阿里云天池牛年读书会

Python编程
——从数据分析到数据科学

分享嘉宾：朝乐门

中国人民大学　副教授、博士生导师、数据科学50人

天池读书会

TIANCHI 天池

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

Python编程：从数据分析到数据科学

从Python学习和编程中常见的误区到面向数据分析的Python编程特点的讲解

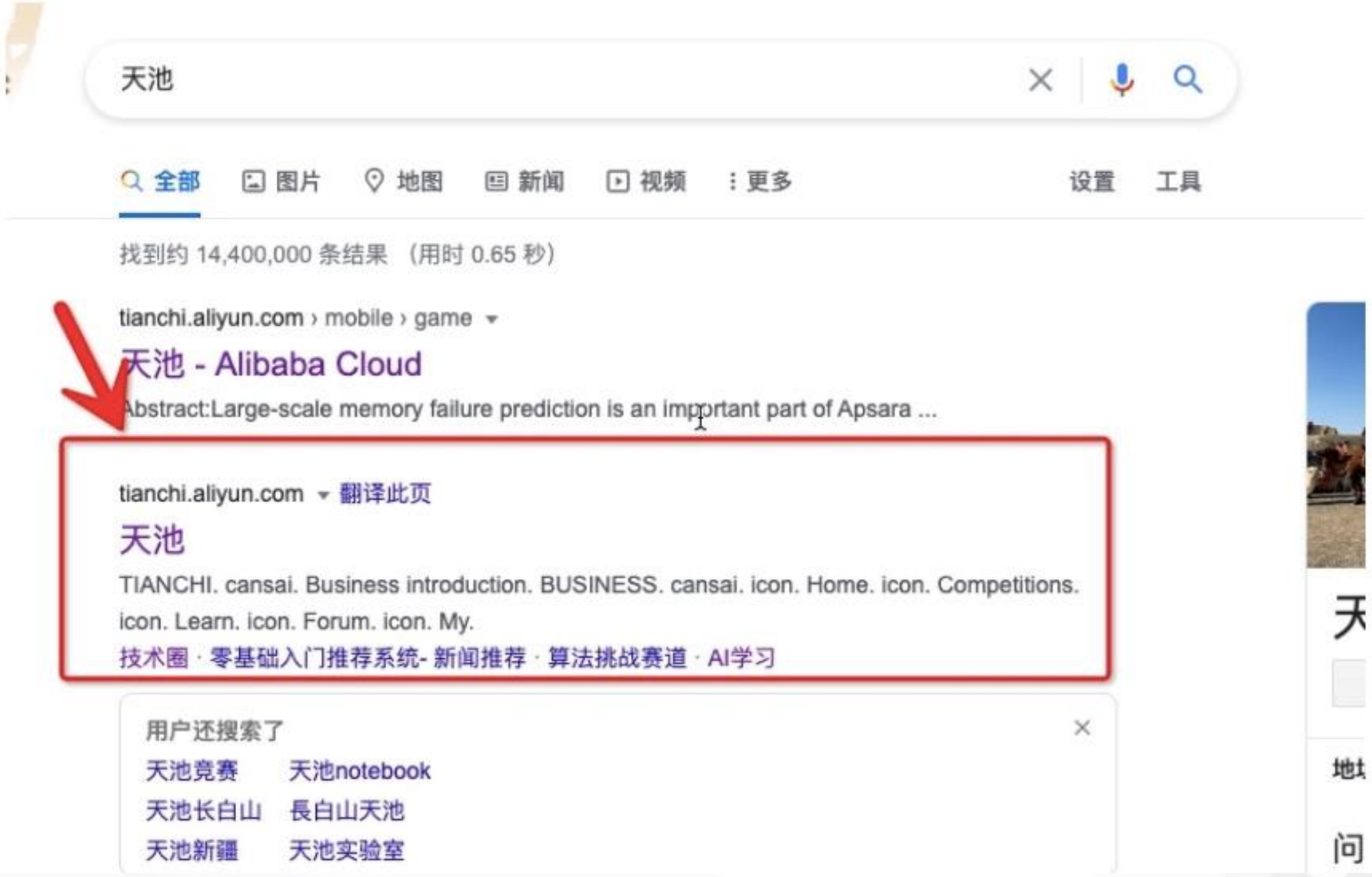
带你从零入门数据科学。

直播嘉宾：朝乐门 中国人民大学

直播时间：7月23日20:00

扫码领取读书会配套学习资源

1）首先需要进入天池官网，大家打开浏览器，搜索 天池 ，找到 tianchi.aliyun.com即可访问进入天池官网；



2）在天池官网，将鼠标移到 天池学习 ，即可出现下拉列表，点击 天池读书会 ，即可进入天池读书会的页面。

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

3）在天池读书会页面，你可以对对应的读书会图书进行提问，优秀的提问还有机会获得赠书，还可以点击配套的训练营或者课程资源进入学习，还有点击实践代码获取读书会的项目实践的代码，跟着我一起进行项目实践和代码学习，同时还有很多其他的读书会，大家也可以观看举办过的读书会的回放，或者预约还没开始的读书会。

