Ink Scrapers

Domingos Lopes

Tingting Chang

Overview

March 28th

Original Dataset

- *author_id (NO COPY RIGHT)*
- *painting_id*
- *art_movement*
- *painting_style*
- *painting_location*
- *medium*
- *painting_dates*
- *painting_title*
- *width*
- *height*
- *nationality*
- *bio_info*
- *first_name*
- *last_name*
- *nationality*
- *birth_year*
- *death_year*

Total # of rows: 207353
Data size: 31GB
** The Athenaeum*

Cleaning Data & EDA

Scraping	• scraping The Web Gallery of Art • scraping The Athenaeum
EDA	• paintings • authors • date • art movement
Data	• color Histogram HSV • PCA • Features from Neural Network
Kmeans	• Color per cluster • Clusters per author • Clusters per art movement

Prediction

Getting Error_rate

Several Algorithms:

- ***KNN***
- ***Extra Tree***
- ***Random Forest***
- ***Gradient Boosting***
- ***XGBoost***
- ***Naive Bayes***

First round of scraping

March 29th

The Web Gallery of Art is a searchable database of European fine arts and architecture (8th-19th centuries), currently containing over 42.400 reproductions. Artist biographies, commentaries, guided tours, period music, catalogue, free postcard and mobile services are provided.

