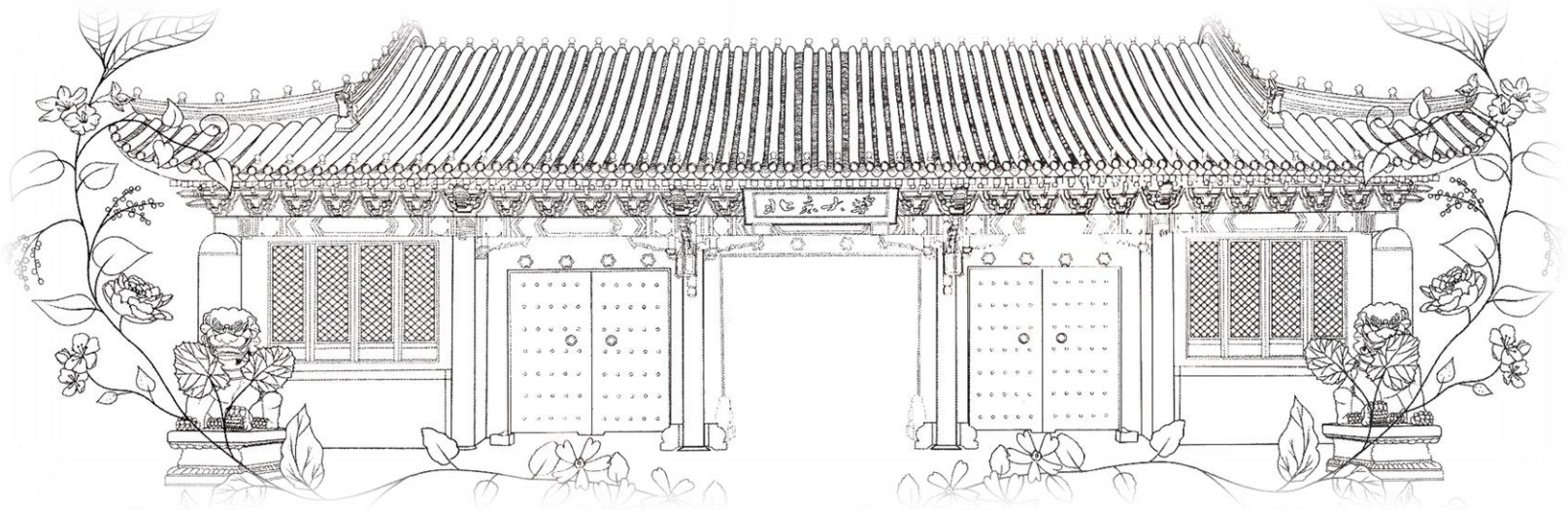
# 《Python数据分析》
# 课程总结与回顾
# （21年春季学期）



# 内容回顾

- 基础篇
  - 数据分析基础（Numpy, pandas等）
  - 数据可视化（matplotlib, seaborn）
- 应用篇
  - 探索式数据分析
  - 机器学习基础
  - 图像数据分析
  - 社会网络分析
  - 时间序列分析
  - 文本数据分析

# 基础篇：数据分析基础

- 数据分析能干什么？
- Python语法回顾
  - 四种数据结构（set, list, tuple, dict）
  - 两种表达形式（列表推导式、生成器表达式）
  - 两种函数（普通函数、匿名函数lambda）
  - 模块化