朝乐门 中国人民大学

直播主题 《Python编程：从数据分析到数据科学》

直播时间 2021年7月23日 20:00

学习资料 Pandas教程

实践项目 威斯康星乳腺癌数据分析及自动诊断

提问 | 学习课程 | 购买地址 | PPT下载 | 实践代码 | 预约直播

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

TIANCHI 天池

目录

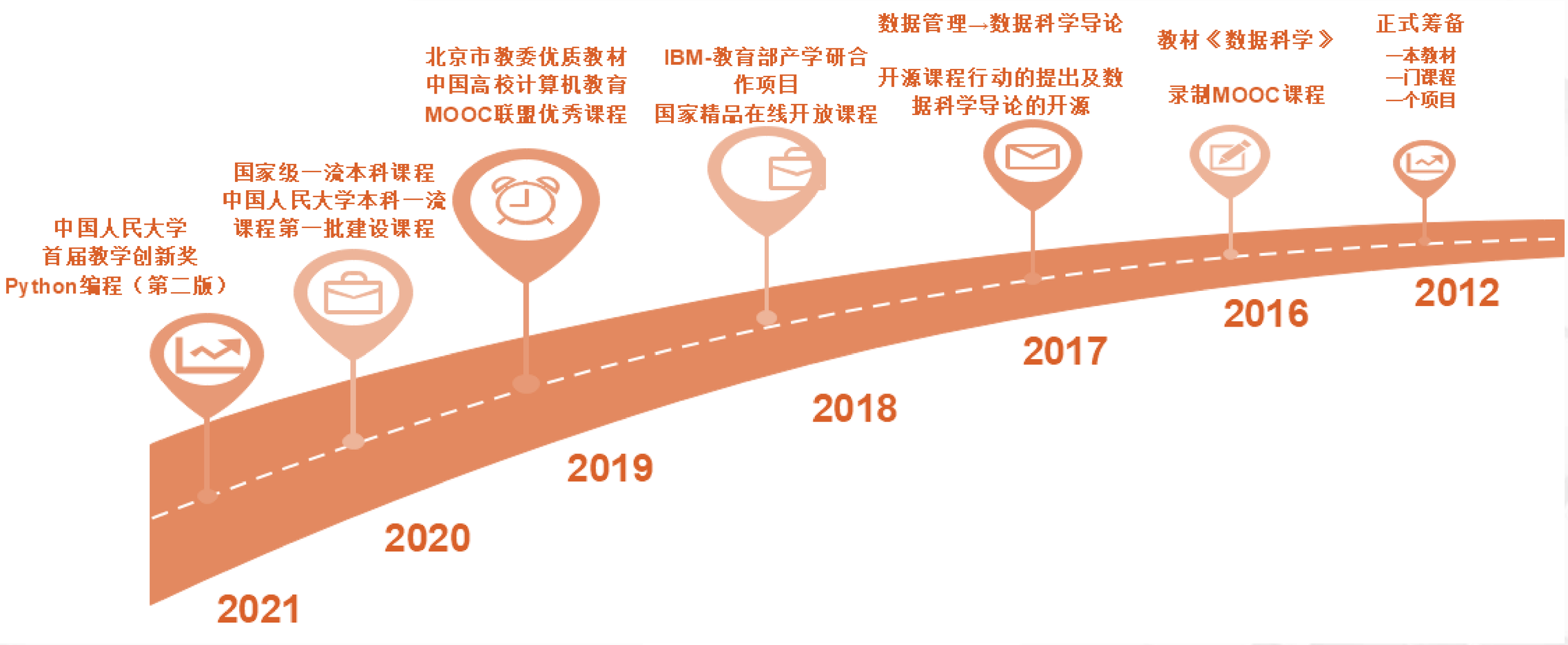
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

分享嘉宾简介

• 朝乐门：中国人民大学副教授、博士生导师；数据科学50人、国家级一流本科课程《数据科学导论》负责人、阿里云MVP、中国计算机学会信息系统专委员会委员、全国高校人工智能与大数据创新联盟专家委员会副主任、国家核心期刊《计算机科学》执行编委、国际期刊《Data Science and Informetrics》副主编；

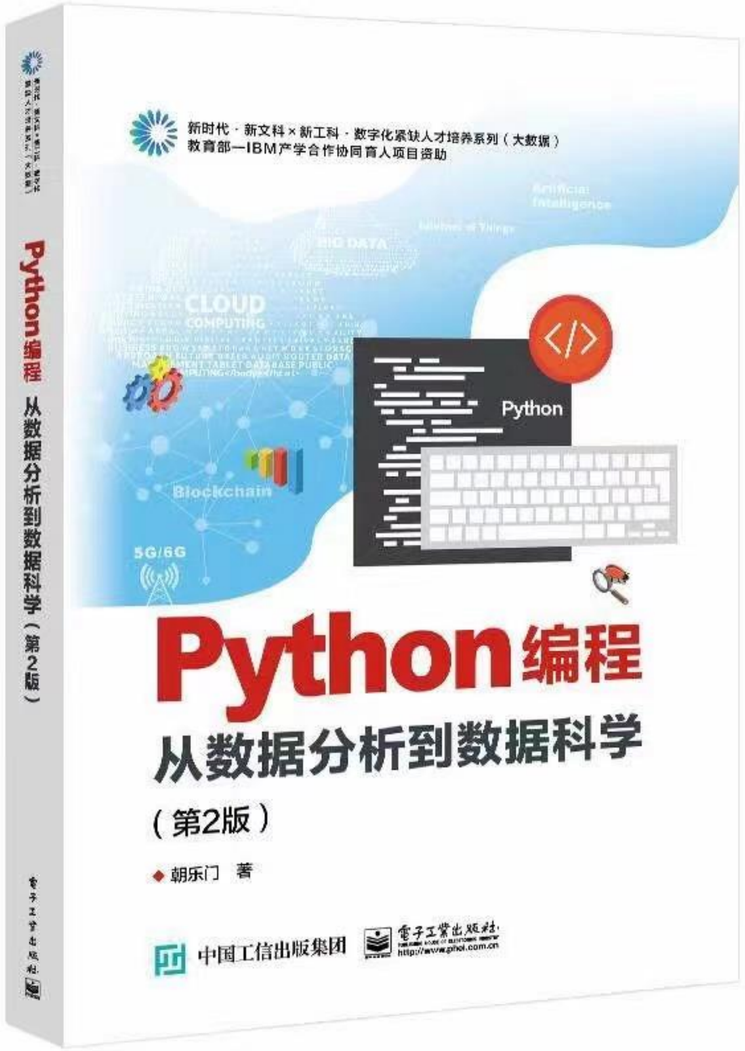
• 朝乐门是我国第一部系统阐述数据科学理念、理论、方法、技术和工具的重要专著——《数据科学》（清华大学出版社，2016）的作者。2019年，其编写的教材《数据科学理论与实践（第二版）》被北京市教委认定为“北京市高等学校优质教材”，本科课程《数据科学导论》被认定为“中国人民大学本科一流专业第一批建设课程”。