ENTER HERE

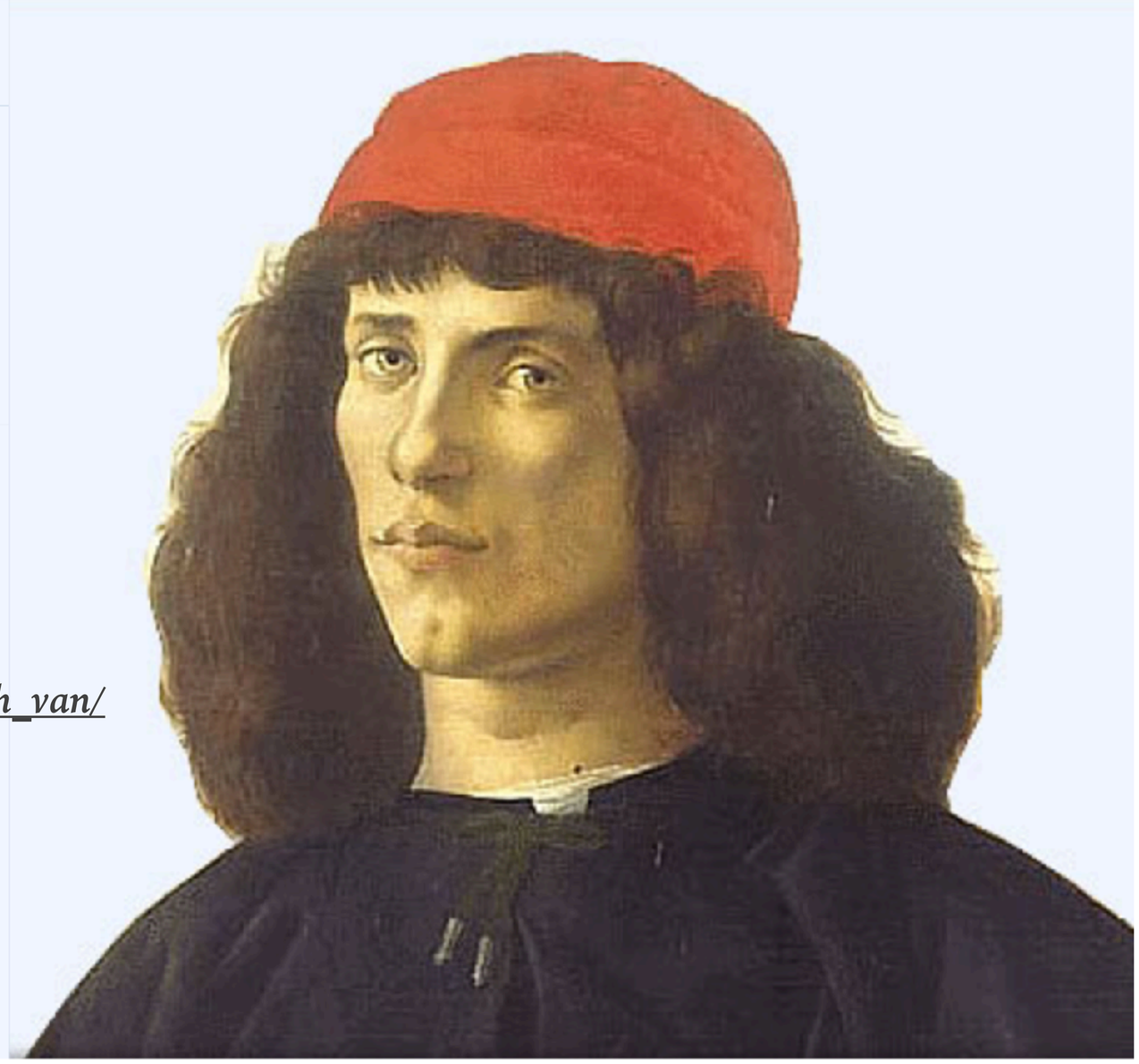
Mobile version

Gallery online since 1996

Total # of paintings: 2325

Data size: 396.3MB

* *http://www.wga.hu/html/g/gogh_van/*



Second round of scraping

March 29th

** http://www.the-athenaeum.org/people/detail.php?ID=789 detail.php?ID=789*

Basic EDA

March 29th

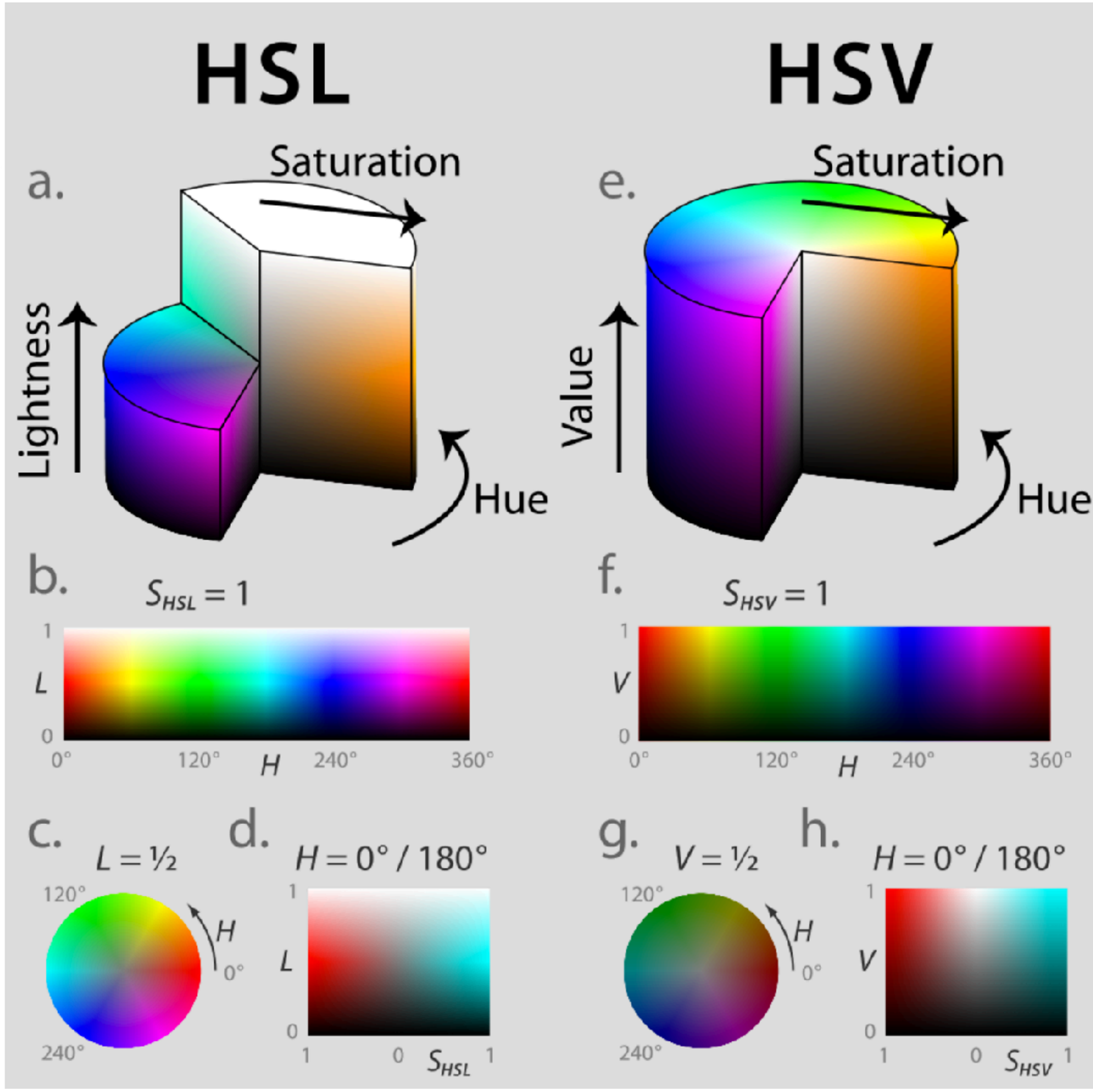
See EDA on Notebook

Color Histogram

- *Multithreading for generating color histogram*
- *Multithreading for resizing images*
- *Google cloud to speed up computation*

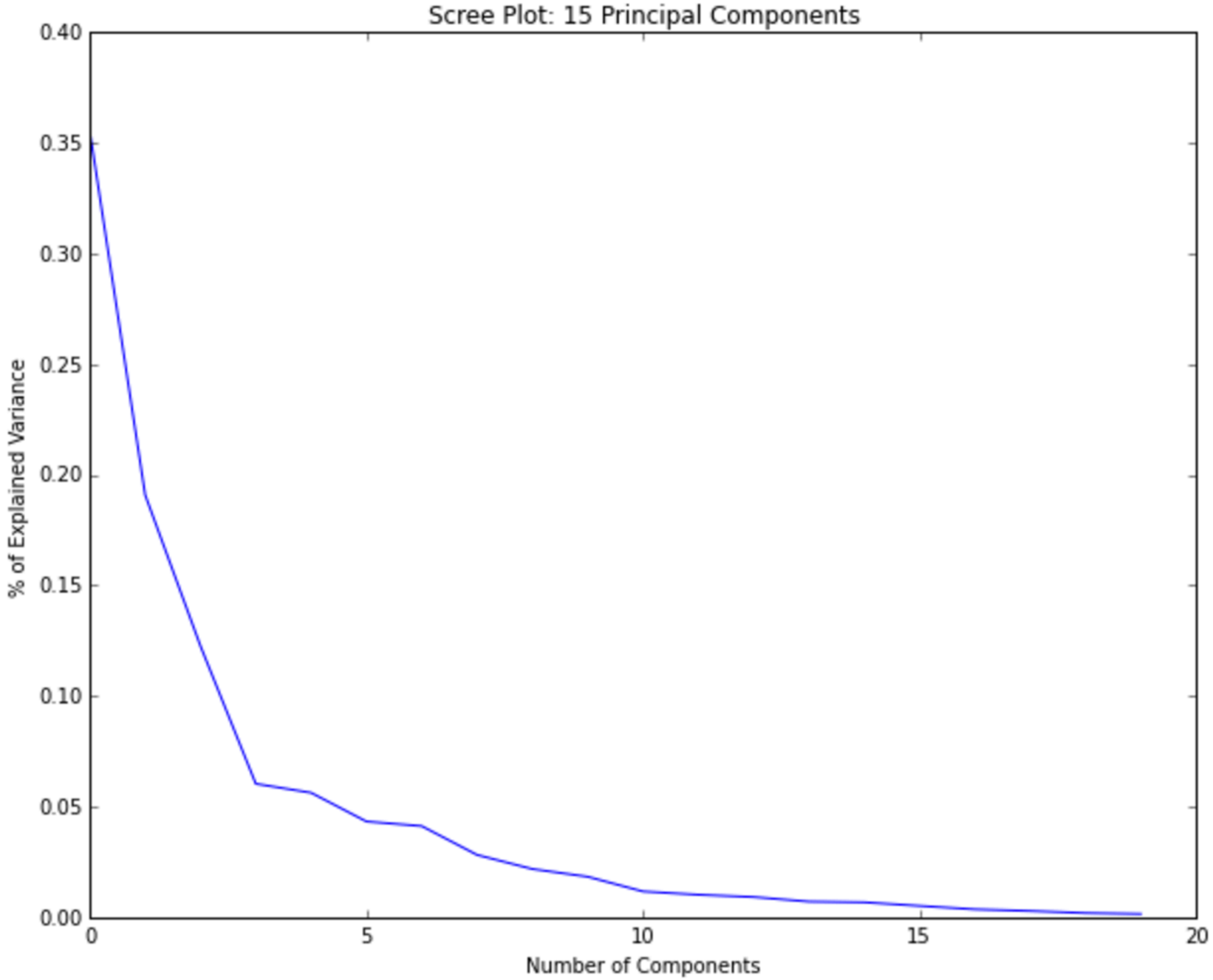
- *These saturated colors have lightness ½ in HSL, while in HSV they have value 1.*
- *hist_settings = [(0, 20, 0, 180), (1, 5, 0, 256), (2, 5, 0, 256)]*

