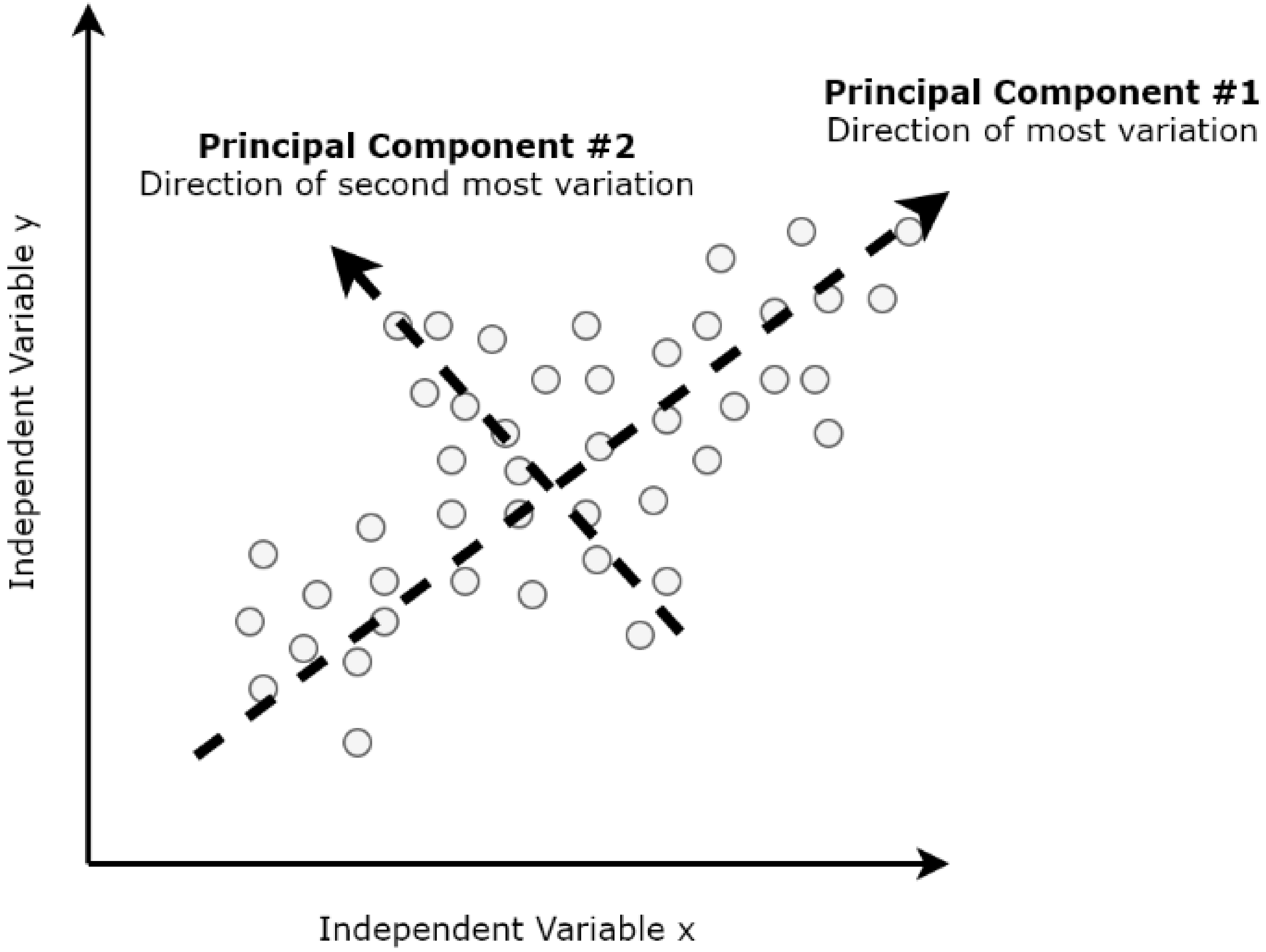
Practice 7

Principal Component Analysis

Introduction to Big Data Analytics(2020)

Problem

- **PCA :** Use RowMatrix in mllib
  - Use predefined function in **pyspark.mllib.linalg.distributed.RowMatrix**



https://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.linalg.distributed.RowMatrix

