Machine Learning in Python: Programming questions

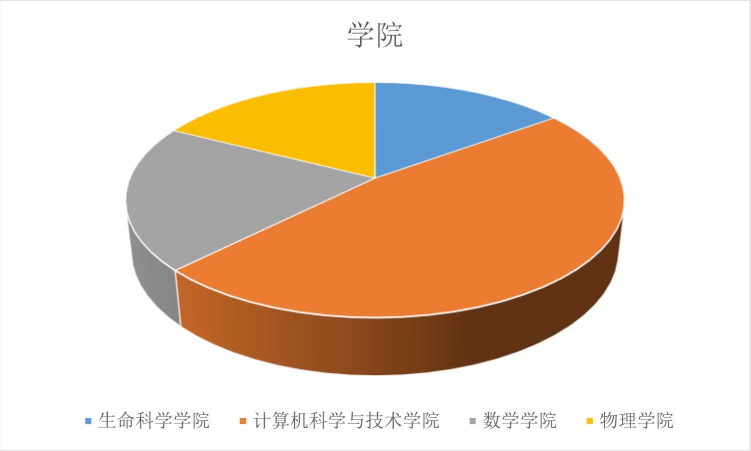
Fengfeng Zhou

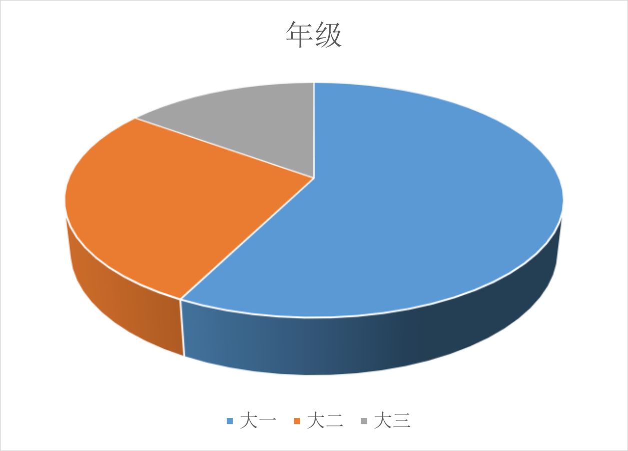
Email: FengfengZhou@gmail.com or ffzhou@jlu.edu.cn

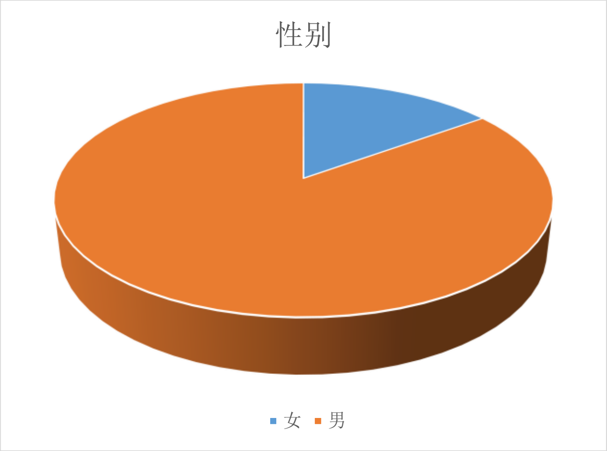
HILab, JLU

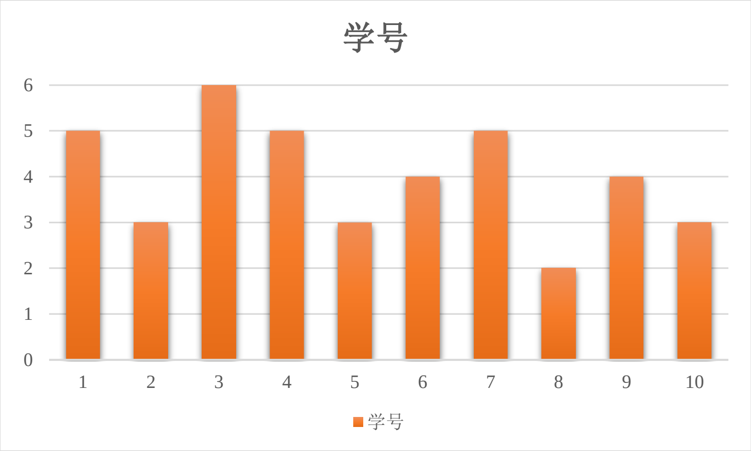
Web: http://healthinformaticslab.org/ffzhou/