hist_26	hist_27	hist_28	hist_29	hist_30	height_width_ratio
0.105572	0.219430	0.264793	0.230477	0.179727	0.693750
0.035917	0.089287	0.095728	0.098622	0.680446	0.779412
0.114563	0.189451	0.218488	0.254685	0.222813	1.310585



```
Out[82]: 0.997030725370565 14
```

PCA



Predictions

March 28th

predict author

model_name	color_hist	pca
naive_bayes	0.599407	0.623145
extra_tree	0.673591	0.660732
random_forest	0.686449	0.668645
knn	0.705839	0.686449
grad_boost	0.747774	0.729970
xgboost	0.794409	0.713155

predict art movement

	model_name	color_hist
0	random_forest	0.653738
1	xgb	0.676024

More things about models:

- *stacking*
- *selected features from Neural Network*

Model comparing

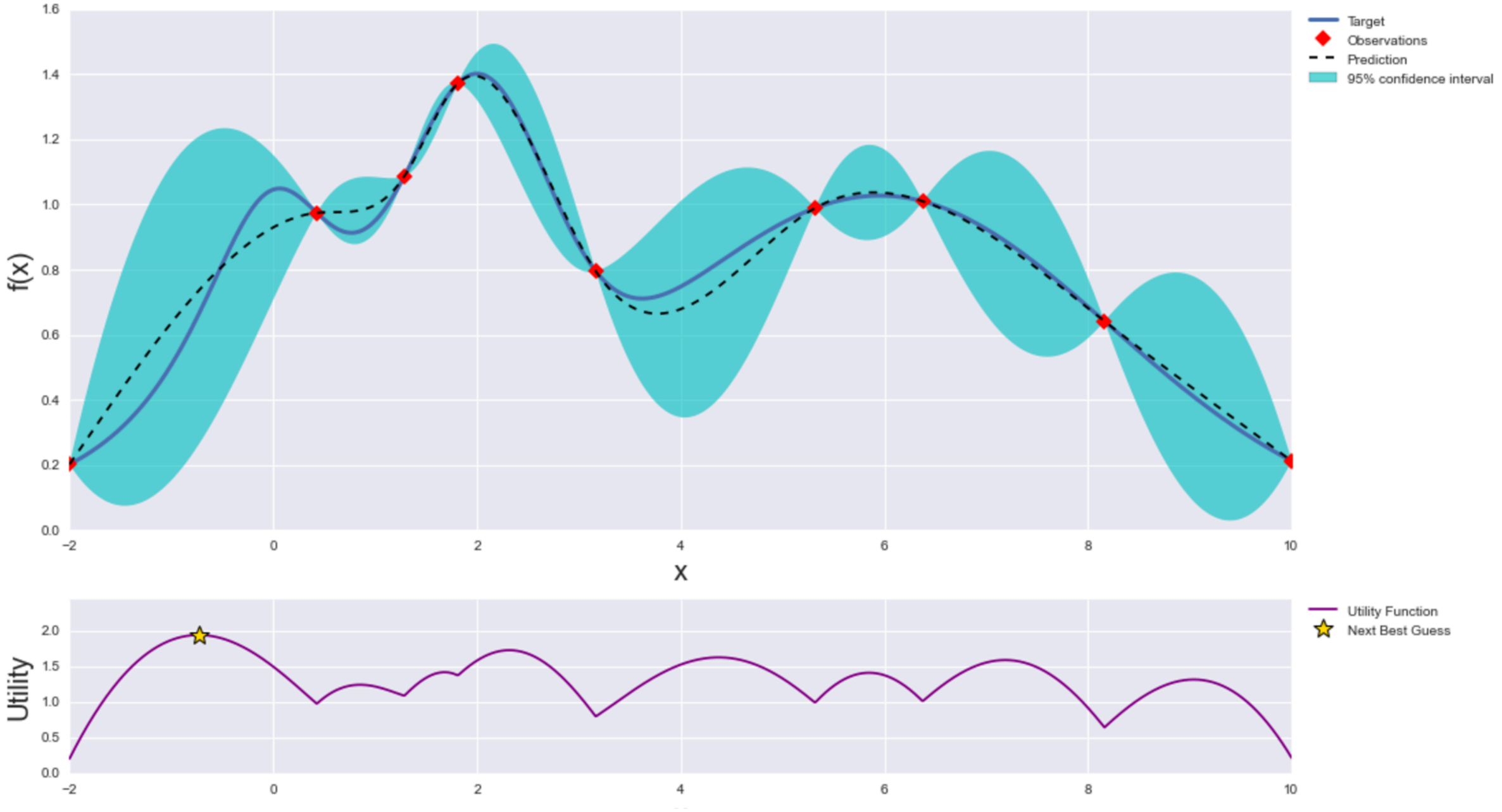
March 28th

xgb	xgb_pca	nbc_res	nbc_pca_res	knn_res	rf_res	rf_pca_res	extra_tree	extra_tree_pca_res	extra_tree_pca_res	grad_boost	grad_boost_
True	True	True	True	True	True	True	True	True	True	True	True
True	True	True	True	True	True	True	True	True	True	True	True
True	True	True	True	True	True	True	True	True	True	True	True
True	True	True	True	True	True	True	True	True	True	True	False
True	True	True	True	True	True	True	True	True	True	True	True
True	True	True	True	True	True	True	True	False	False	True	True
True	True	True	True	True	True	True	True	False	False	True	True
True	True	True	True	True	True	True	True	False	False	True	True
True	True	True	True	True	True	True	True	False	False	True	False
True	True	True	True	True	True	True	True	False	False	True	True