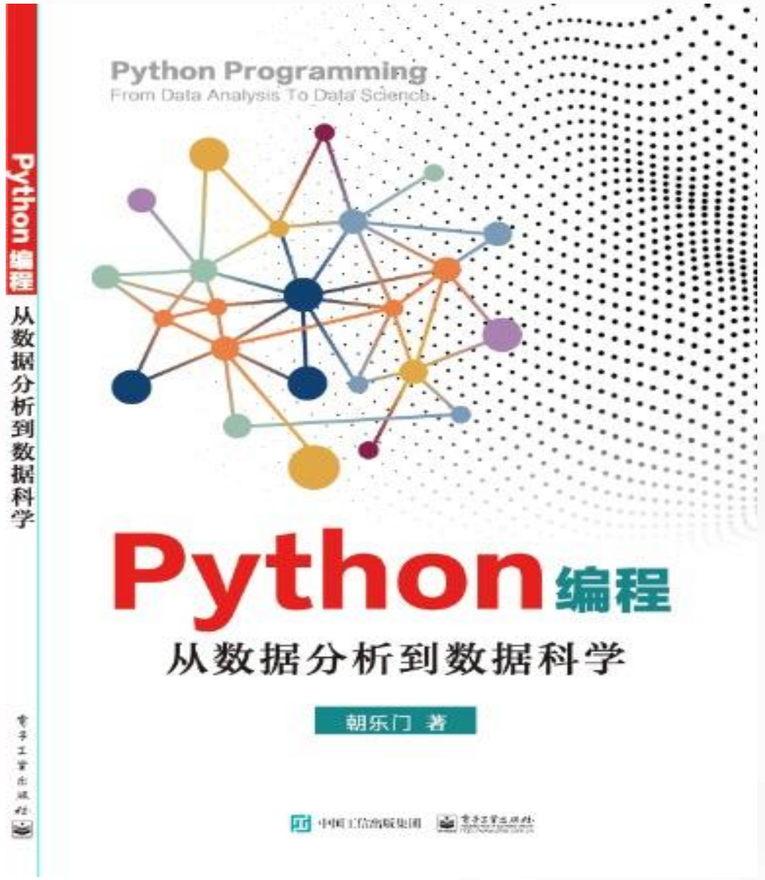
分享嘉宾简介——朝乐门



直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

图书简介

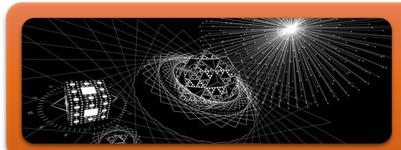




直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

Python学习的误区或困境

“将Python当作Java/C来教（或学）”，换一个“新语言”来讨论“老问题”



数科类专业与计科类专业中选用的Python教材没什么区别

只关注语法，不关注背后的思想和逻辑

“多数学生都已有C/Java等语言的基础，Python属于“第二外语”

“先讲知识点，后练习代码”式老套路，主次颠倒；

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

如何写出高质量的Python代码

写Python代码的两个境界

- 能run就ok?
- Pythonic coding

Python的Zen

- import this

PEP8-Style Guide for Python Code

- adapted from *Guido's Python Style Guide*
- https://legacy.python.org/dev/peps/pep-0008/

Google Style Guides系列

- Google Python Style Guide
- https://github.com/google/styleguide/blob/gh-pages/pyguide.md

图书内容

1.Py语法要点

- Python for DA
- 数据分析/数据科学项目中常用Python语法要点

2. 数据准备

- 随机数
- 多维数组
- 数据框
- 可视化
- Web爬取

3. 算法与模型

- 机器学习
- 统计学
- 图像处理
- 自然语言理解

4.大数据分析

- Spark
- MongoDB
- MLib

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

TIANCHI 天池

图书特色

本书特色

1.本书是专门为数据科学、大数据分析和大数据应用类人才编写的Python语言教材。

2.本书改变了同类教材中普遍存在的“将Python当作Java/C来教（或学）”的现状，突显Python在数据分析和数据科学中的特殊语法和新思维。

3.本书改变了传统教材的“以先讲知识点，后写代码”式编写风格，首次将代码放在中心位置，配有最必要的文字提示，做到主次分明，一目了然，便于学习。

4.本书主要讲解大数据人才常用的Python语言及第三方扩展库的基础知识、思路、方法、经验和技巧，打通了从Python到数据分析再到数据科学的通道，改变了传统图书中对Python、数据分析和数据科学三个知识领域的拆分式讲解读模式。

5.本书创新性地以Markdown的方式来组织内容编写，采用全彩印刷，给读者直观的认识，让知识获取更加生动。

6.本书按照朝乐门老师提出的“开源课程（Open-Source Course，OSC）行动倡议”，为高校教师提供开源社区，并以开源课程模式共同建设与维护课程资源，包括教学大纲、教学方案、PPT、源代码、原始数据、习题、勘误表以及其他参考资料。

名家推荐

数据科学与大数据技术专业的申报与建设在我国方兴未艾，亟需一系列专业核心教材。《Python编程：从数据分析到数据科学》是一本极具特色的优秀教材，主要特色包括：

1. 定位明确。以数据科学与大数据技术专业的人才培养需求为立足点，充分借鉴国外一流大学相关课程的建设经验，知识内容的选择与组织符合专业的教学需要。

2. 简明扼要。设有Q&A、图表、思路、注意、提示等栏目，深入浅出地讲解了面向数据科学的Python编程中的重点、难点与疑点。

3. 实用性强。采取以代码为中心的编写风格，内容涉及Python基础语法、数据规整化、数据可视化、自然语言处理、Web爬取、统计分析、机器学习、Spark编程、NoSQL数据库、初学者常见错误及纠正方法、数据科学岗位面试题以及扩展阅读书目，全景展现了数据科学人才应必备的知识与技能。

—— 杜小勇（中国人民大学教授，中国计算机学会数据库专业委员会主任）

《Python编程：从数据分析到数据科学》一书定位明确、内容丰富、结构完整、特色鲜明，理论与实践并重，符合数据科学与大数据技术、大数据管理与应用等大数据类专业人才培养的新需求，是一本非常值得推荐的优秀教材。

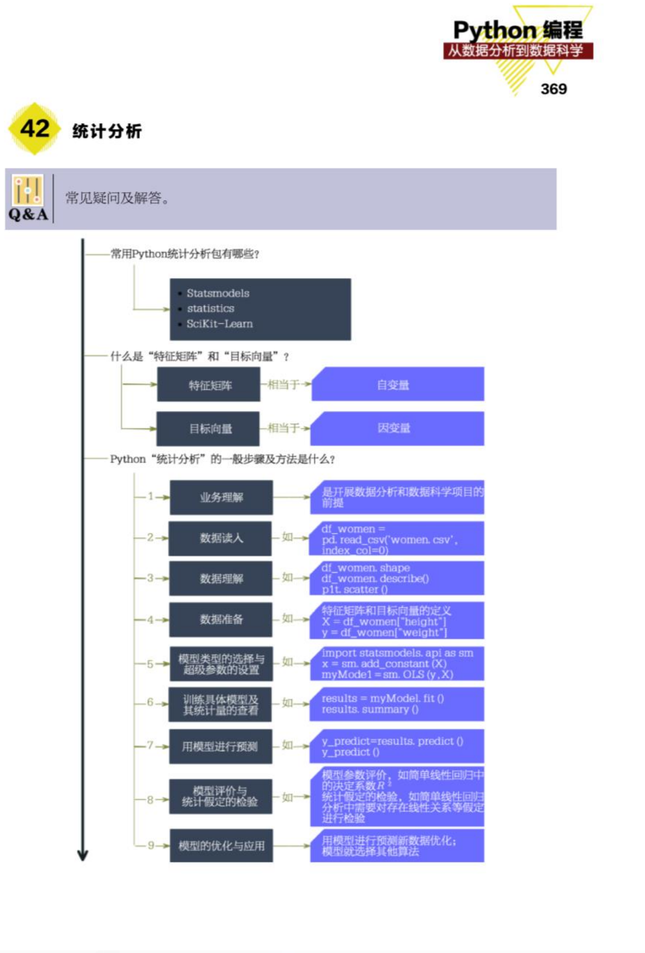
——陈钟（北京大学教授，教育部高等学校计算机类专业教学指导委员会副主任）

从顶层设计到知识体系的精心讲解，这本书充分体现了数据科学专家独有的3C精神——原创性（Creative）设计、批判性（Critical）思维和好奇性（Curious）提问。本书不但凝聚了作者的智慧与心血，而且探索了Python教材编写方式的重要创新，是继《数据科学》《数据科学理论与实践》之后又一力作。

—— 邢春晓（清华大学教授，中国计算机学会信息系统专委会副主任）

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

学习方法



42.7 拟合优度评价

In [18]:

提示 评价回归直线的拟合优度——计算 R^2（决定系数）。

```
results.rsquared
```

思路 R^2 的取值范围为[0,1]，越接近 1，说明“回归直线的拟合优度越好”。

Out [18]: 0.9910098326857505

42.8 建模前提假定的讨论

In [19]:

注意 在基于统计学方法完成数据分析和数据科学任务时，不仅需要进行模型优度的评价，还需要重点分析统计方法的“应用前提假定”是否成立。

思路 通常，统计分析都是建立在一个或多个“前提假定条件”之上。以简单线性回归为例，其“前提假定条件”如下。

```
#（1）X 和 y 之间存在线性关系→检验方法为计算 F 统计量
#（2）残差项（的各期）之间不存在的自相关性→检验方法为计算 Durbin-Watson 统计量
#（3）残差项为正态分布的随机变量→检验方法为计算 JB 统计量
```

In [20]:

提示 查看 F 统计量的 p 值。

```
results.f_pvalue
```


50 继续学习本书内容的推荐资源

50.1 重要网站

[1] Python 官网：有很多权威资料，如 Python Tutorial 等，URL：Python.org。
[2] Pypi 官网：Python 包索引（Python packages index），可以看到每个包的帮助文档，URL：https://pypi.org/project/pip/。
[3] LearnPython.org：学习 Python 的著名网站。
[4] DataCamp：用 Python 学习数据科学的著名网站：https://www.datacamp.com/。
[5] PyData：Python数据分析的著名社区，URL：https://groups.google.com/forum/#!forum/pydata。
[6] pystatsmodels：Python 统计分析的著名讨论社区，https://groups.google.com/ forum/#!forum/pystatsmodels。
[7] Python cheat sheet：一图讲解 Python 知识，URL：https://ehmatthes.github.io/pcc/cheatsheets/README.html。
[8] Python2 和 Python3 的区别：https://wiki.python.org/moin/Python2orPython3。
[9] PEP 8，Python 写代码规范（Style Guide for Python Code），URL：https://www.python.org/dev/peps/pep-0008/。
[10] Kaggle：有很多数据科学和 Python 相关的竞赛、数据集等，URL：https://www.kaggle.com。
[11] Python Weekly：Python 每周一报，报告内容包括新闻、文章、新版本发布、工作岗位等，建议订阅。URL：https://www.pythonweekly.com/。
[12] GitHub：有很多 Python 开源项目，可以参见《Top 20 Python AI and Machine Learning projects on Github》等相关研究报告，建议读者多参与开源项目。当然，也有 Python 之父 Guido van Rossum 等大牛发起的开源项目。
[13] Stack Overflow：虽然不是只针对 Python 和数据科学的网站，但是本书作者最喜欢的网站之一，URL：https://stackoverflow.com/。

