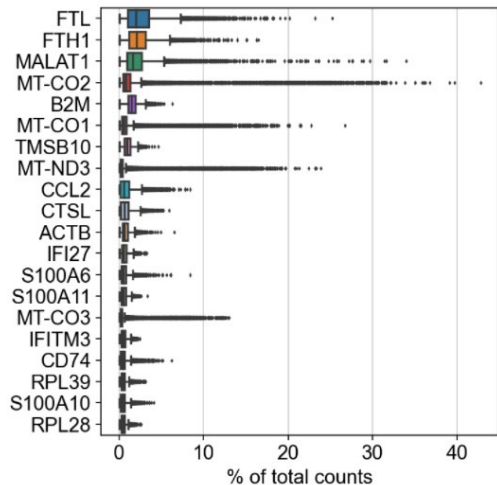
Row : Consists of Cells
Column: Consists of Genes

Data	Gene 1	Gene 2	Gene 3	Gene 4
Cell1	0	5	0	3
Cell 2	2	1	0	1

Data Matrix

[illegible]

Data Exploration



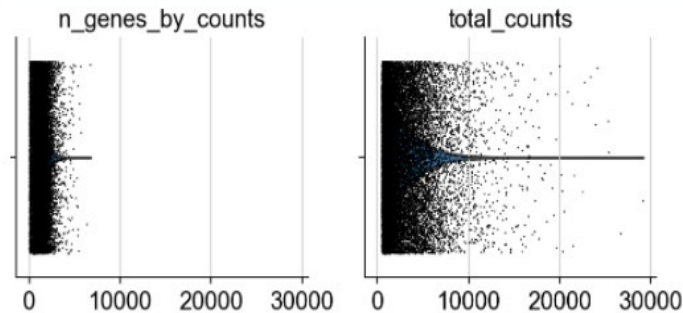
Top 20 genes

Top Genes and Their Cell Count

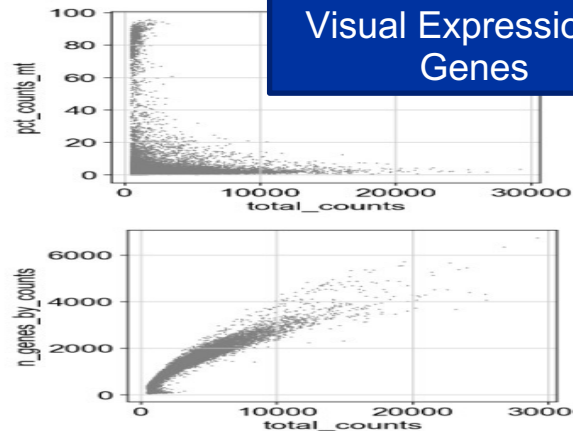
```
1 adata.var.sort_values(by=['n_cells'],ascending=False)
```

	gene_ids	feature_types	genome	n_cells
FTH1	ENSG000000167996	Gene Expression	GRCh38_nCoV	17934
FTH1	ENSG000000087086	Gene Expression	GRCh38_nCoV	17926
MALAT1	ENSG000000251562	Gene Expression	GRCh38_nCoV	17892
MT-CO2	ENSG000000198712	Gene Expression	GRCh38_nCoV	17874
B2M	ENSG000000166710	Gene Expression	GRCh38_nCoV	17866
...	...	...	...	...
AC021851.1	ENSG000000272800	Gene Expression	GRCh38_nCoV	3
AC022568.1	ENSG000000253474	Gene Expression	GRCh38_nCoV	3
MEF2C-AS2	ENSG000000245864	Gene Expression	GRCh38_nCoV	3
LINC01979	ENSG000000262585	Gene Expression	GRCh38_nCoV	3
AL021368.3	ENSG000000272541	Gene Expression	GRCh38_nCoV	3