https://github.com/fmfn/BayesianOptimization

https://scikit-optimize.github.io/optimizer/index.html

Gaussian Process and Utility Function After 9 Steps



BayesianOptimization

Initialization

Step	Time	Value	max_depth	max_features	n_estimators
1	00m04s	0.56358	2.6284	6.4277	530.4468
2	00m06s	0.62865	5.9855	6.9211	558.7029
3	00m05s	0.58832	3.5851	5.7812	594.1779
4	00m06s	0.59401	3.9791	5.6445	588.9019
5	00m04s	0.63013	5.9280	5.7369	322.3853

Bayesian Optimization

Step	Time	Value	max_depth	max_features	n_estimators
6	00m19s	0.66452	10.0000	7.0000	100.0000
7	00m22s	0.66576	10.0000	7.0000	720.0000
8	00m11s	0.66329	9.7643	6.9618	206.3093
9	00m14s	0.66230	9.9508	6.9063	402.5351
10	00m08s	0.56408	2.0179	5.5760	150.1354
11	00m12s	0.66477	9.9179	6.6891	260.0988
12	00m18s	0.67219	9.9510	6.8468	674.1693
13	00m14s	0.66502	9.9072	6.8666	366.7500
14	00m15s	0.66749	9.9940	6.9664	447.9864
15	00m11s	0.66997	9.9504	5.1304	231.8717
16	00m18s	0.66576	9.9891	5.2606	697.9774
17	00m19s	0.66205	9.9973	6.6935	645.5310
18	00m14s	0.66452	9.9927	6.9834	293.4353
19	00m17s	0.66873	9.9906	6.7586	427.7723
20	00m15s	0.66799	9.9790	6.9790	340.8515
21	00m14s	0.66180	9.9828	6.9739	240.0485
22	00m17s	0.66947	10.0000	5.0000	470.7071
23	00m17s	0.66526	10.0000	5.0000	276.1377
24	00m21s	0.66056	9.9358	5.0326	663.2862
25	00m23s	0.66650	10.0000	7.0000	488.8521