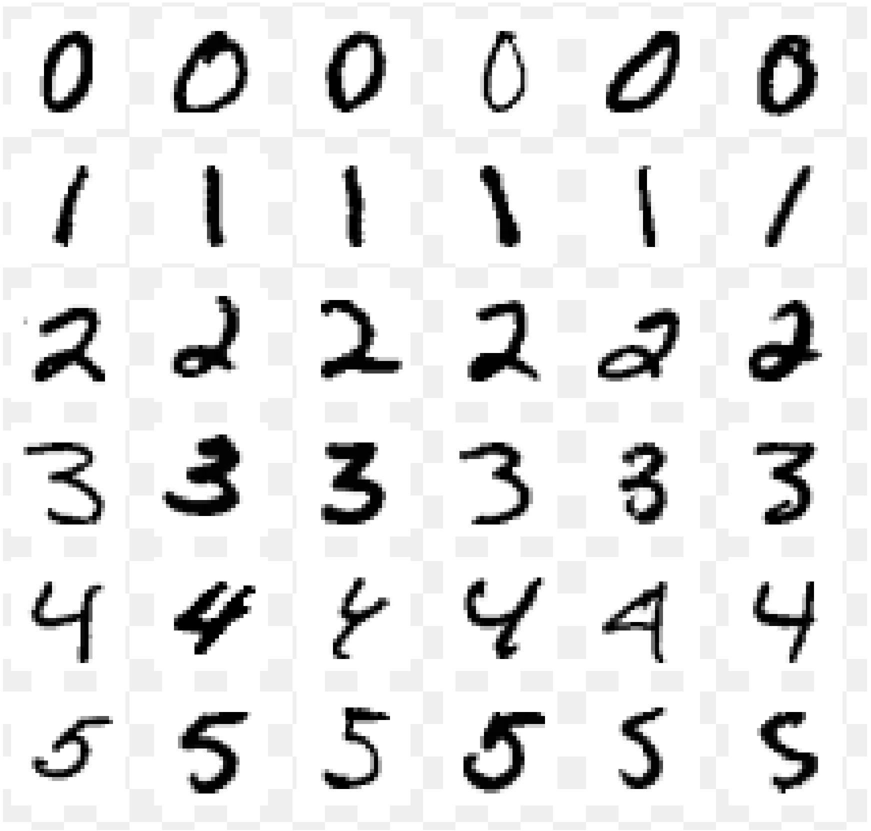
Dataset for PCA

- **MNIST** : Recognition of handwritten digits
  - There are 10 handwritten digits(0~9) in bitmap format.

- 784 Features (pixel values)

1. Pixel 1
2. Pixel 2
…
…
…
63. Pixel 63
784. Pixel 784

* The MNIST Database :



- **You can download the dataset using *sklearn.datasets.fetch_openml* library**

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_openml.html

Practice 7

1. Use predefined classes in *pyspark.mllib.linalg.distributed.RowMatrix* for PCA

   - Row matrix makes the RDD data be row-oriented distributed matrix
   - It has many sub-functions, and you need to use *computePrincipalComponents* to get principal component of Row matrix.
   - *For example,*

```
pc_rdd = mat.computePrincipalComponents(16)
```

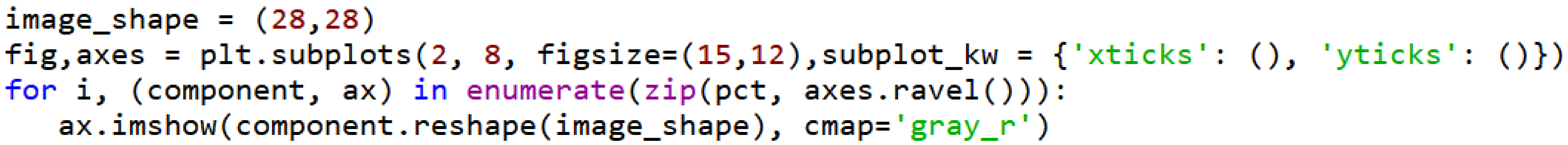
   - In this example, mat means transformed Row Matrix of MNIST dataset

2. Reduce the number of the dataset features from *784* to *16*

https://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.linalg.distributed.RowMatrix

Practice 7

3. Visualize principal components after implementing PCA on MNIST dataset.
    - You need to visualize the principal component of MNIST dataset in *28x28* bitmap.
    - Print out **first 16 pictures in 2x8** matrix.
    - For example,



```
image_shape = (28,28)
fig,axes = plt.subplots(2, 8, figsize=(15,12),subplot_kw = {'xticks': (), 'yticks': ()})
for i, (component, ax) in enumerate(zip(pct, axes.ravel())):
    ax.imshow(component.reshape(image_shape), cmap='gray_r')
```

Practice 7

4. **You need to use predefined arguments we suggest.**

- **Number of data points: 10,000**

  Use first ten thousands(10,000) data points as datasets

- **Number of partitions: 300**

  You can split data when you make it RDDs.

  For example, “ ***RDD = sc.parallelize(Data, numPartition)*** ”

Submission

1. You have to submit “**result.png**” file on iCampus.
2. In your **result.png** file, there must be figures of principal components of MNIST dataset.
3. Deadline: *May 28th 23:59 P.M.*
4. Your **result.png** file must be like following