

# 大数据分布式计算与教学



中 央 财 经 大 学  
统 计 与 数 学 学 院  
李 丰

feng.li@cufe.edu.cn

# 目录

1 教学背景

2 实验环境

3 案例教学

4 教学效果



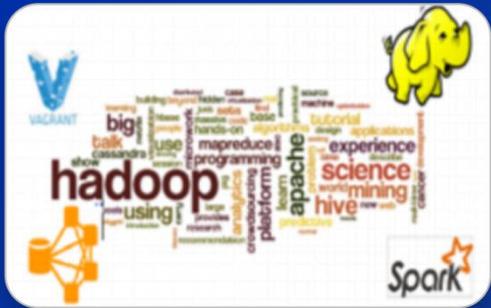
# 1

## 教学背景

## 大数据教学痛点



# 100+ 常用的 机器学习算法



# 10+ 流行的 大数据平台



# 18学时 教会学生 “大数据”

# 案例教学目标

从大数据到价值的实践能力

大数据决策的能力

案例  
教学

熟悉大数据分析平台

理解机器学习算法

应用深度学习工具

# 教学对象



**应用统计专业（大数据方向）**  
**硕士一年级**      **专业必修课**



**统计学专业**      **限选课**  
**本科三年级**



**非统计专业**      **选修课**  
具备计算机基础、对大数据分析有需求的



# 2

## 实验环境

# 教学环境

## 虚拟实验环境



### 大数据分布式计算平台

学生无需在自己电脑部署这些平台，便于案例能够迅速进行

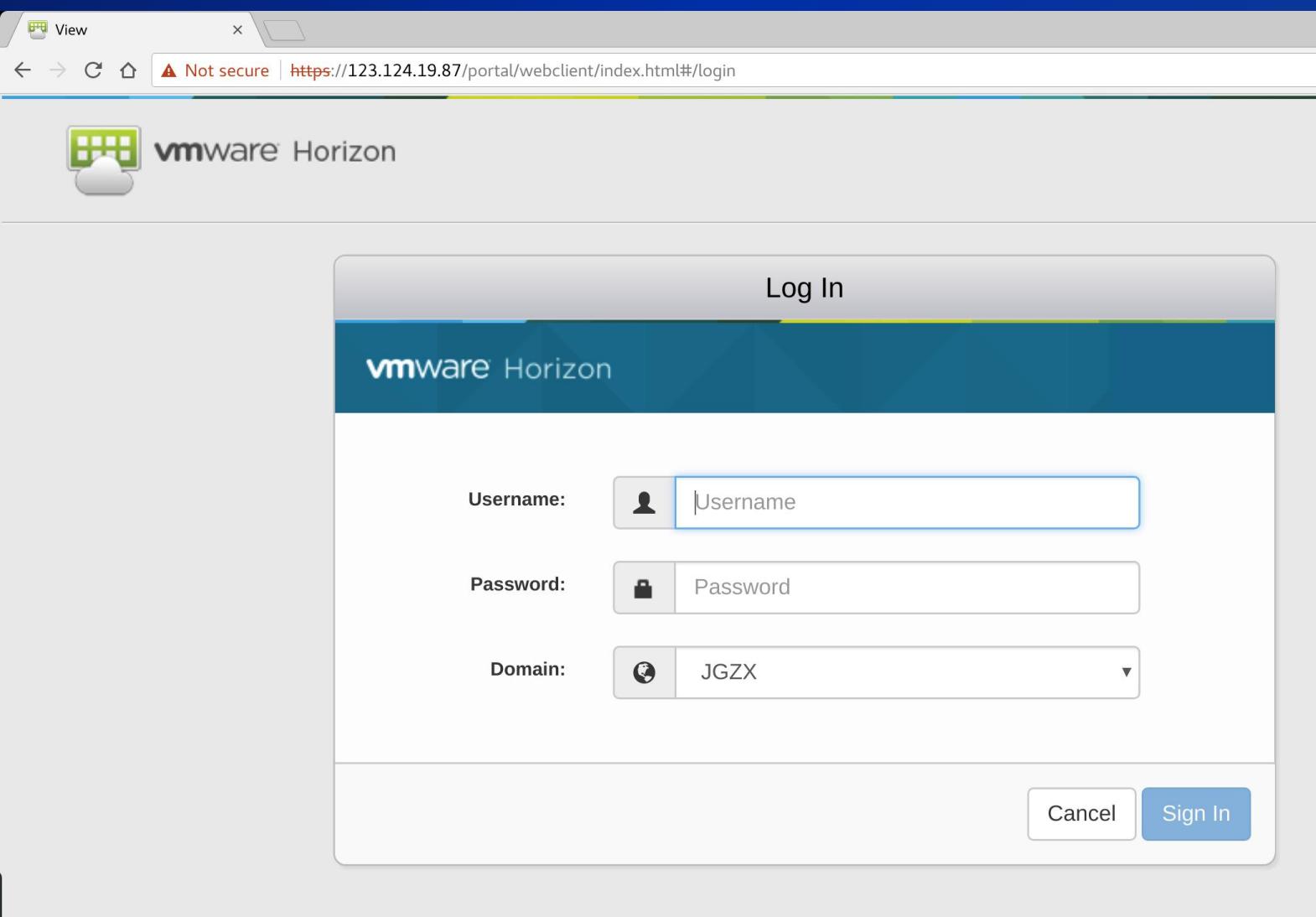
### 虚拟云桌面

通过客户端或者网页实时查看实验流程，系统实时反馈实验进度

### 前期开发案例库

学生通过完整的案例学习，优秀案例会补充到已有案例库

# 大数据分布式计算平台



随时随地任何终端登录个人云桌面

# 大数据分布式计算平台

VMware Horizon View Client

高性能计算共享桌面

123.124.19.87

高性能计算共享桌面

这台电脑 Mozilla Firefox

回收站 PuTTY

李丰 R i386 3.4.0

Acrobat Reader DC R x64 3.4.0

FileZilla Client Filezilla

PuTTY Configuration

Category: Session

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address) 192.168.113.164 Port 22