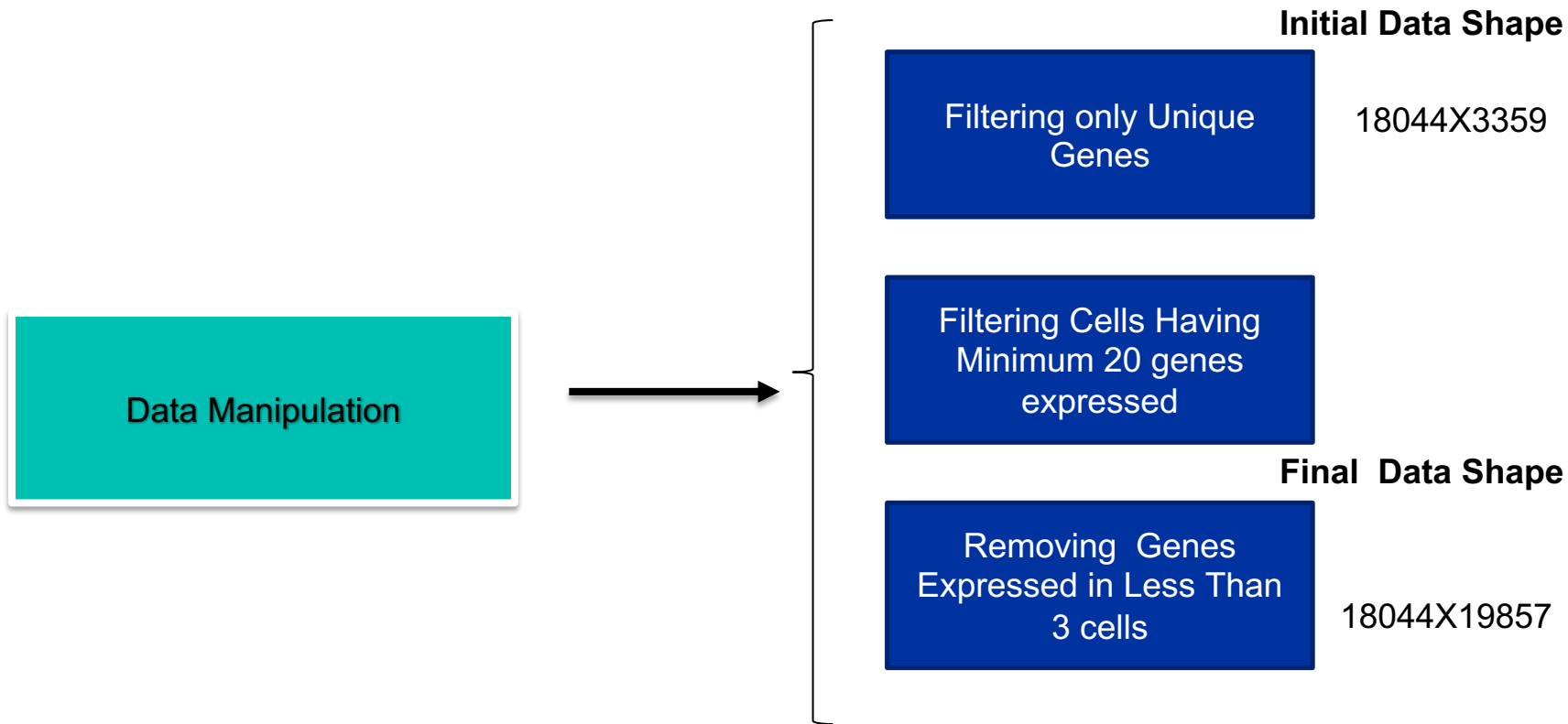
19857 rows × 4 columns



Visual Expression of Genes



Data Manipulation



Data Transformation

```
1 #Normalization
2 from sklearn import preprocessing
3 data=preprocessing.normalize(d, norm='l2', axis=1, copy=True)
```

Normalization of Data



Normalized Data Range

0.0

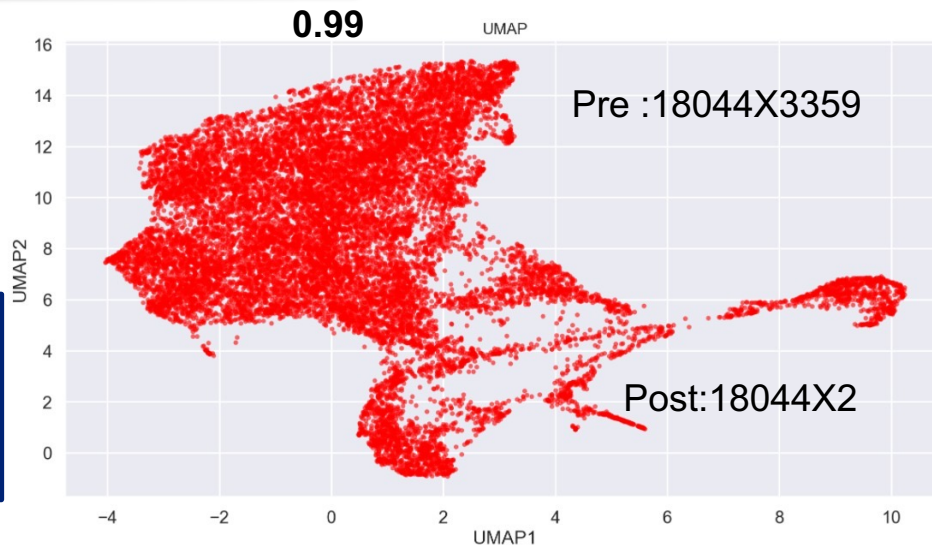


NumPy



umap-learn

Dimension Reduction Step For Easy
Computation and Visualization



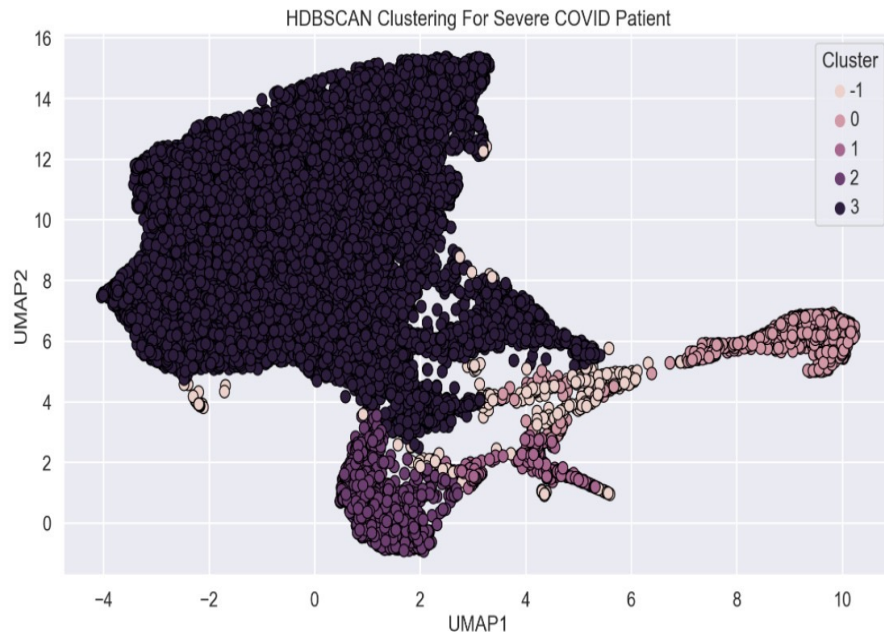
Machine Learning Approach

Aim: We want to group the cells to see if they can form clusters of similar characteristics

Algorithm: HDBSCAN
Package Name: hdbscan

Evaluation Metrics: Correlation , Inter Cluster Distance
Package Name: Numpy

Scatter Plot Visualization After Clustering



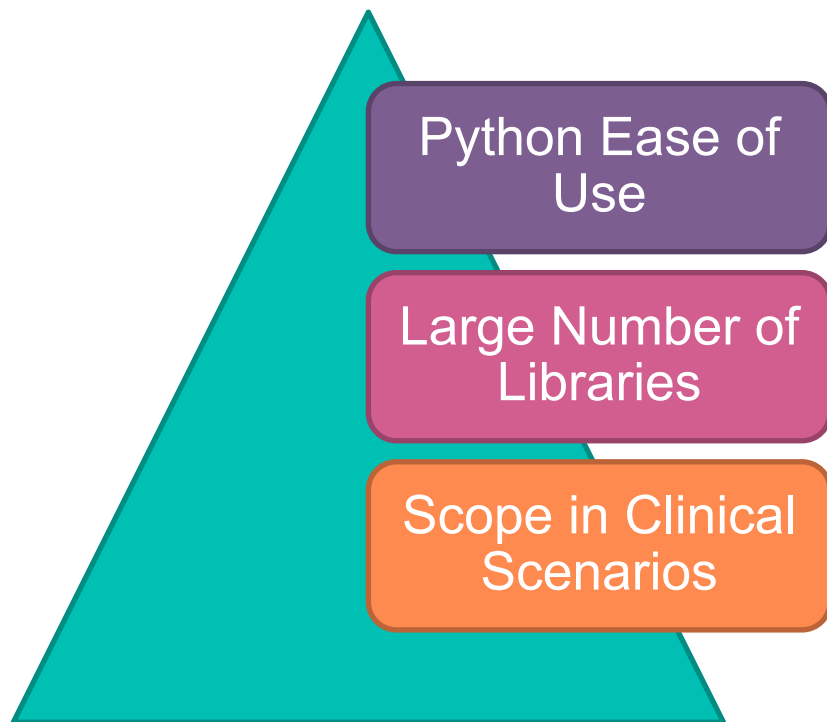
Some points to consider

Slow Computation

High Memory
Consumption

Problem With
Threading

Wrapping Up



References

1. Paper Link: <https://www.nature.com/articles/s41591-020-0901-9>
2. Data Link: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145926>
3. Python : python.org
4. Scanpy: <https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html>



Thank You



**The Global Healthcare
Data Science Community**

Contact Channels

📞 UK +44 1843 609600
✉ office@phuse.global
🌐 phuse.global

Social Media

🐦 [@phusetwitta](https://twitter.com/phusetwitta)
📘 [/phusebook](https://facebook.com/phusebook)
📺 [/phusetube](https://youtube.com/phusetube)
🌐 [/company/phuse](https://linkedin.com/company/phuse)



Q/A

Name: Soumita Chel
Email: soumita.chel@bayer.com