Who takes this course – 选修 2022

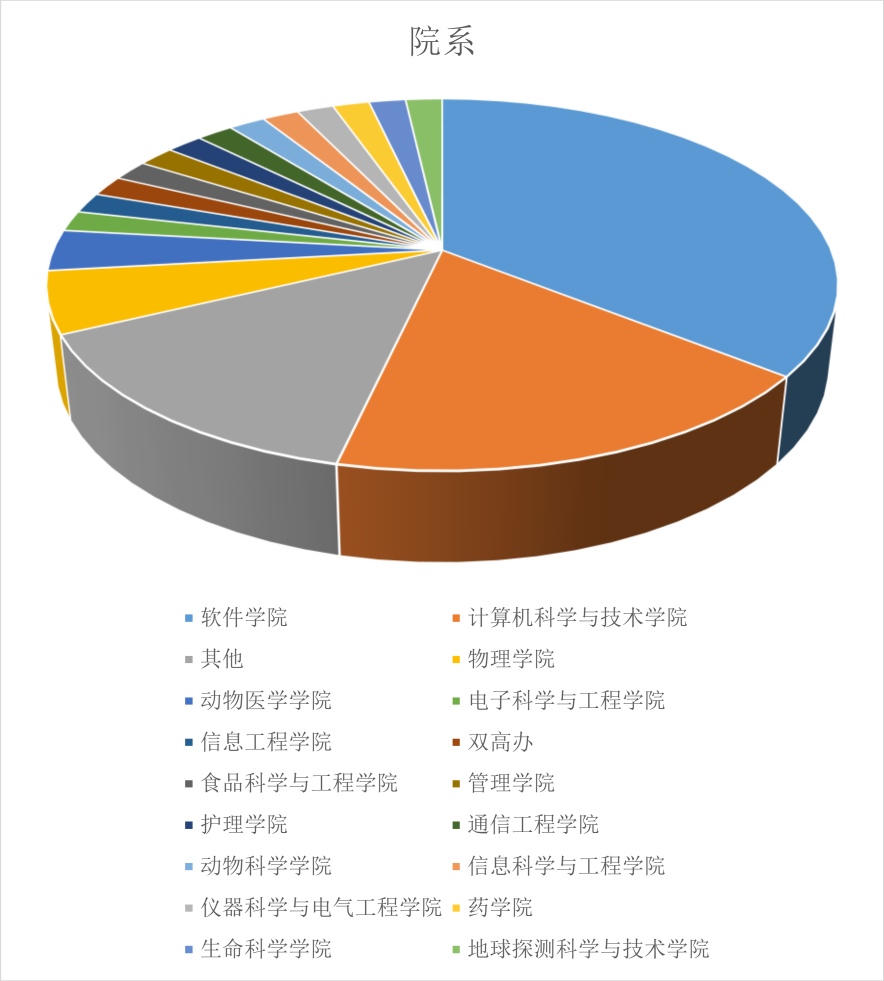








Who takes this course – 旁听 2022



This course equips you with

- Python programming skills on machine learning;
- Skills of how to do a research project on a biomedical big data;
- Knowledge of how to write a scientific paper;

This course encourages you to

- Ask novel questions (or sometimes naïve ones);
- Organize an interdisciplinary team and communicate with your team members from different majors;
- Lead an team and motivate your team members;

Scoring rules

- 40%: regular tests (coding, raising questions, etc)
- 60%: team project (making your question, solving and presenting your solution for 15 min, etc)
  - ✓ Team size: **5** max
- **20% extra**: proving you are better than the existing literature.
- Simple version: McTwo, bonus version: TCGA **(+10%)**

- HILab bonus: I'll help the top 3 teams expand their work into SCI journal articles, if you are interested.

Scoring rules

- Email: **ffzhou@jlu.edu.cn**
- Papers at http://www.ncbi.nlm.nih.gov/pubmed/
- Coding Q&A: https://stackoverflow.com/questions
- Online doc like
  - https://github.com/search?q=python
- No **plagiarism**!

Communication courtesy

- Subject logo: [JLU-HI2023T] [student #]
- Email:
  - ✓ Name, student #, test #
  - ✓ Your solution
- Always "reply" the email (copying the previous emails in the bottom).

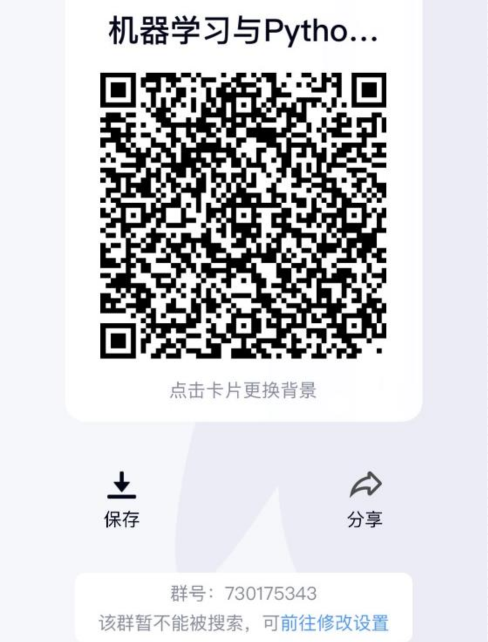
1) 多位同学发作业到我的另一个邮箱了！
2) JLU_HI2019T
3) JLU HI2019T
4) JLU-HI2019
5) 部分同学"坚持不懈"的直到学期结束还在犯同样的错误

Communication courtesy

- Code
- Data
- Readme
- Captured screen of your result window

Course data

- **Web link:** https://pan.baidu.com/s/1YneRMRj4FYVrUZkR0Jcxxw
- **Password:** nhj9



机器学习与Pytho...

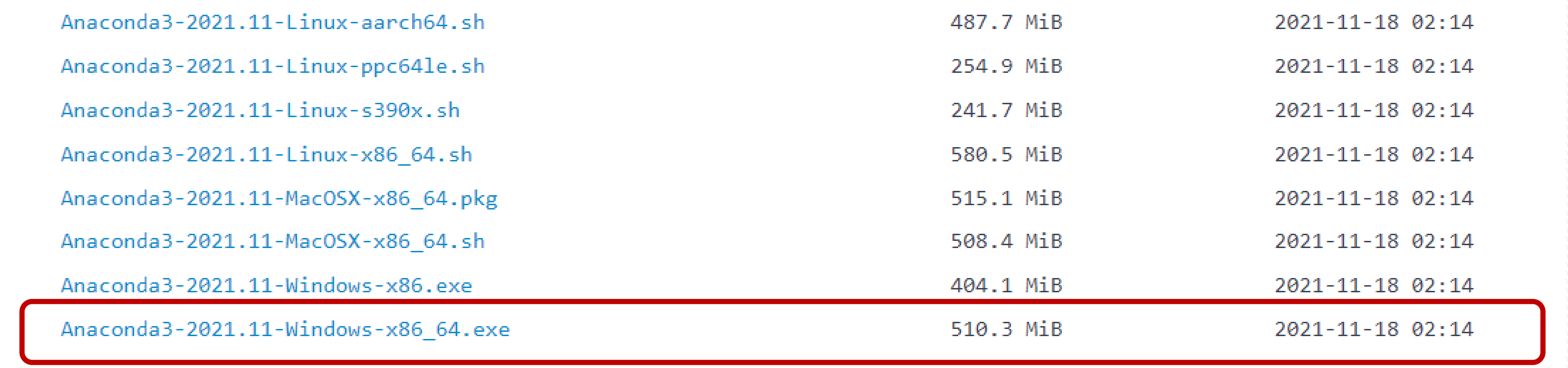
点击卡片更换背景

群号：730175343

该群暂不能被搜索，可前往修改设置

Course data

- **Software: Anaconda**
  https://mirrors.tuna.tsinghua.edu.cn/help/anaconda/



Anaconda3-2021.11-Linux-aarch64.sh	487.7 MiB	2021-11-18 02:14
Anaconda3-2021.11-Linux-ppc64le.sh	254.9 MiB	2021-11-18 02:14
Anaconda3-2021.11-Linux-s390x.sh	241.7 MiB	2021-11-18 02:14
Anaconda3-2021.11-Linux-x86_64.sh	580.5 MiB	2021-11-18 02:14
Anaconda3-2021.11-MacOSX-x86_64.pkg	515.1 MiB	2021-11-18 02:14
Anaconda3-2021.11-MacOSX-x86_64.sh	508.4 MiB	2021-11-18 02:14
Anaconda3-2021.11-Windows-x86.exe	404.1 MiB	2021-11-18 02:14
Anaconda3-2021.11-Windows-x86_64.exe	510.3 MiB	2021-11-18 02:14

Course data

- HILab newbie training projects

Question 1

- Define a function
  - ✓ Factorial of *n*: *n*!=1×2 ×…×*n*
- Test cases
  - ✓ 5
  - ✓ 12
  - ✓ 0
  - ✓ -1
  - ✓ “Hello world”

Question 2

- Ask two binary classification questions of a given cancer type, excluding these example questions
  - ✓ Man versus woman
  - ✓ Asian versus Caucasian

Question 3

- Define a function:

  (Sample, Class, Feature, Matrix) = fLoadDataMatrix(FileName)

	f1	F2	Class
S1	0.5	1.9	1
S2	3.2	0.8	0

Question 4

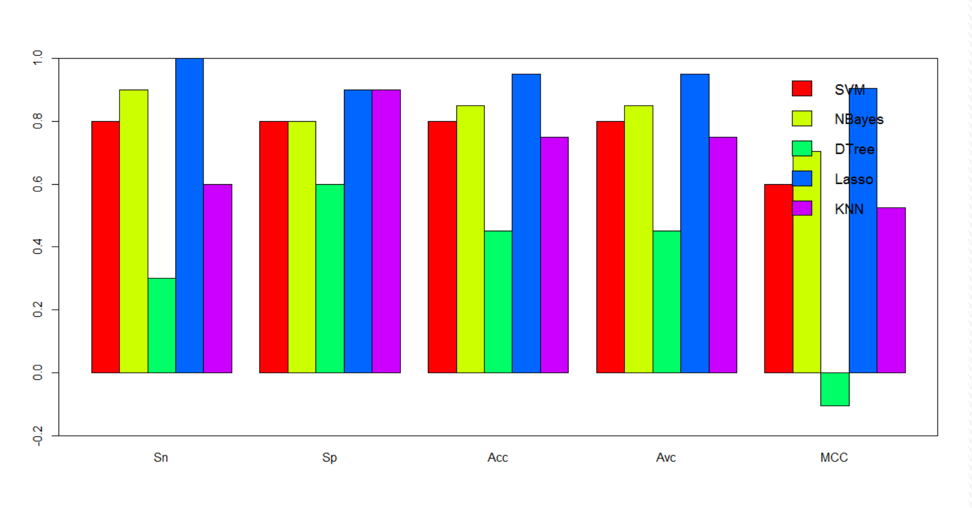
- Find the top 10 ranked features, and print their names.
- (Tvalue, Pvalue) = stats.ttest_ind(dataP, dataN)
- Data file: ALL3.txt

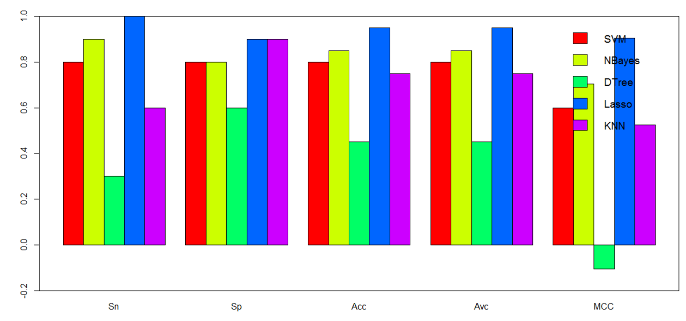
Question 5

- Dot plots of t-test based
  - ✓ Rank-1 vs Rank-2
  - ✓ Rank-9 vs Rank-10
  - ✓ Rank-1000 vs Rank-1001
  - ✓ Rank-10000 vs Rank-10001
- Data file: ALL3.txt

Question 6

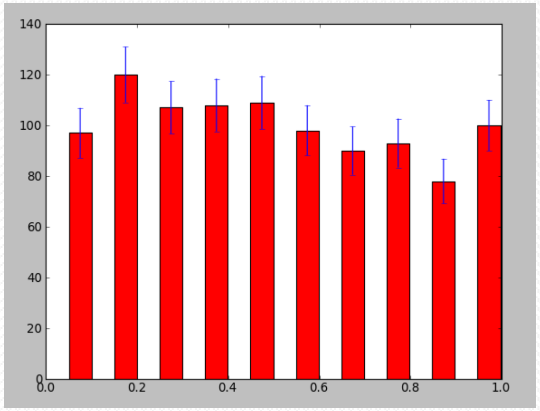
- Histogram of SVM/Nbayes/KNN
  - ✓ Top-1
  - ✓ Top-10
  - ✓ Top-100
  - ✓ Bottom-100
- Data file: ALL3.txt





Question **bonus**

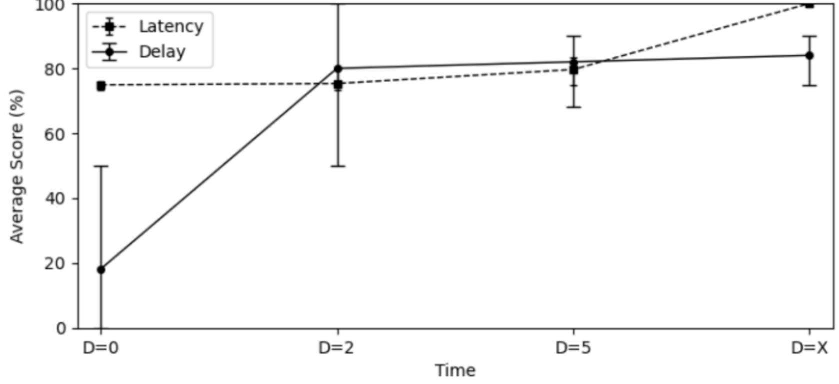
- Histogram of SVM/Nbayes/KNN
  - ✓ Top-1
  - ✓ Top-10
  - ✓ Top-100
  - ✓ Bottom-100
- 10 random runs for the error bars
- Data file: ALL3.txt



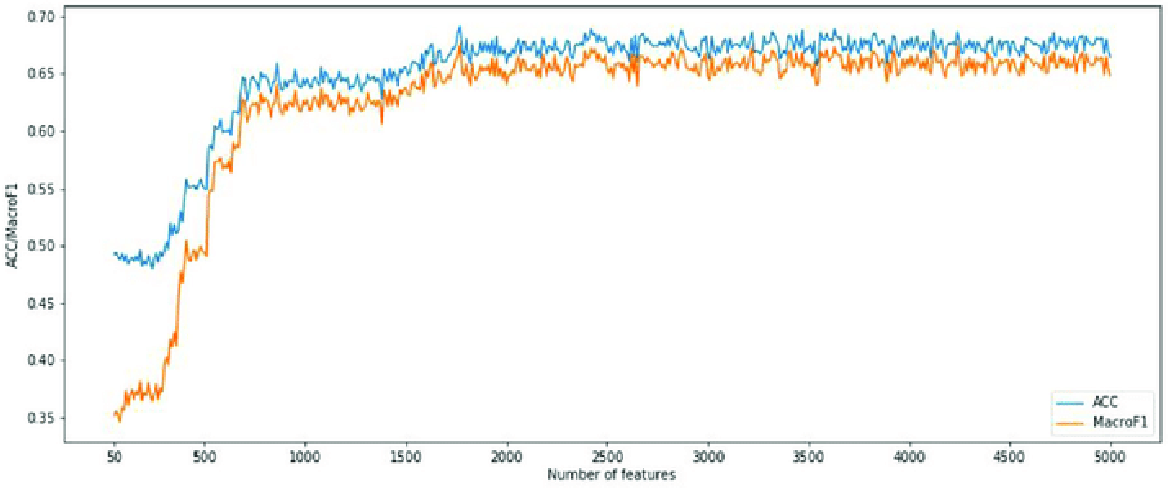
```
import numpy as np
import pylab as plt

data = np.array(np.random.rand(1000))
y,binEdges = np.histogram(data,bins=10)
bincenters = 0.5*(binEdges[1:]+binEdges[:-1])
menStd     = np.sqrt(y)
width      = 0.05
plt.bar(bincenters, y, width=width, color='r', yerr=menStd)
plt.show()
```

https://stackoverflow.com/questions/11774822/matplotlib-histogram-with-errorbars

Question bonus

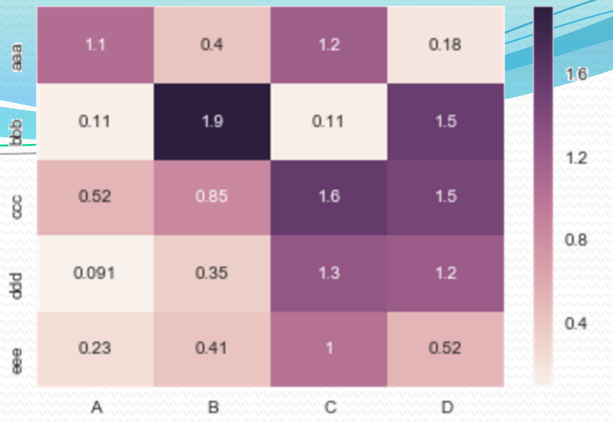


- Incremental Feature Selection of SVM/Nbayes/KNN
  - ✓ Top-ranked 100 features
  - ✓ Line plot
- 10 random runs for the error bars
- Data file: ALL3.txt



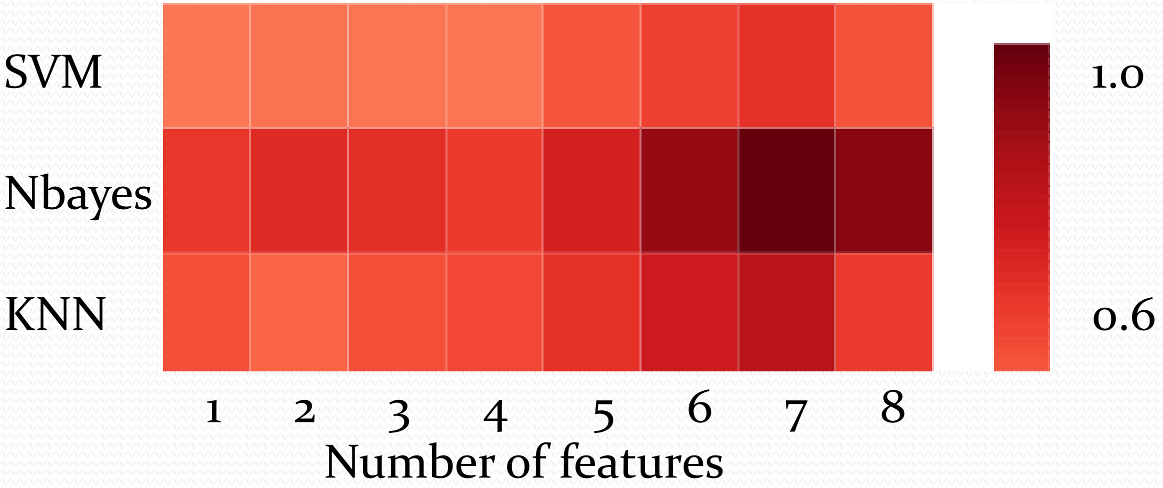
https://stackoverflow.com/questions/63419636/show-error-bar-in-multi-line-plot-using-matplotlib?r=SearchResults&s=13|106.7049

Question bonus



	A	B	C	D
aaa	1.1	0.4	1.2	0.18
bbb	0.11	1.9	0.11	1.5
ccc	0.52	0.85	1.6	1.5
ddd	0.091	0.35	1.3	1.2
eee	0.23	0.41	1	0.52

- Incremental Feature Selection of SVM/Nbayes/KNN
  - ✓ Top-ranked 100 features
  - ✓ Heatmap
- Data file: ALL3.txt



https://stackoverflow.com/questions/12286607/making-heatmap-from-pandas-dataframe?r=SearchResults&s=2|95.1606

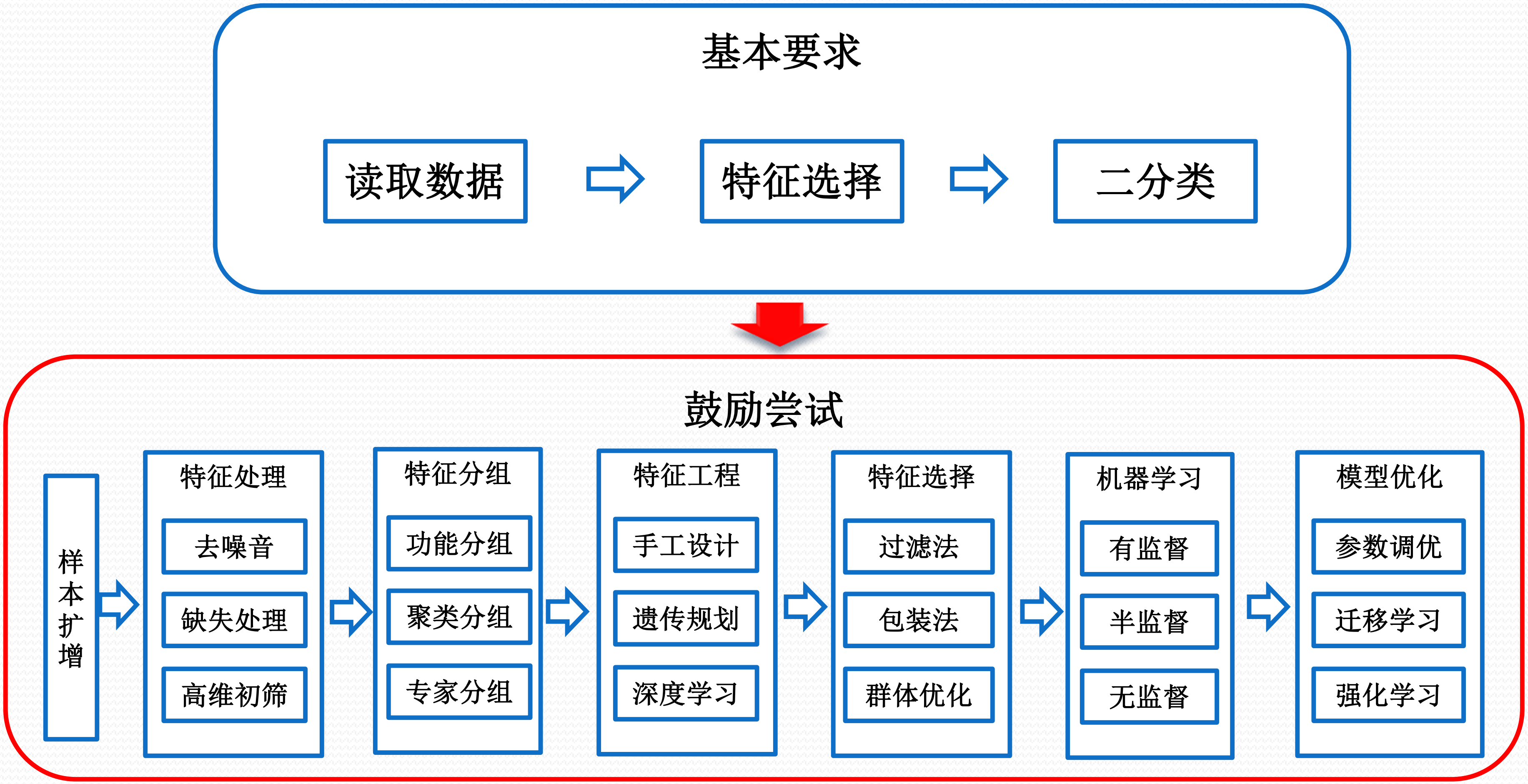
Question 7

- Start your course project, if you have not!
- The date of the final will be determined on the last class of rehearsal.
- You are already working on your course project, if you did the six take-home tests step-by-step.

Question 8

Presentation of your team project

What you can do after this course



基本要求

读取数据 → 特征选择 → 二分类

鼓励尝试

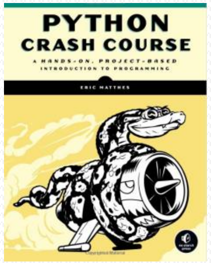
样本扩增 → 特征处理 → 特征分组 → 特征工程 → 特征选择 → 机器学习 → 模型优化

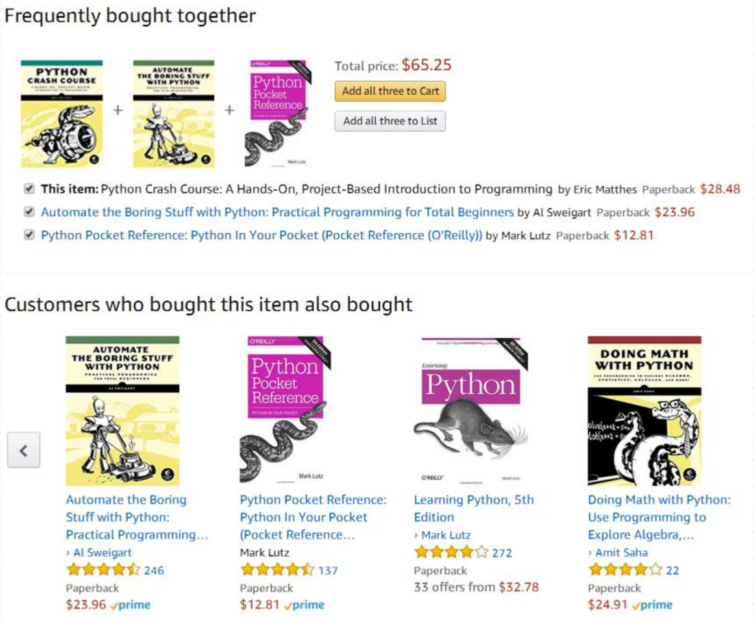
- 特征处理
  - 去噪音
  - 缺失处理
  - 高维初筛
- 特征分组
  - 功能分组
  - 聚类分组
  - 专家分组
- 特征工程
  - 手工设计
  - 遗传规划
  - 深度学习
- 特征选择
  - 过滤法
  - 包装法
  - 群体优化
- 机器学习
  - 有监督
  - 半监督
  - 无监督
- 模型优化
  - 参数调优
  - 迁移学习
  - 强化学习

Data science – financial big data

- Trends of stock market
- Bankruptcy risk
- Long term and short term benefits of a stock

Data science – marketing big data





Frequently bought together

Total price: $65.25

Add all three to Cart

Add all three to List

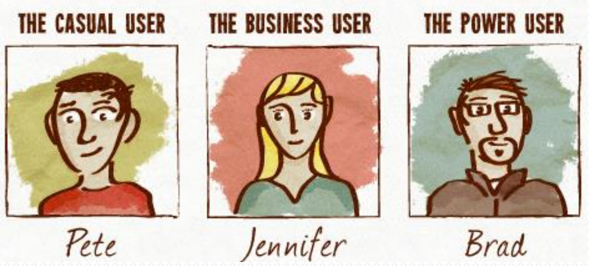
☑ **This item:** Python Crash Course: A Hands-On, Project-Based Introduction to Programming by Eric Matthes Paperback $28.48

☑ Automate the Boring Stuff with Python: Practical Programming for Total Beginners by Al Sweigart Paperback $23.96

☑ Python Pocket Reference: Python In Your Pocket (Pocket Reference (O'Reilly)) by Mark Lutz Paperback $12.81

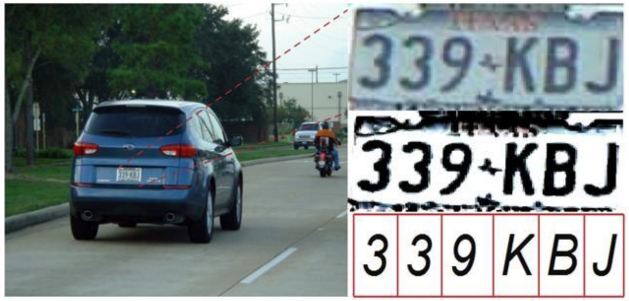
Customers who bought this item also bought

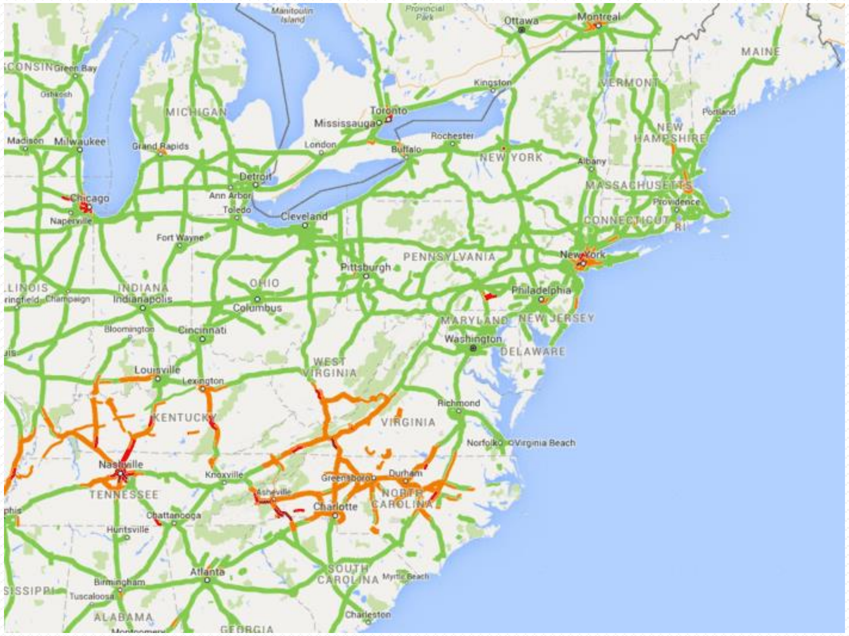
Automate the Boring Stuff with Python: Practical Programming…	Python Pocket Reference: Python In Your Pocket (Pocket Reference…	Learning Python, 5th Edition	Doing Math with Python: Use Programming to Explore Algebra,…
› Al Sweigart	Mark Lutz	› Mark Lutz	› Amit Saha
★★★★½ 246	★★★★½ 137	★★★★☆ 272	★★★★☆ 22
Paperback	Paperback	Paperback	Paperback
$23.96 ✓prime	$12.81 ✓prime	33 offers from $32.78	$24.91 ✓prime



- Advertise similar items
- Recommend “you may like” items
- Send discount codes for items not immediately needed

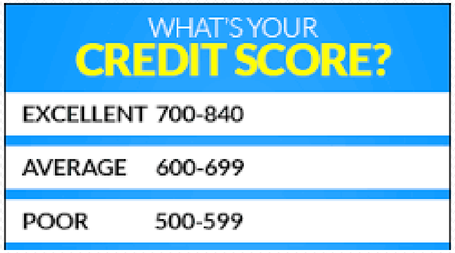
Data science – traffic big data





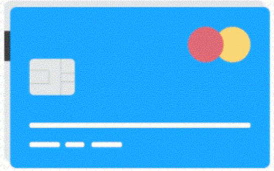
- ● Monitor traffic and predict traffic loads in advance
- ● Detect the license plat data of speeding cars
- ● Recommend better navigation routes

Data science – credit big data



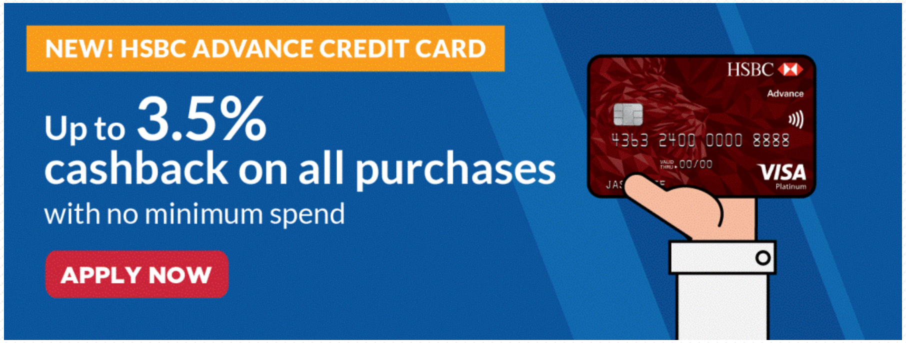
WHAT'S YOUR
CREDIT SCORE?

EXCELLENT	700-840
AVERAGE	600-699
POOR	500-599



3.0% or 5.8%

0.1% or 10.1%



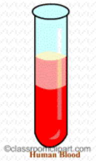
NEW! HSBC ADVANCE CREDIT CARD

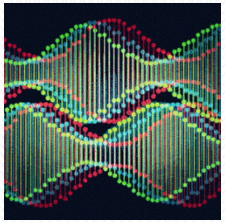
Up to 3.5%
cashback on all purchases
with no minimum spend

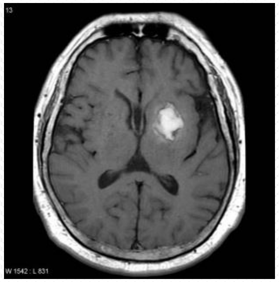
APPLY NOW

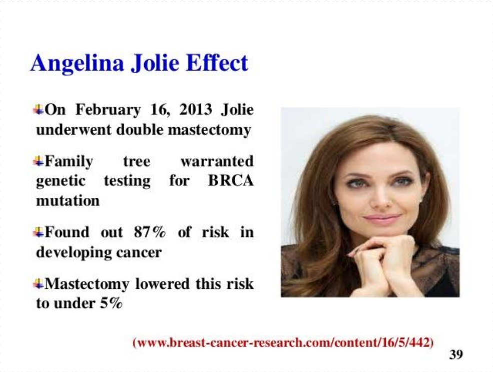
- Determine the amount and interest of a loan
- Approve a credit card with better benefits
- Rent an apartment or even get a decent job

Data science – biomedical big data









Angelina Jolie Effect

- On February 16, 2013 Jolie underwent double mastectomy
- Family tree warranted genetic testing for BRCA mutation
- Found out 87% of risk in developing cancer
- Mastectomy lowered this risk to under 5%

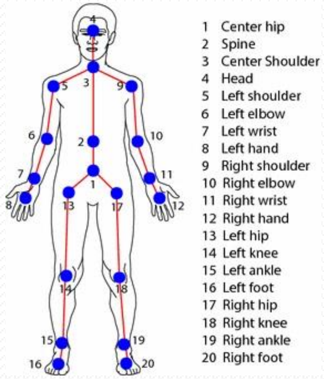
(www.breast-cancer-research.com/content/16/5/442)

39

- Determine disease lesion sites, best treatments, and personal disease risks, etc.
- Detect personal talents

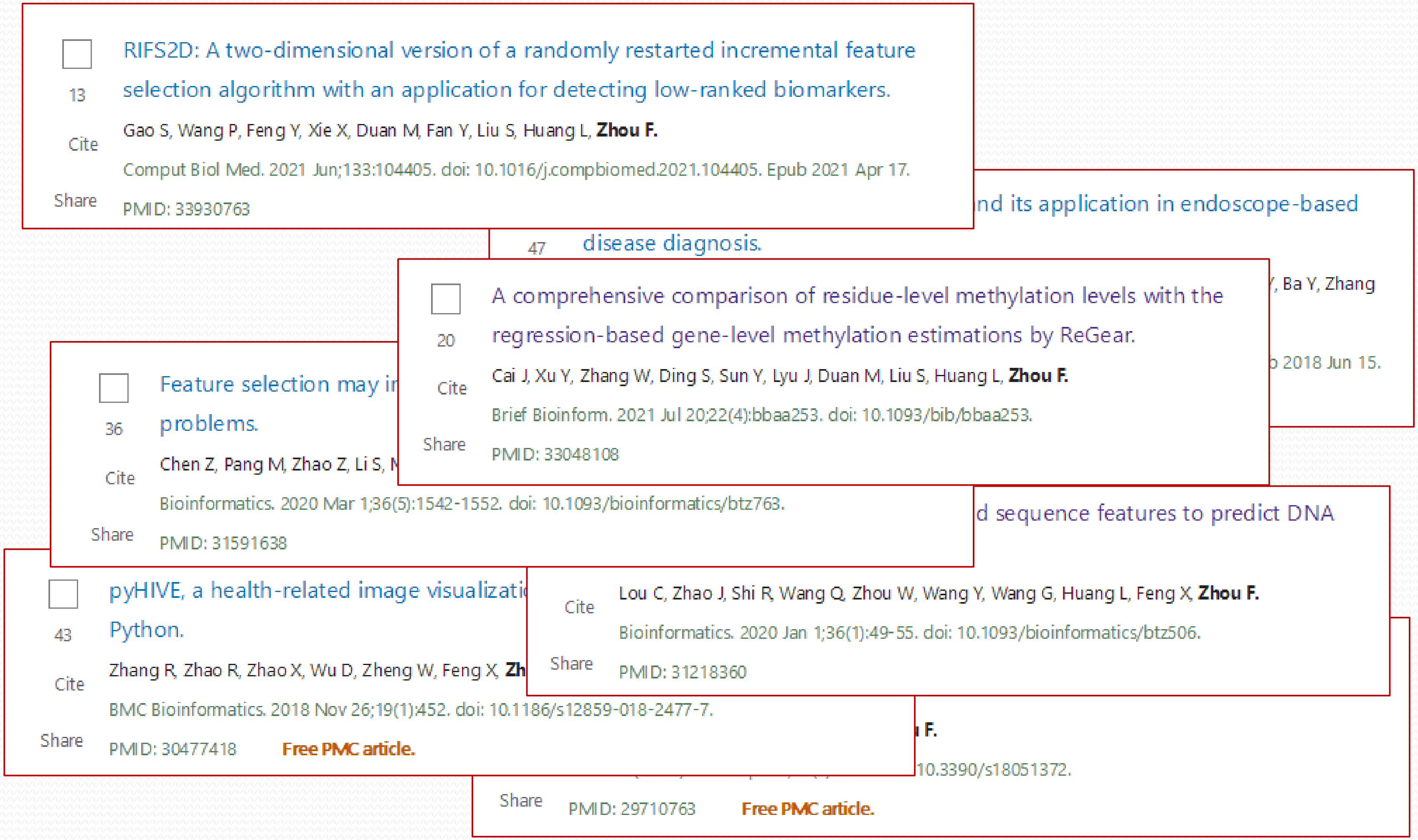
Well, a possibility needs to be achieved by **hard working**!

Data science – entertainment big data



- Predict audience interests
- Effective ad targeting
- Predicting gait popularity

You are the next undergrad that I’m proud of!



□ RIFS2D: A two-dimensional version of a randomly restarted incremental feature selection algorithm with an application for detecting low-ranked biomarkers.

13

Cite

Share

Gao S, Wang P, Feng Y, Xie X, Duan M, Fan Y, Liu S, Huang L, **Zhou F.**

Comput Biol Med. 2021 Jun;133:104405. doi: 10.1016/j.compbiomed.2021.104405. Epub 2021 Apr 17.

PMID: 33930763

nd its application in endoscope-based disease diagnosis.

47

, Ba Y, Zhang

2018 Jun 15.

□ A comprehensive comparison of residue-level methylation levels with the regression-based gene-level methylation estimations by ReGear.

20

Cite

Share

Cai J, Xu Y, Zhang W, Ding S, Sun Y, Lyu J, Duan M, Liu S, Huang L, **Zhou F.**

Brief Bioinform. 2021 Jul 20;22(4):bbaa253. doi: 10.1093/bib/bbaa253.

PMID: 33048108

□ Feature selection may i problems.

36

Cite

Share

Chen Z, Pang M, Zhao Z, Li S, M

Bioinformatics. 2020 Mar 1;36(5):1542-1552. doi: 10.1093/bioinformatics/btz763.

PMID: 31591638

d sequence features to predict DNA

Cite

Share

Lou C, Zhao J, Shi R, Wang Q, Zhou W, Wang Y, Wang G, Huang L, Feng X, **Zhou F.**

Bioinformatics. 2020 Jan 1;36(1):49-55. doi: 10.1093/bioinformatics/btz506.

PMID: 31218360

□ pyHIVE, a health-related image visualizati Python.

43

Cite

Share

Zhang R, Zhao R, Zhao X, Wu D, Zheng W, Feng X, **Zh**

BMC Bioinformatics. 2018 Nov 26;19(1):452. doi: 10.1186/s12859-018-2477-7.

PMID: 30477418 **Free PMC article.**

F.

10.3390/s18051372.

Share

PMID: 29710763 **Free PMC article.**

Team project requirements

Fengfeng Zhou
Email: FengfengZhou@gmail.com or ffzhou@jlu.edu.cn
HILab, JLU
Web: http://healthinformaticslab.org/ffzhou/

Homework 7 – pipeline

Combine your Python scripts of the previous homework into a pipeline, with the following requirements

- One main interface script with all the functional modules integrated, all parameters in one config file
- Report all the results as output files
- A manual document

Remember to project an example data file! So that I can run the script directly.

Homework 7 – pipeline (1)

One main interface script with all the functional modules integrated, all parameters in one config file with the command line:

```
python ./pipeBinClass.py pbc.conf
```

```
# Configuration for a binary classification project
# Team leader: XXX (student number)
# Team member: YYY (student number)
# Contact: TeamLeader-Email
# Date: 2020-05-08

Input: DataMatrix.csv
OutputDir: output-dir

# config for t-test
MaxFeature: 20

# Dot-plot
DP-ImageFormat: JPG
DP-DOI: 300
DP-Output: Hello1.jpg
```

Homework 7 – pipeline (2)

Report all the results as output files (in the output directory “output-dir”)

- Features ranked by t-test in a text file
- Dot-plot image files
- Histogram image files
- …

Homework 7 – pipeline (3)

A manual document

- Describe the syntax of the configuration file
- Describe the output data files

Homework 8 – team project

Documents of the team project

- PPT (with all members’ names, student numbers, who is the team leader)
- The pipeline and configuration file used to generate this PPT
- Data file