Connection type:  Raw  Telnet  Rlogin  SSH  Serial

Load, save or delete a stored session

Saved Sessions

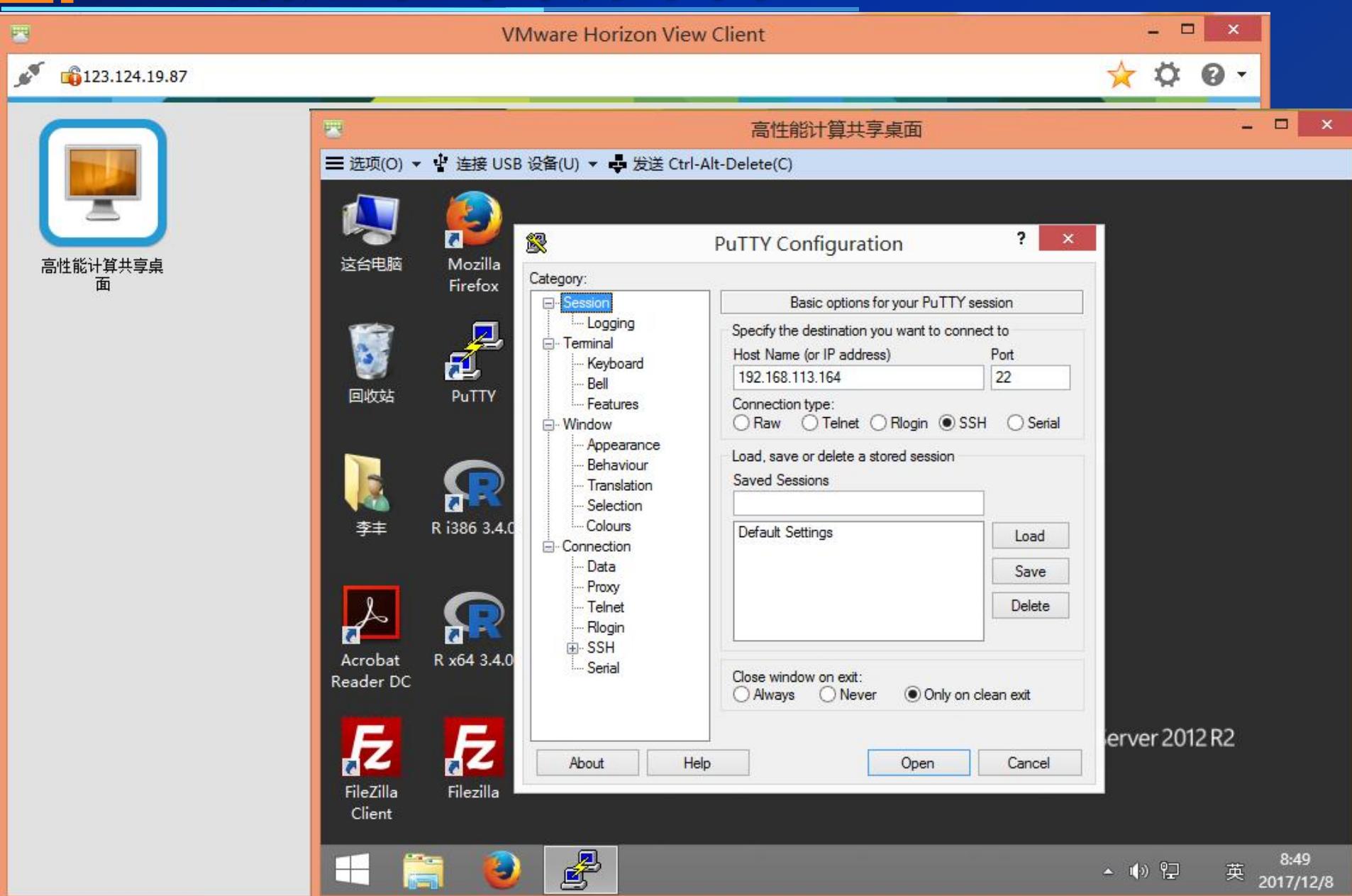
Default Settings Load Save Delete

Close window on exit:  Always  Never  Only on clean exit

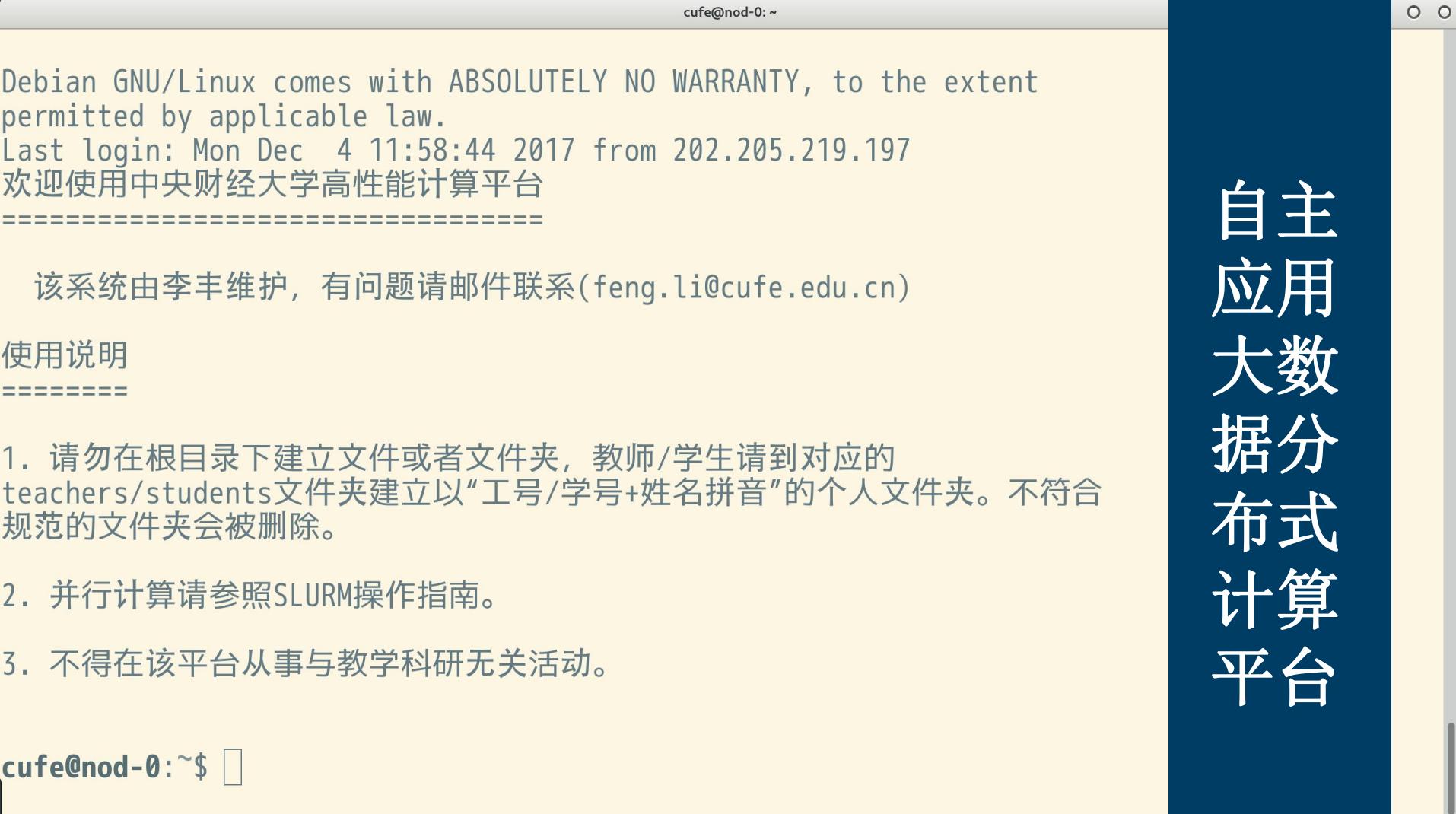
About Help Open Cancel

server 2012 R2

8:49 英 2017/12/8



# 大数据分布式计算平台



cufe@nod-0: ~

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent  
permitted by applicable law.

Last login: Mon Dec 4 11:58:44 2017 from 202.205.219.197

欢迎使用中央财经大学高性能计算平台

=====

该系统由李丰维护，有问题请邮件联系(feng.li@cufe.edu.cn)

使用说明

=====

1. 请勿在根目录下建立文件或者文件夹，教师/学生请到对应的  
teachers/students文件夹建立以“工号/学号+姓名拼音”的个人文件夹。不符合  
规范的文件夹会被删除。
2. 并行计算请参照SLURM操作指南。
3. 不得在该平台从事与教学科研无关活动。

cufe@nod-0:~\$ □

自主  
应用  
大数  
据分  
布式  
计算  
平台

# Hadoop交互界面

```
cufe@nod-0:~$ hadoop
Usage: hadoop [--config confdir] [COMMAND | CLASSNAME]
  CLASSNAME           run the class named CLASSNAME
or
  where COMMAND is one of:
    fs                  run a generic filesystem user client