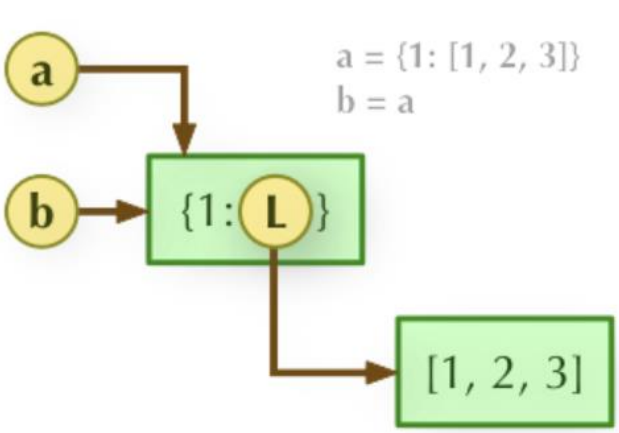
# 基础篇：数据分析基础

- 列表推导式和生成器表达式有何区别？
  - 括号的形式
  - 遍历的次数

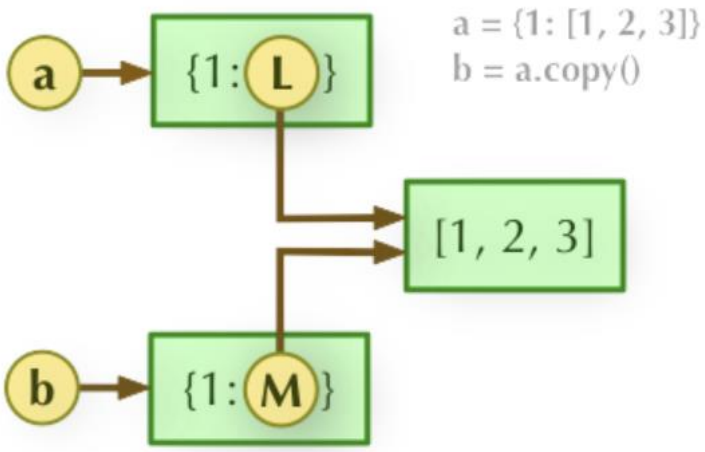
# 基础篇：数据分析基础

- 赋值、浅拷贝、深拷贝

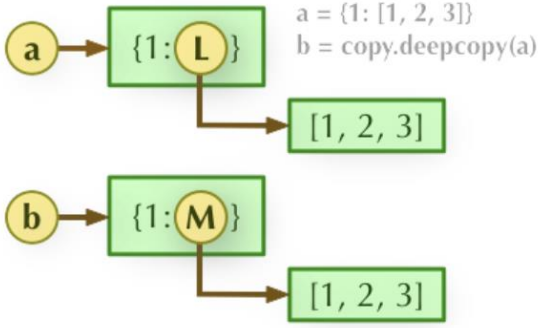
**b = a:** 赋值引用，a 和 b 都指向同一个对象。




**b = a.copy():** 浅拷贝，a 和 b 是一个独立的对象，但他们的子对象还是指向统一对象（是引用）。




**b = copy.deepcopy(a):** 深度拷贝，a 和 b 完全拷贝了父对象及其子对象，两者是完全独立的。



# 基础篇：数据分析基础

- 匿名函数

```
sum = lambda arg1, arg2: arg1 + arg2
print("运行结果：", sum( 10, 20 ))
print("运行结果：", sum( 20, 20 ))
```

# 基础篇：数据分析基础

- Numpy简介
  - 创建与打印数组（np.array()）
  - 基本运算（+、-、两种乘法、通用函数）
  - 索引、切片和迭代
  - 数组的形状操作、分割和组合

# 基础篇：数据分析基础

- pandas基本数据结构的生成
  - Series：传list或dict进去
  - DataFrame：dict套list、dict套dict
- 索引是不可变的，但可以重新索引reindex()，填补方式有ffill()和mfill()
- 从坐标轴删除条目：drop()，注意有inplace()

# 基础篇：数据分析基础

- pandas索引、选择和过滤
  - Series：用整数下标索引切片、标签切片(包含end)索引
  - DataFrame：
    - df[‘列名’], df[[‘列名1’,’列名2’]]
    - df[行整数下标或下标切片]
    - 通过轴标签：df.loc[label], df.loc[:, label], df.loc[label_1, label_2]
    - 通过整数下标：df.iloc[where], df.iloc[:, where], df.iloc[where _1, where _2]

# 基础篇：数据分析基础

- pandas算术和数据对齐
  - NA值会传播
  - add, sub, div… (fill_value可以设置)
- DataFrame和Series之间的操作
  - 默认地：DataFrame和Series间的算术运算Series的索引将匹配DataFrame的列，并在行上扩展
- 函数应用和映射：
  - Series：map方法
  - DataFrame：apply用于某一行或列、applymap用于每一个元素

# 基础篇：数据分析基础

- 排序
  - sort_index() vs. sort_values()，默认升序
- 排名
  - rank()方法，可以指定行或列

# 基础篇：数据分析基础

- pandas描述性统计
  - 汇总：count(), describe(), max()…
  - 唯一值：unique(), value_counts()…
  - 成员判断：isin()

# 基础篇：数据分析基础

- pandas缺失值处理
  - dropna()：默认剔除所有包含缺失值的行
    - 参数可调：how='all'只剔除全部NAN的行，thresh设置阈值
  - fillna()：可对每列填充不同值，可指定原地修改
    - 注意inplace, axis, method, value等参数
  - isnull(), notnull()

# 基础篇：数据分析基础

- 数据读写
  - read()简单顺序读取, readline()每次读取一行, readlines()按行读取所有内容
  - 使用pandas读文件
    - 表头、跳几行、缺失值…
  - 使用pandas写文件
    - to_csv, to_json, …

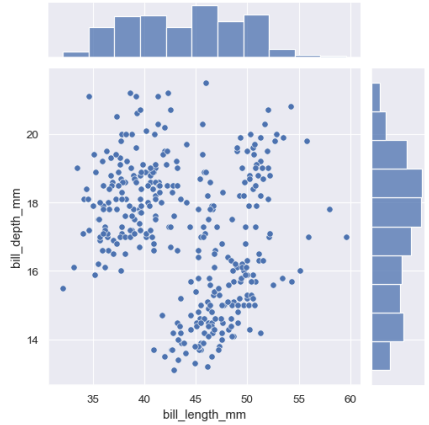
# 基础篇：数据分析基础

- 数据清洗
  - 去重
  - 利用函数或映射进行转换
- 数据聚合和分组
  - Groupby()：最简单的分组法、使用字典或Series分组、使用函数分组、使用索引级别分组

# 基础篇：数据可视化

- Matplotlib
  - 基本函数，如plot, scatter, xlim, xlabel, grid, annotate, text, title, legend
  - 统计图形的绘制：柱状图（+堆积, 多数据并列, 条形）、直方图、阶梯图、饼图、极线图、气泡图、箱线图
  - 完善统计图形：图例、标题、刻度（刻度定位器、刻度格式器）、标签、子图

# 基础篇：数据可视化




- seaborn
  - 单变量分布可视化：直方图、密度分布、累积分布…
  - 多变量间关系的可视化：散点图、折线图（含分面）、热力图、等高线、jointplot、散点矩阵图…
  - 定类变量的可视化：定类散点图、定类箱线图、定类小提琴图…
  - 可视化中的美学因素

# 应用篇：探索式数据分析

- 探索式数据分析 vs. 验证式数据分析

| EDA | CDA |
| --- | --- |
| • No hypothesis at first<br>• Generate hypothesis<br>• Uses graphical methods (mostly) | • Start with hypothesis<br>• Test the null hypothesis<br>• Uses statistical models |

- 一个变量的分析
- 两个变量的分析
- 三个或三个以上变量的分析

| 比较项目 | 参数检验 | 非参数检验 |
| --- | --- | --- |
| 检验对象 | 总体参数 | 总体分布和参数 |
| 总体分布 | 正态分布 | 未知 |
| 数据类型 | 连续数据 | 连续数据或离散数据 |
| 检验效能 | 较高 | 较低 |

- 探索式数据分析 vs. 验证式数据分析
- 一个变量的分析
  - 描述性统计：最值、均值、百分位数、众数
  - 分布
    - 表格或柱状图
    - 直方图
    - 密度分布
    - 累积分布
    - 正态分布如何检验：数值法、图示法、统计检验法（SW检验、AD检验）
  - 对于同一个变量的多组样本
    - 参数检验
    - 非参数检验
- 两个变量的分析、三个或三个以上变量的分析

# 应用篇：探索式数据分析

- 探索式数据分析 vs. 验证式数据分析
- 一个变量的分析
- 两个变量的分析
  - 散点图
  - 相关系数
  - 分组
- 三个或三个以上变量的分析

# 应用篇：探索式数据分析

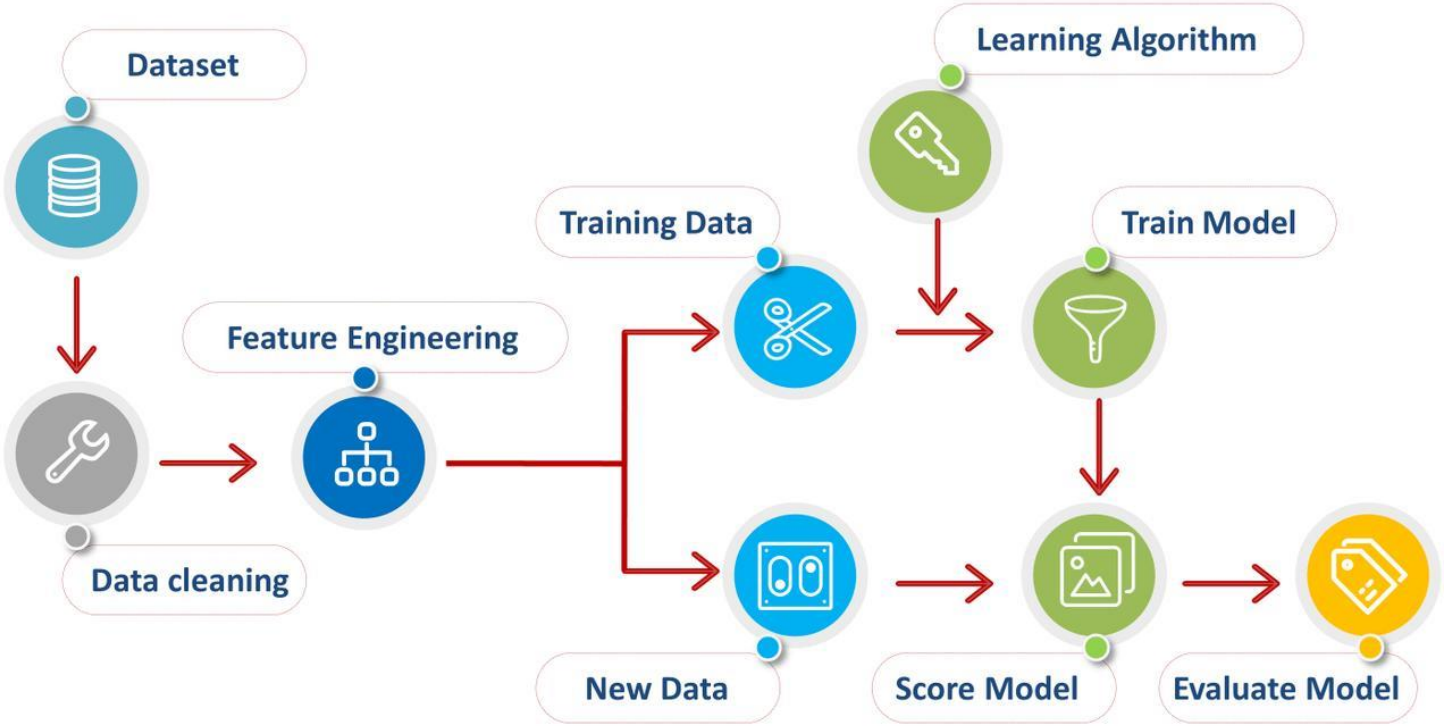
- 探索式数据分析 vs. 验证式数据分析
- 一个变量的分析
- 两个变量的分析
- 三个或三个以上变量的分析
  - 只能用于3个变量：珍珠图/气泡图、热力图等
  - 可以用于3个或以上变量：
    - 两两分别看（相关系数图、散点图矩阵等）
    - 因子分析
    - 聚类分析
    - 判别分析
    - 回归分析

# 因子分析的步骤

- 分析KMO和巴特球形检验
  - 分析KMO值；如果此值高于0.8，则说明非常适合进行因子分析；如果此值介于0.7~0.8之间，则说明比较适合进行因子分析；如果此值介于0.6~0.7，则说明可以进行因子分析；如果此值小于0.6，说明不适合进行因子分析
  - 如果Bartlett检验对应*p*值小于0.05也说明适合进行因子分析。
- 描述因子提取情况和方差解释率等
  - 特征值（Eigenvalue）>1的因子一般可以保留。
  - 描述总共提取的因子个数；分析每个因子旋转后的方差解释率和累积总共方差解释率。
- 分析loading载荷系数值
  - 通过因子载荷系数值（经验阈值0.3），分析出每个因子与题项的对应关系情况；结合因子与题项对应关系，对各个因子进行命名。

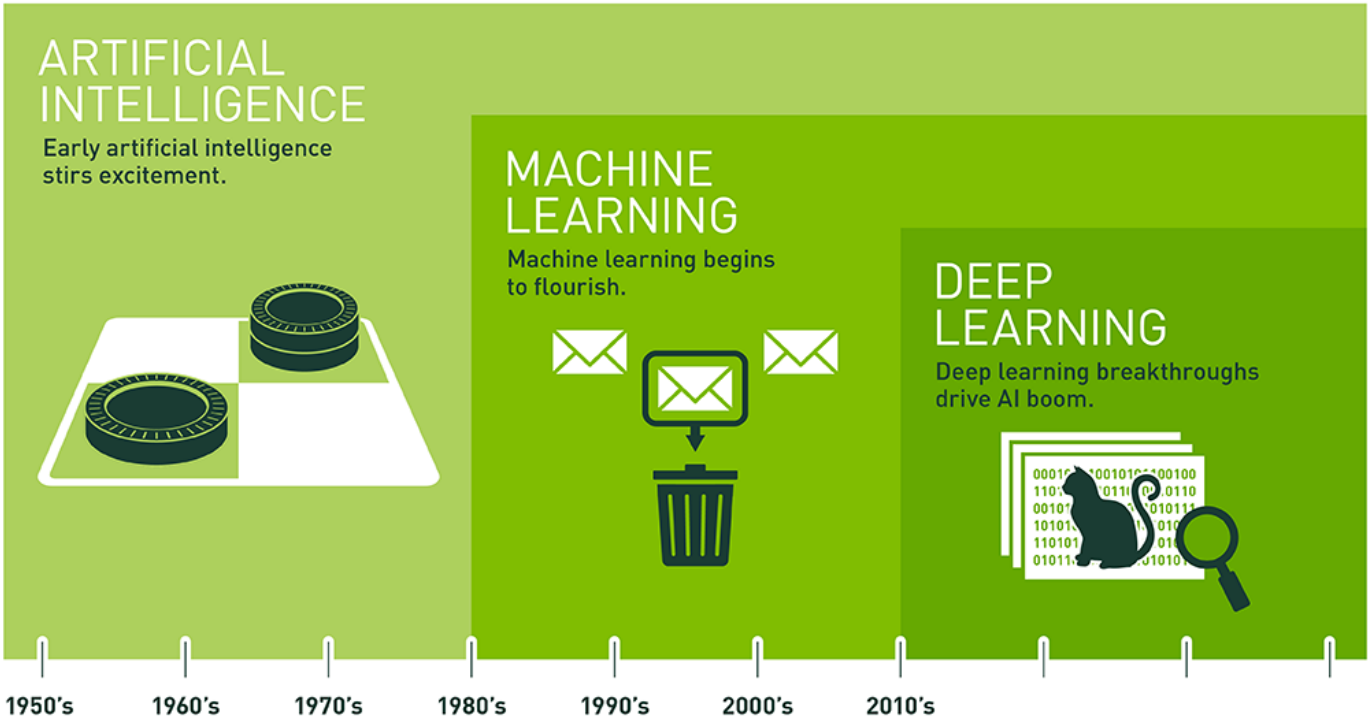
# 应用篇：机器学习基础

- 机器学习的概念。
- 三个关键词：算法、经验、性能
- 机器学习的基本流程。



# 应用篇：机器学习基础

- 人工智能、机器学习及深度学习的关系




Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

# 应用篇：机器学习基础

- 机器学习的经典问题
  - 概念、适用场景及常用算法
  - 分类、回归、聚类
- 利用scikit-learn进行预处理
  - 标准化、归一化、离散化、标称特征编码
- 利用scikit-learn进行机器学习
  - 分类、回归、聚类
  - 模型评价指标
- 关联规则挖掘
  - 支持度、置信度

# 应用篇：图像数据分析

- 图像基础知识
  - 图像处理技术：
    - 形成、存储和显示
    - 图像处理
  - 图像处理技术路径
  - 图像的种类：
    - 位图、矢量图
    - 彩色图、灰度图
  - 图像分辨率
  - 位图的表示方式

# 应用篇：图像数据分析

- 图像数据基本操作
  - 图像读写
  - 图像切片
  - 改变图像大小
  - 图像旋转
  - 对比度调节
  - 图像颜色直方图
- 高级任务
  - 边缘检测
  - 基于神经网络的图像识别
  - 卷积神经网络的原理与典型结构

# 应用篇：社会网络分析

- 社会网络的表示：边列表、邻接矩阵
- 社会网络的重要理论：六度分割理论
- 社会网络的常用指标：
  - 微观层面：中心性（点度中心性、中介中心性、接近中心性）、PageRank…
  - 中观层面：聚类系数…
  - 宏观层面：直径、密度…

# 应用篇：社会网络分析

- 社会网络中的一些现象：偏好连接性、同质性、传递性、核心-边缘结构
- 社会网络分析的应用
  - 链路预测
  - 排名
  - 社区探测
  - 分类/聚类
  - 恢复能力
  - 结构分析
  - 信息级联
  - 传染病模型

# 应用篇：时间序列分析

- 时间序列数据（Time series data），是一批按照时间先后顺序排列的统计数据
- 截面数据（Cross-section data），是一批发生在同一时间截面上的数据
  - 工业普查数据、人口普查数据、家庭调查数据等
- 面板数据（Panel data），是时间序列数据与截面数据的合成体。
  - 1978-1999年我国各省市城镇居民消费结构的调查资料

# 应用篇：时间序列分析

- Python中时间和日期的处理
- 时间序列基础
- 日期的范围、频率和转化
- 时间区间和区间算术
  - 时间区间（Period）和时间戳（Timestamp）有何差别？
- 时间序列的重采样
  - 上采样：插值
  - 下采样：Groupby
  - 其他采样
- 移动窗口函数

# 应用篇：文本数据分析

- Python中文本/字符串的处理
  - 通用序列操作
  - 字符串格式化与转义
  - 字符串函数（方法）
  - 正则表达式
- 文本数据分析及其任务
  - 词性标注、词义标注、分词/切词（中文：最大匹配法）、信息抽取、自动文摘、信息检索、情感分析、表示学习、机器翻译、人机对话
  - 难点：知识体系缺乏 -> 歧义

# 期末考试

- 时间：6月28日14:00-16:00
- 地点：待定
- 形式：闭卷考试
- 题目：基础篇（约30分）+应用篇（约70分），简答题为主，含10分代码题，给代码备忘录 Cheet sheet
- 要求：《北京大学本科考试工作与学习纪律管理规定》
  - http://www.dean.pku.edu.cn/web/rules_info.php?id=8

所有作业必须在规定上课日期的课前提交（如上课时间为某天下午 15:10，则必须在当天 15:09 前提交到指定位置）。除遇不可抗力（不包括时间管理不善、课程冲突、数据或文档丢失等问题），如作业迟交在 24 小时以内，总分扣除 20%；迟交在 48 小时以内，总分扣除 40%；迟交在 72 小时内，总分扣除 60%；迟交在 96 小时内，总分扣除 80%；迟交 96 小时以上，该次作业不计入总分。

# 小组作业

- 5月24日15:09前：提交小组选题给助教
- 6月7日：课堂汇报
  - 每组10min，额外≤5min的Q&A，小组全体（线下）成员都要上台
  - 教室：7-8节：二教301，10-11节：二教304
  - 请务必确认汇报人的电脑可以正常连接投影
- 7月3日（周六）晚上11:59
  - 上交小组作业全部内容（小组报告、PPT、代码[如有]等），每组只需要组长提交即可，注明组号

# 小组作业：远程同学的汇报

- 如果组内有至少一位线下同学，由线下同学展示
- 如果组内全部为线上同学：
  - 使用“腾讯会议”远程连线

# 欢迎对本课程提出建议！

- ☺

谢谢！