TSNE VS PCA

March 5th

TSNE

embed high-dimensional data into low dimensions,

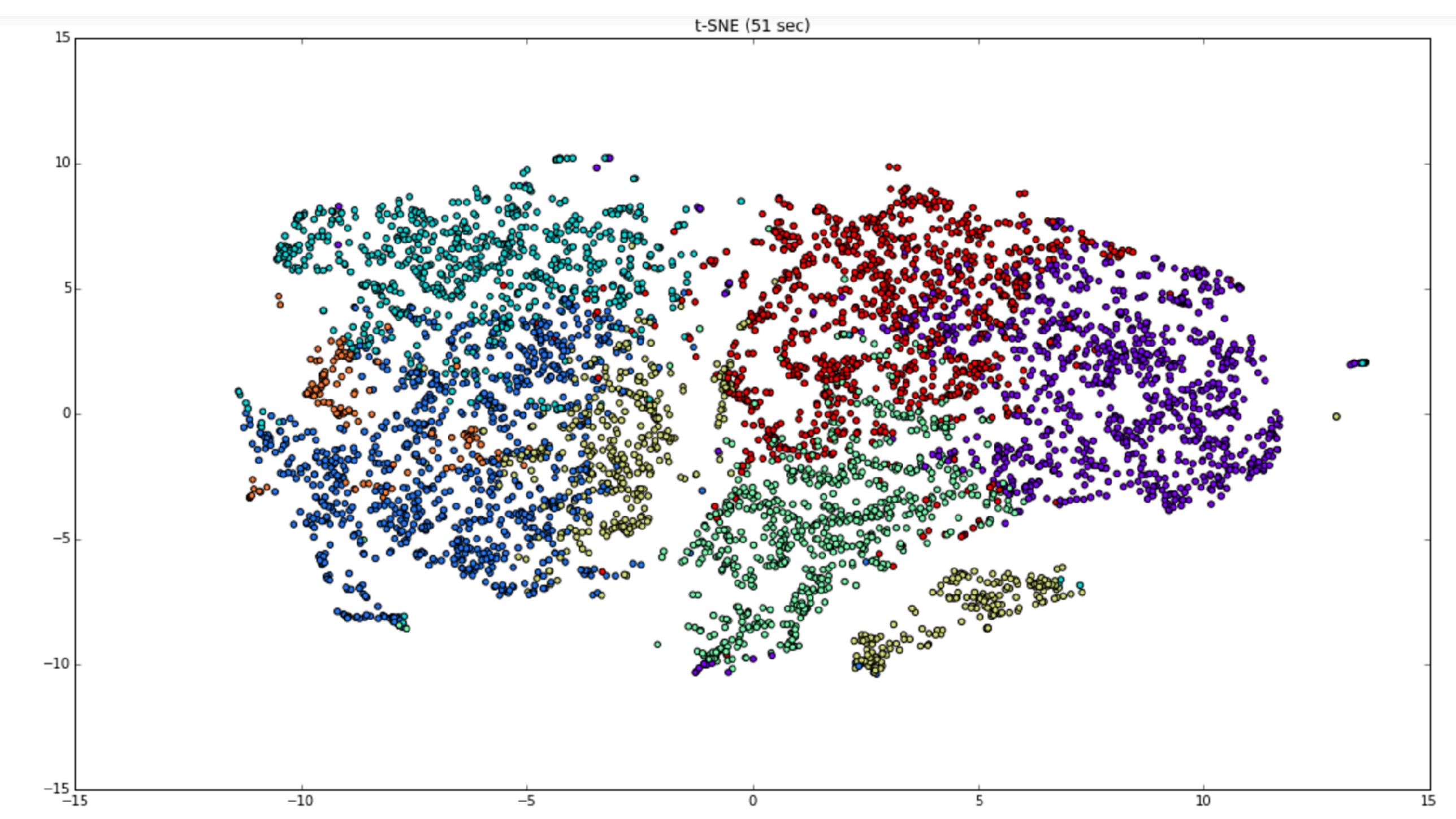
t-distributed stochastic neighbor

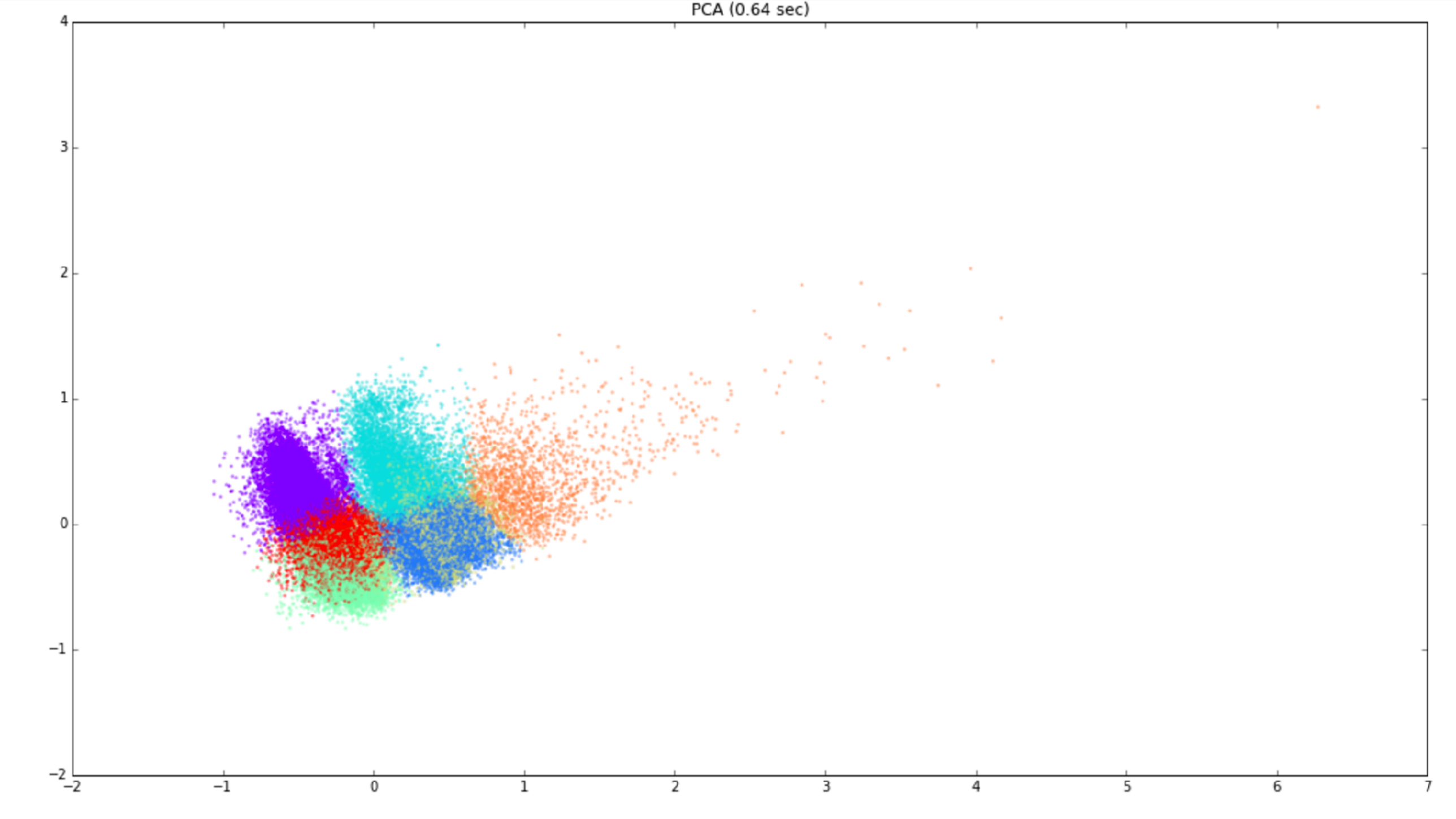
find an estimate on the number of expected clusters

PCA

embed high-dimensional data into low dimensions,

linear and parametric method.

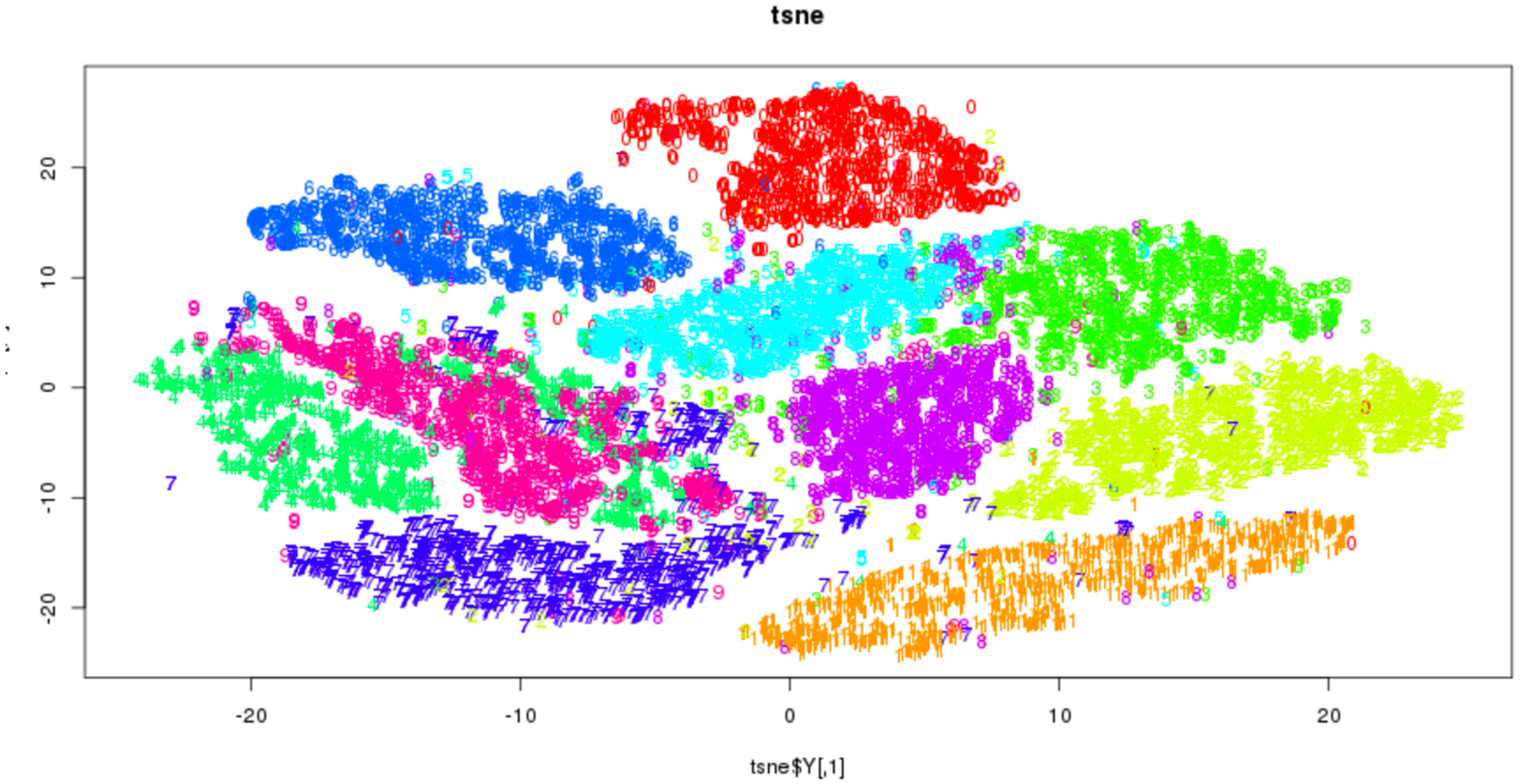

t-SNE (51 sec)
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15


PCA (0.64 sec)
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4
5
6
7

TSNE

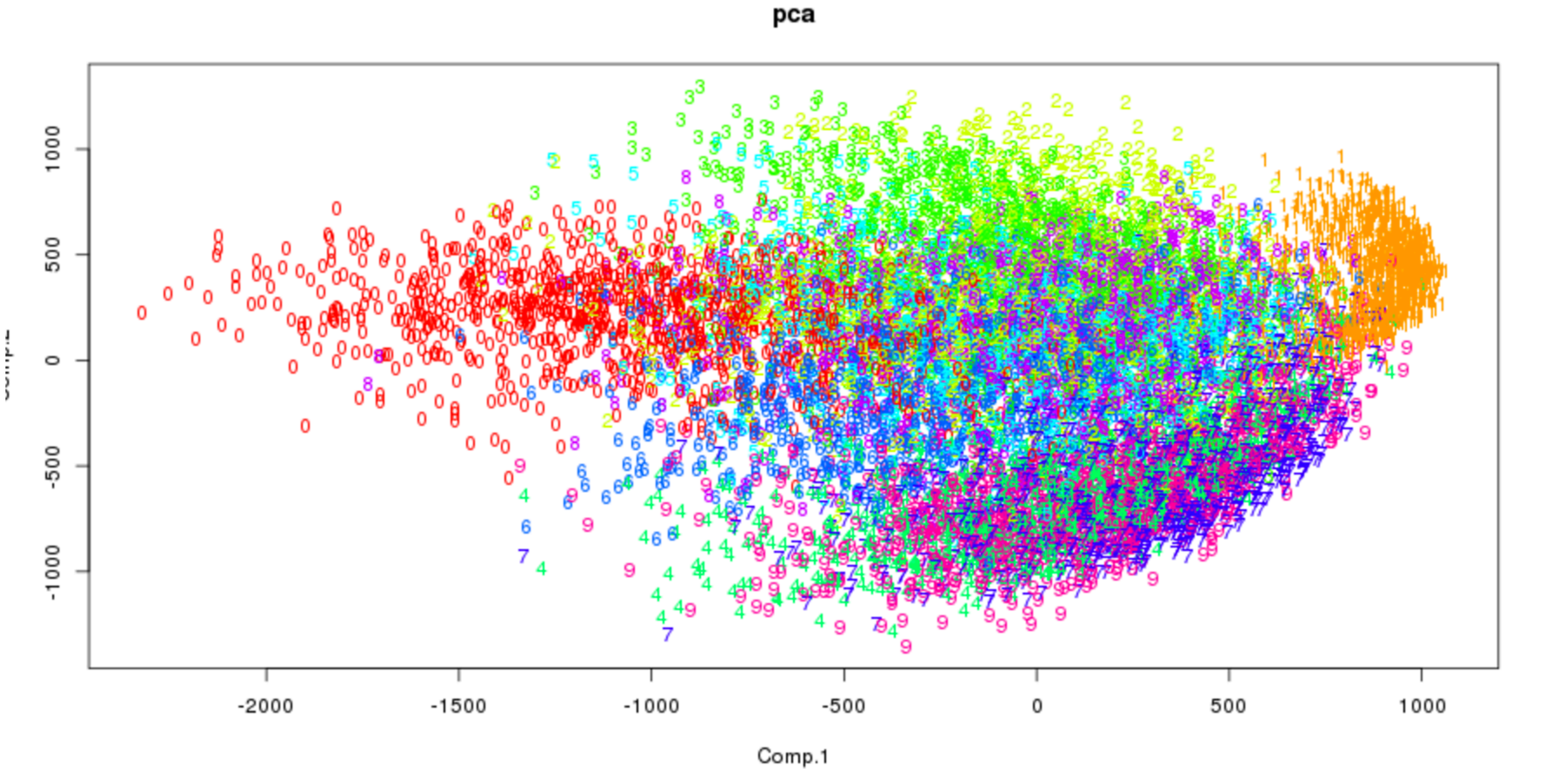
https://www.kaggle.com/puyokw/digit-recognizer/clustering-in-2-dimension-using-tsne/code

March 29th



PCA

March 29th



More Kmeans on Notebook

Convolutional Neural Networks

March 29th

Well suited for image data:

- *Color histograms contain no spatial information*
- *Other potential distinguishing features, for classification purposes, are very hard to detect*
- *Can, in theory, approximate any function: this means, with an adequate design, has the potential to solve any problem*

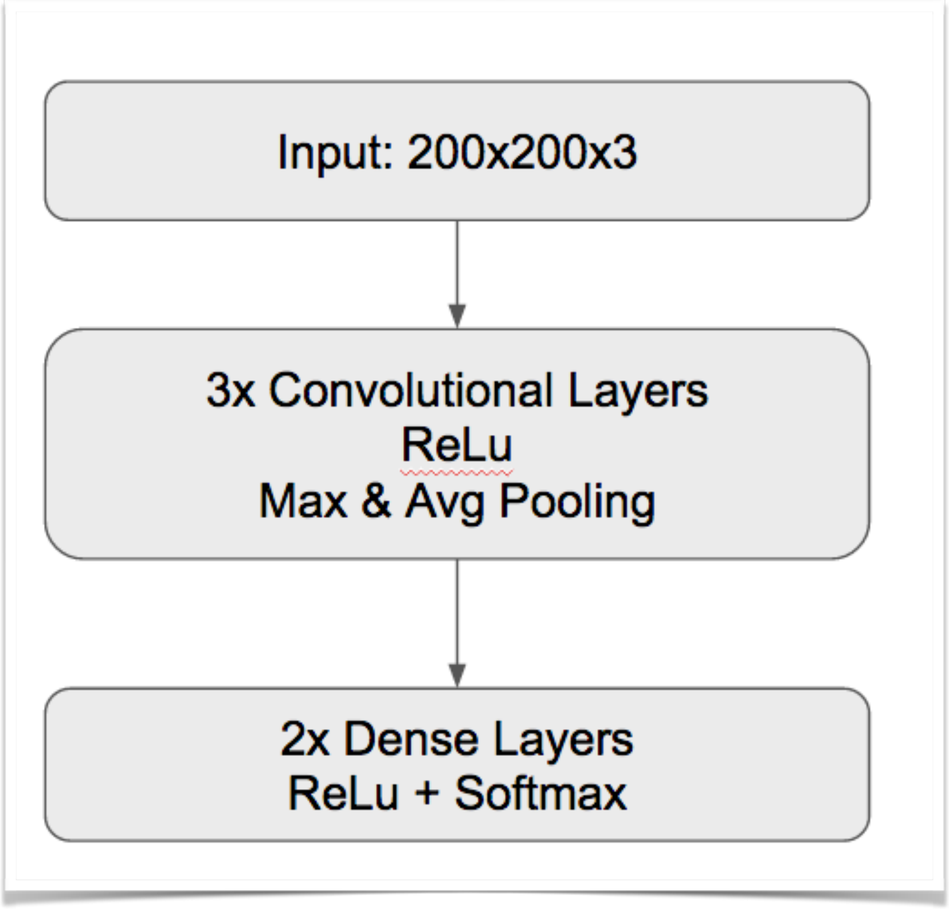
However:

- *Hard to train*
- *Hard to tune*
- *May suffer from vanishing gradients*
- *Demand a different combination of hardware resources*

Our network - Initial Configuration

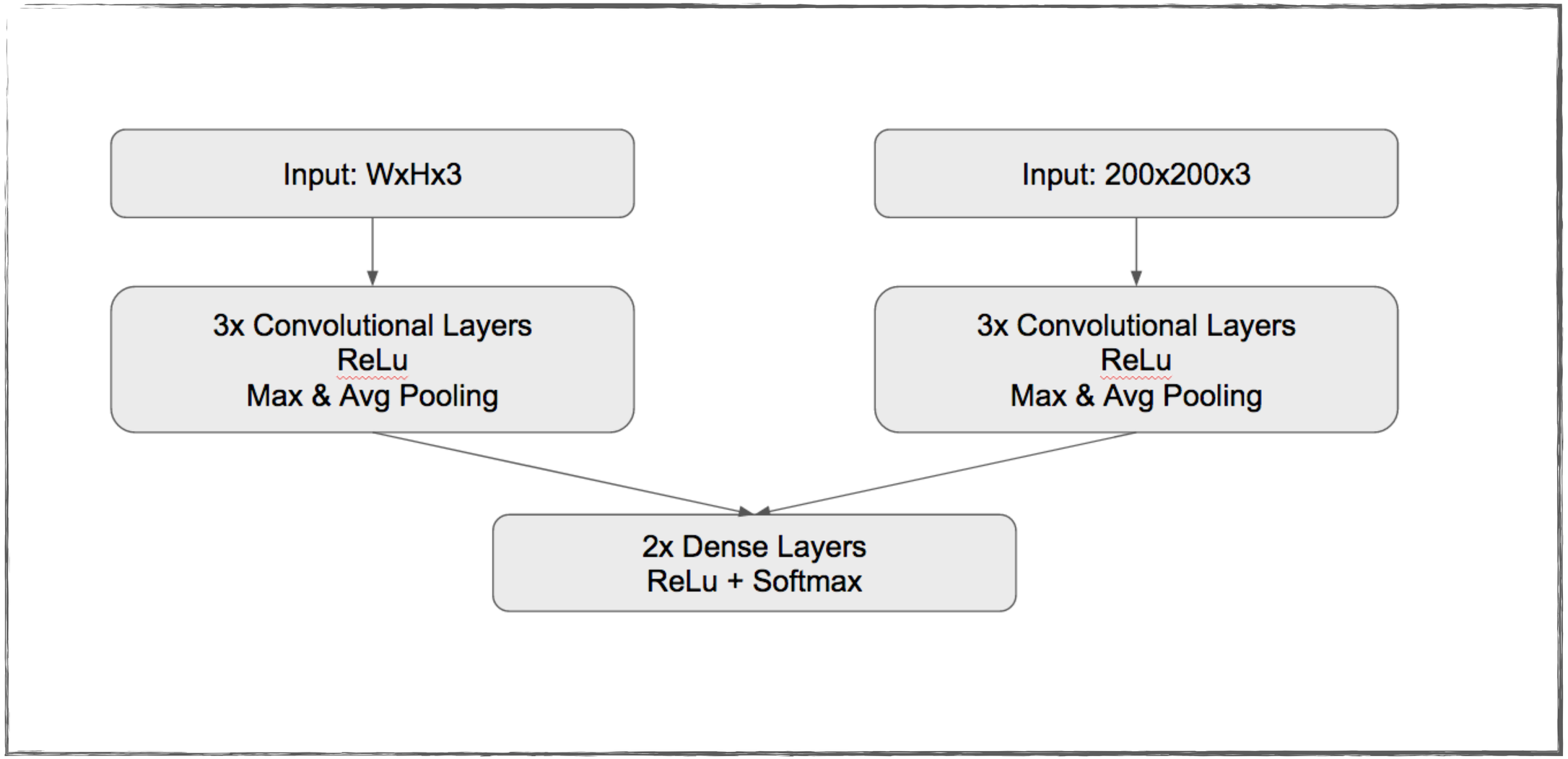
March 29th

- *Resized all paintings to 200x200*
- *Used batch generation*
- *Also included the height / width ratio*



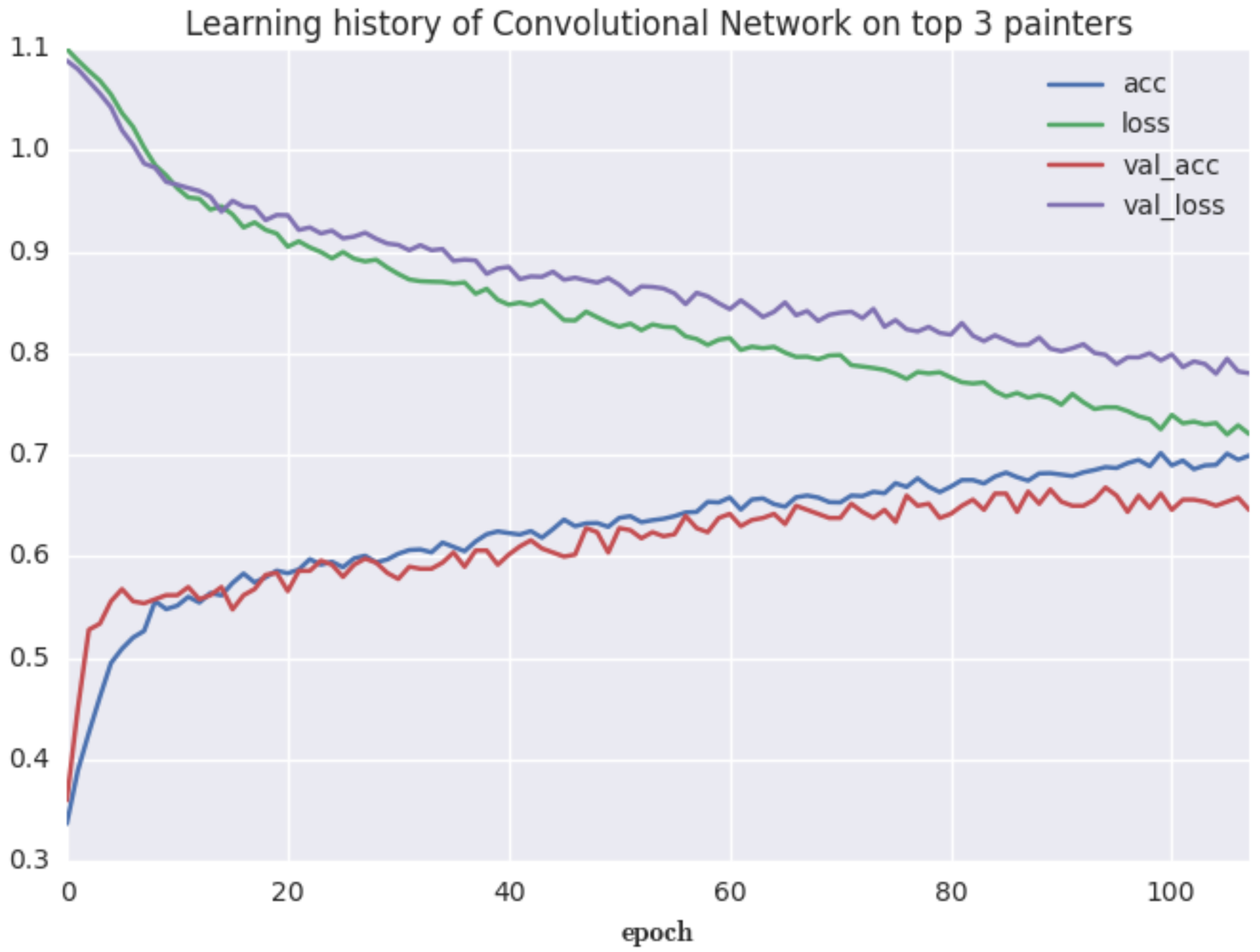
Our network - Current Configuration

March 29th



Training results

March 29th



Google Cloud Computing

March 29th

- *Gives access to high computing power*
- *Allows the users to scale resources according to their needs*
- *Makes it easy to share data between different instances*
- *GPU support still in beta*
- *Free trial credit*

Future Work

March 29th

- *Stacking model*
- *Predict the year of paintings*
- *Emulate painting style*

Thanks~