    version             print the version
    jar <jar>            run a jar file
                          note: please use "yarn jar" to launch
                          YARN applications, not this command.
    checknative [-a|-h]   check native hadoop and compression libraries availability
    distcp <srcurl> <desturl> copy file or directories recursively
    archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
    classpath            prints the class path needed to get the
                        Hadoop jar and the required libraries
    credential          interact with credential providers
    daemonlog           get/set the log level for each daemon
    trace               view and modify Hadoop tracing settings

Most commands print help when invoked w/o parameters.
cufe@nod-0:~$
```

# 大数据案例选题三要素

海量真实数据 + 大数据平台 + 机器学习算法

## 数据来源



股票收益数据  
新闻文本数据  
房地产数据  
法律文书

## 数据获取工具



Scrapy  
BeautifulSoup  
Requests  
rvest

## 大数据平台



Hadoop  
Spark  
Mahout  
.....

## 计算机语言



Python  
R  
Scala  
.....

## 分布式机器学习算法



聚类分类  
主题模型  
分布式SVD  
Tensor  
.....

# 往年实验案例选题方向

选题方向	案例名称
自然语言处理	<ul style="list-style-type: none"> <li>• 基于法律文书的上市公司纠纷研究</li> <li>• 利用大数据分布式计算从海量文本数据中发掘新闻情绪与股票收益的关联</li> </ul>
智能推荐算法实现	<ul style="list-style-type: none"> <li>• 亚马逊在线书店新书预测策略</li> <li>• 基于RHadoop的电影推荐系统的应用</li> <li>• 基于物品的协同过滤算法的推荐系统</li> <li>• 基于气候信息的零售商店药品销售与推荐</li> </ul>
信用风险度量与管理	<ul style="list-style-type: none"> <li>• P2P平台信用风险评估</li> <li>• 基于文本与财经数据的国有企业债务违约测度</li> </ul>
大数据聚类分析	<ul style="list-style-type: none"> <li>• 基于DSBSCAN对科技股票涨跌情况聚类</li> <li>• 基于MapReduce的股票聚类分析</li> <li>• 基于欧洲五大联赛2014-2015赛季数据的线性判别分析</li> <li>• 基于朴素贝叶斯分类法的电子邮件分类研究</li> <li>• 基于KNN模型的游戏舰队分类</li> <li>• 基于MapReduce对某大学女生支出行为的聚类分析</li> </ul>
.....	<ul style="list-style-type: none"> <li>• 我国零售业上市公司财务绩效研究——基于MapReduce的并行主成分分析</li> <li>• 运用全数据逻辑回归和 MapReduce并行计算方法研究消费者橙汁品牌购买偏好</li> <li>• 中国电影票房预测——KNN方法的MapReduce实现</li> <li>• 基于泊松回归及MapReduce并行计算的付费搜索广告研究</li> <li>• 应用商店APP应用流行趋势</li> <li>• