# 需求点：

* 数据的在线获取
  + 数据源？
  + It will source its data online, but should optimize the use of downloaded data, to avoid excessive network traffic. （压缩文件下载，本地解压？）(分析后认为，比较好的方法是，下载数据之前，先检查本地文件，如果本地文件恰好有这家公司（company\_symbol.csv）的同时，其中的内容又刚好涵盖查询时间段，就从这个文件中截取数据，否则再下载新数据)
  + You should also provide the ability to source data for any company.（爬虫？）
* 历史数据调取及描述性统计、可视化
  + for any company the user selects
  + 特定时期的特定股票描述性统计信息
  + 优化后的可视化of the raw data, but also of transformations, such as moving averages
* 指定日期收市价预测
  + basic modelling (such as regression)
  + The user specifies the modelling period (i.e. the training data);
  + They also specify a date for which they require a prediction;
  + A linear model is built, using the specified period;
  + The prediction is produced, along with the model’s RMSE and R2 value (co-efficient of determination).
  + Any other prediction approaches you may choose to implement (e.g. non-linear regression models) will add credit to your project.
* 交互界面
  + The program should provide a comprehensive but easy to use and intuitive interface (text or graphical).

# 实现策略

* 数据源
  + 10/31更新：经查询，使用pandas-datareader可以在线获取股票数据，直接返回dataframe，非常方便，所以不再使用爬虫方法
    - Data\_source\_pdr()
  + 可更新的Company list，采用pandas-datareader自带的get\_nasdaq\_symbols方法，返回symbols dataframe，以更新文件
    - Data\_source\_getsymbols()
  + Company list：Downloaded as companylist.csv
  + 爬虫获取股价历史信息
    - Yahoo finance 提供自当日起一年内的股价信息
    - <https://finance.yahoo.com/quote/NVDA/history?p=NVDA>
    - 主键为Companylist.Symbol
* 历史数据调取及描述性统计、可视化
  + 用户传入公司名，实时下载数据，计算描述性统计信息
  + 可视化
    - Raw time-series 原始时间序列
    - Linear trend lines 股价折线图
    - Moving Averages(e.g. MA(n), with user-selectable n); ？
    - Weighted Moving Averages？
    - Moving Average Convergence/Divergence (MACD);？
    - And so on （详细指标可能会加分）
  + (输出对应图形及报表可能会加分)
* 股价预测（收市价预测）
  + 自定训练集（时间段）及预测时间点
  + 线性回归模型， 输出RMSE及R2 (co-efficient of determination).（必须），使用其他模型加分
  + （输出对应报表可能会加分）
* 交互界面
  + 主菜单设计 第一印象得分（加分项：日期，问候语，tips）
  + 完整流程：数字导航器，返回上级菜单，返回主菜单
  + 异常输入处理与报错 用户友好性测试
  + 程序运行前后的文件管理 清除？或下载文件的实时更新（同名股价文件覆盖）

# 项目进度控制

* 截止日期: 2020/11/29
  + 源代码 + PDF报告
* 起始日期：2020/10/30
* 项目阶段
  + 10/30~10/31：需求分析及开发框架
  + 11/1~11/5：历史数据及预测模块
  + 11/5~11/10：可视化模块及预测模块优化
  + 11/10~11/15：接口设计及流程优化，UI优化
  + 11/15~11/18：测试阶段
  + 11/18~11/20：使用手册及开发历程说明

# 开发日志

* 10/30：数据抓取
  + 爬虫实现
    - data\_source\_selenium(company\_symbol)
    - 事实上只需要获得下载URL即可，但是被Yahoo Finance动态网页隐藏了 至少直接抓取HTML内容并不完全
    - Selenium 实现爬取动态网页
    - 速度慢
  + 10/31：数据抓取
  + 已改使用pandas-datareader
  + 部分数据获取部分异常测试
  + 编写了dt\_verify模块，用于处理时间
    - Dt\_format(date)：检查日期字符串格式
    - Dt\_compare(start\_date, end\_date)：检查前后日期顺序
  + 11/12: 截止至，描述性统计及可视化已完成
    - 开始开发预测模型