50.2 重点图书

[1]《Python Data Science Handbook》（Jake VanderPlas）：一本 Python 和数据科学相结合的好书，推荐认真阅读。

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

实战演示、互动交流

大家可以使用手机扫左侧海报二维码，或者**电脑**访问下方地址进入天池读书会页面，点击今天读书会中的 **实践代码** 和我一起进行项目实践学习，天池为大家准备好了代码和运行环境，非常方便。

https://tianchi.aliyun.com/specials/promotion/activity/bookclub

朝乐门 中国人民大学

直播主题 《Python编程：从数据分析到数据科学》

直播时间 2021年7月23日 20:00

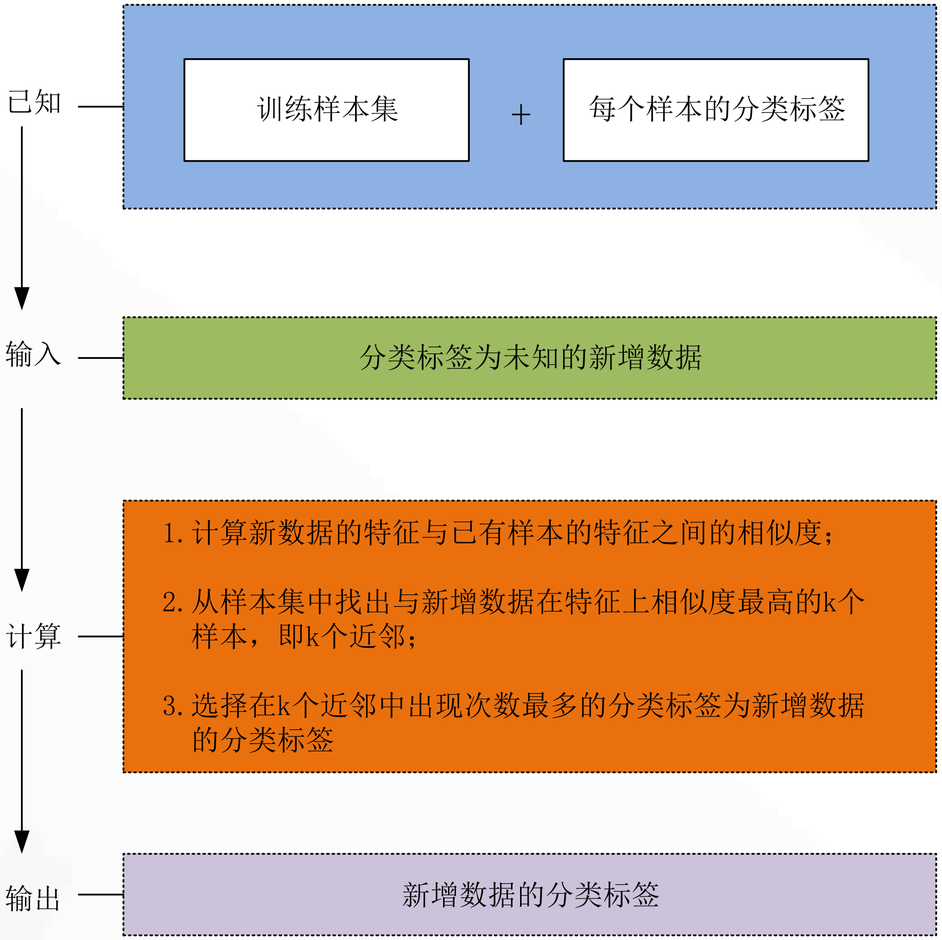
学习资料 Pandas教程

实践项目 威斯康星乳腺癌数据分析及自动诊断

提问 | 学习课程 | 购买地址 | PPT下载 | 实践代码 | 预约直播

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

KNN算法



K=4

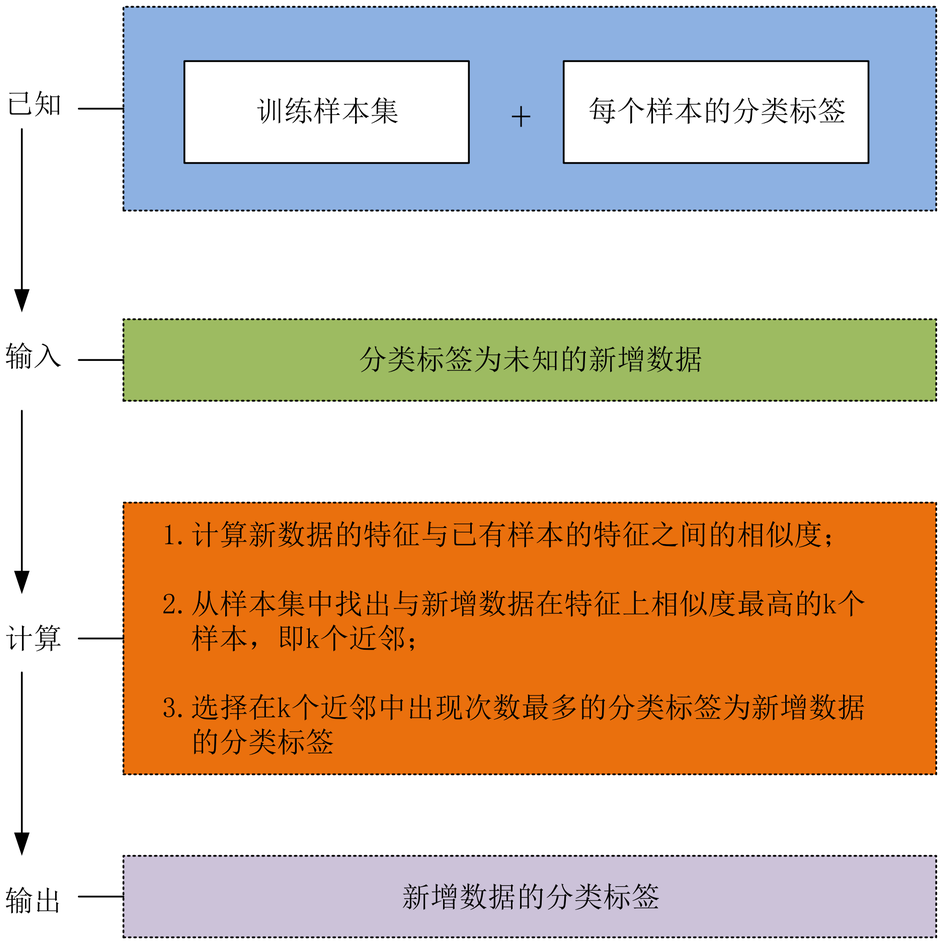
电影名称	打斗镜头	接吻镜头	电影类型
California Man	3	104	爱情片
He’s Not Really into Dudes	2	100	爱情片
Beautiful Woman	1	81	爱情片
Kevin Longblade	101	10	动作片
Robo Slayer 3000	99	5	动作片
Amped II	98	2	动作片

当遇到一部未看过的电影（不知道剧情，但知道其中的打斗次数和接吻次数分别为18和90）时，如何知道它是爱情片还是动作片？

$$d = \sqrt{(3-18)^2 + (104-90)^2} = 20.5$$

电影名称	与未知电影的距离
California Man	20.5
He’s Not Really into Dudes	18.7
Beautiful Woman	19.2
Kevin Longblade	115.3
Robo Slayer 3000	117.4
Amped II	118.9

k-近邻算法



K=4

电影名称	打斗镜头	接吻镜头	电影类型
California Man	3	104	爱情片
He’s Not Really into Dudes	2	100	爱情片
Beautiful Woman	1	81	爱情片
Kevin Longblade	101	10	动作片
Robo Slayer 3000	99	5	动作片
Amped II	98	2	动作片

当遇到一部未看过的电影（不知道剧情，但知道其中的打斗次数和接吻次数分别为18和90）时，如何知道它是爱情片还是动作片？

$$d = \sqrt{(3-18)^2 + (104-90)^2} = 20.5$$

电影名称	与未知电影的距离
California Man	20.5
He’s Not Really into Dudes	18.7
Beautiful Woman	19.2
Kevin Longblade	115.3
Robo Slayer 3000	117.4
Amped II	118.9

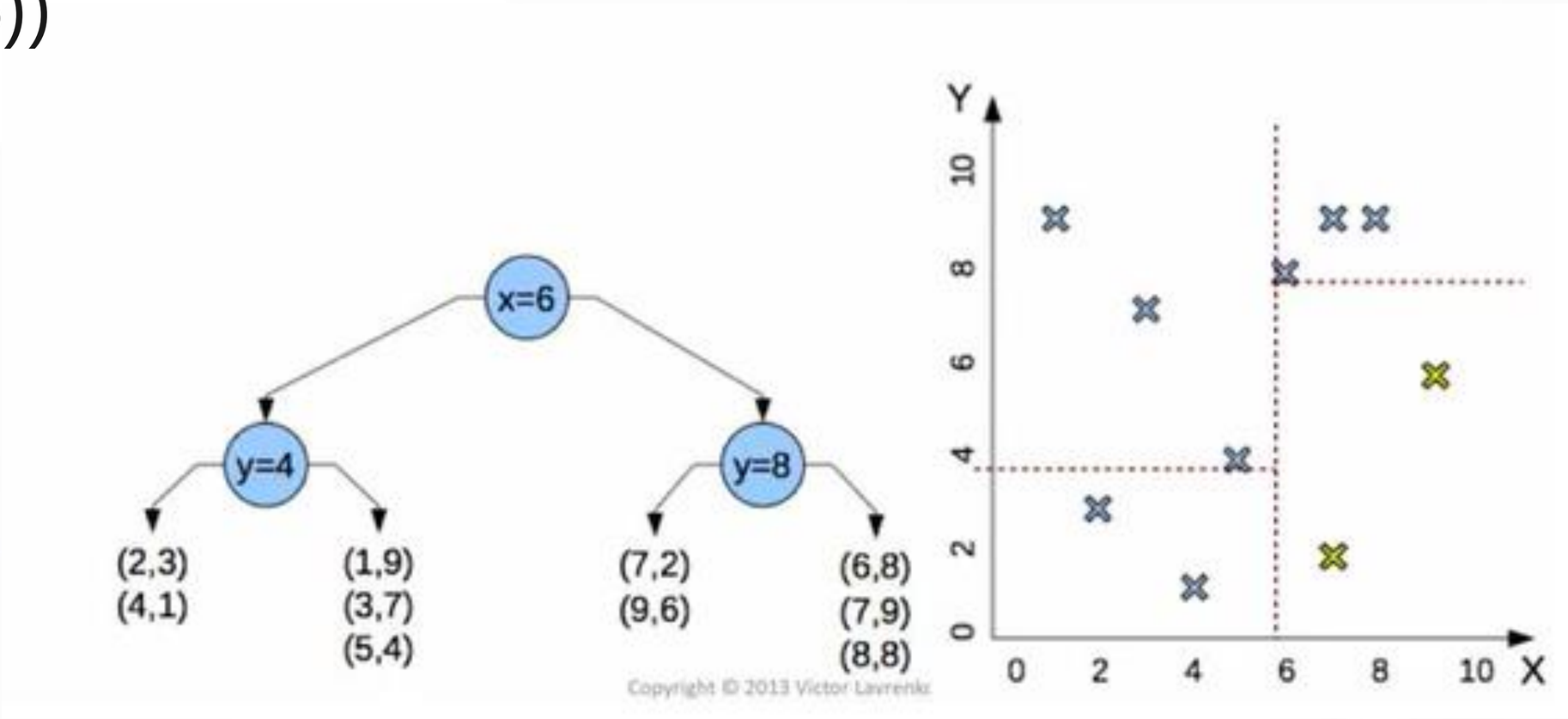
距离计算及闵氏距离

$$d_{12} = \sqrt[p]{\sum_{k=1}^{n} |x_{1k} - x_{2k}|^p}$$

- 其中，p是一个变参数：
- 当p=1时，是曼哈顿距离；
- 当p=2时，是欧氏距离；
- 当p→∞时，是切比雪夫距离。

K-D Tree

{(1,9),(2,3),(4,1),(3,7),(5,4),(6,8),(7,2),(8,8),(7,9),(9,6))}



bc_data.csv

来源

- 威斯康星乳腺癌数据库（Wisconsin Breast Cancer Database）”

主要属性

- ID：病例的ID；
- Diagnosis（诊断结果）： M 为恶性（212 个），B 为良性（357个）
- 细胞核的10个特征值：radius（半径）、texture（纹理）、perimeter（周长）、面积（area）、平滑度（smoothness）、紧凑度（compactness）、凹面（concavity）、凹点（concave points）、对称性（symmetry）和分形维数（fractal dimension）等。同时，为上述10个特征值分别提供了三种统计量，分别为均值（mean）、标准差（standard error）和最大值（worst or largest）。

bc_data.csv

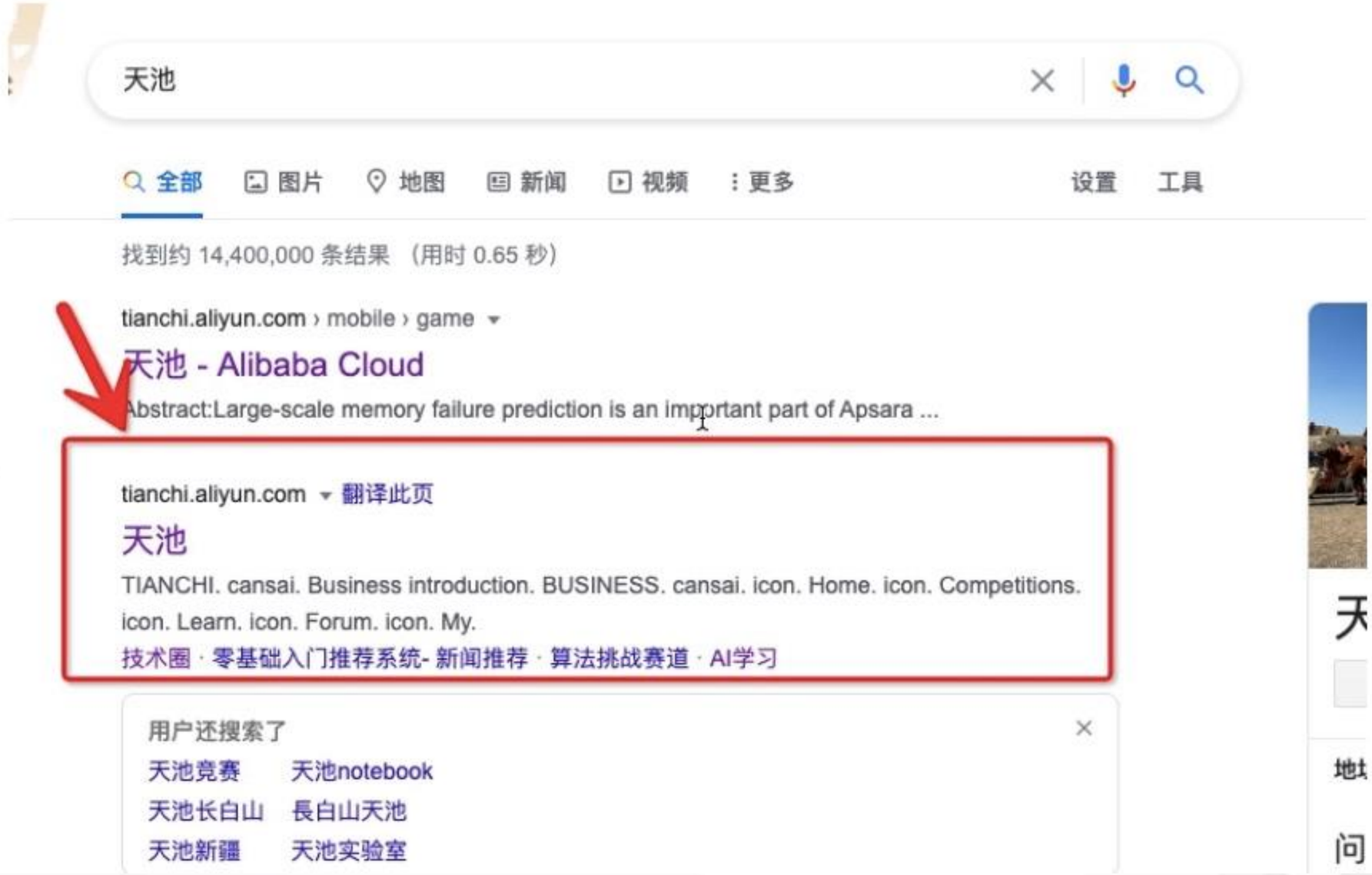
id	diagnosis	radius_mean	texture_mean	perimeter_	area_mean	smoothnes	compactne	concavity_	concave p	symmetry_	fractal_dim	radius_se	texture_se	perimeter_	area_se	smoothnes	c
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	4(		

TIANCHI 天池

Q&A

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

1）首先需要进入天池官网，大家打开浏览器，搜索 天池 ，找到 tianchi.aliyun.com即可访问进入天池官网；



2）在天池官网，将鼠标移到 天池学习 ，即可出现下拉列表，点击 天池读书会 ，即可进入天池读书会的页面。

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

3）在天池读书会页面，你可以对对应的读书会图书进行提问，优秀的提问还有机会获得赠书，还可以点击配套的训练营或者课程资源进入学习，还有点击实践代码获取读书会的项目实践的代码，跟着我一起进行项目实践和代码学习，同时还有很多其他的读书会，大家也可以观看举办过的读书会的回放，或者预约还没开始的读书会。

朝乐门 中国人民大学

直播主题 《Python编程：从数据分析到数据科学》

直播时间 2021年7月23日 20:00

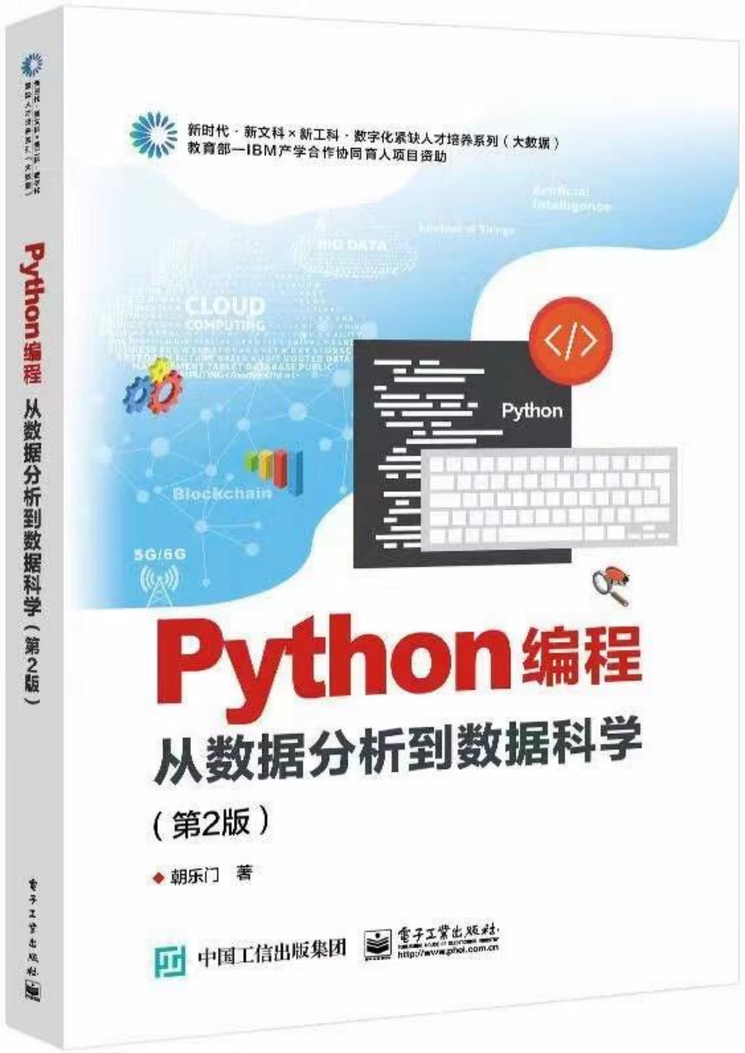
学习资料 Pandas教程

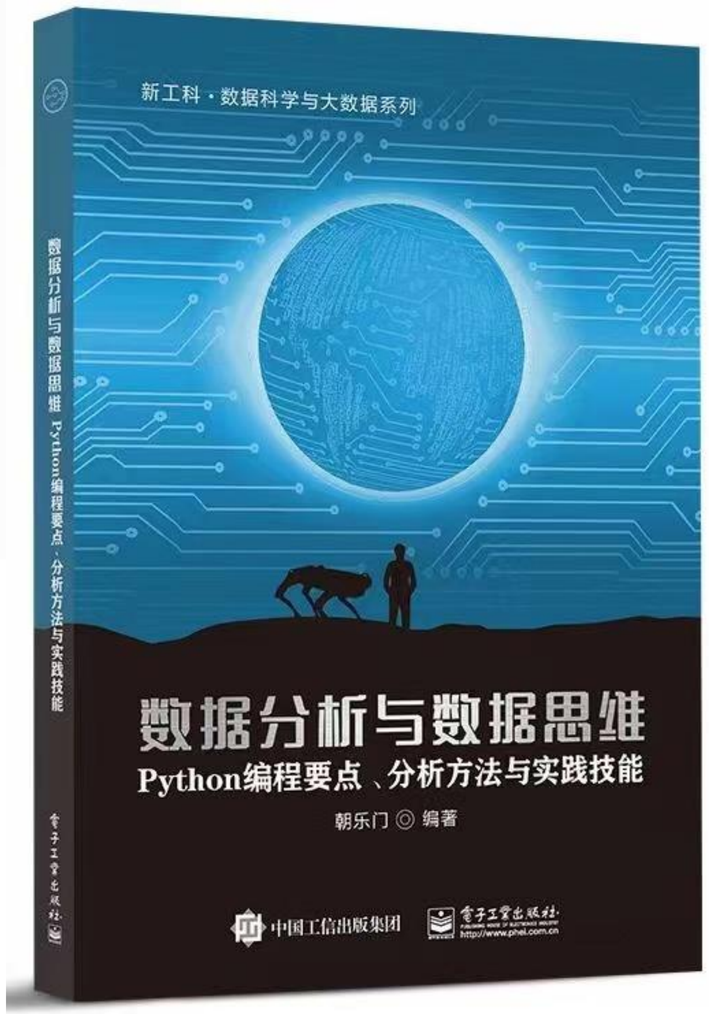
实践项目 威斯康星乳腺癌数据分析及自动诊断

提问 | 学习课程 | 购买地址 | PPT下载 | 实践代码 | 预约直播

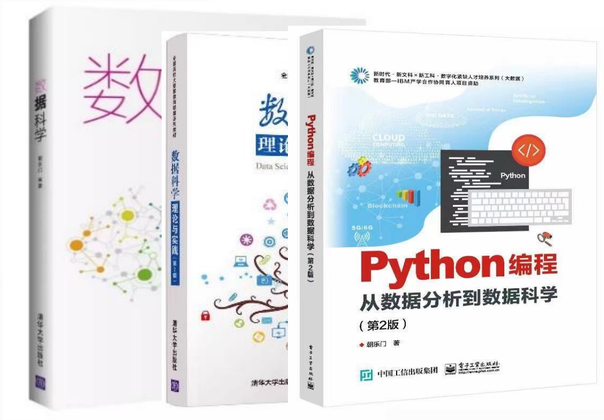
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

总结





直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub



参考书目

微信公众号

chaolemen

@

ruc.edu.cn

主讲人联系方式

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub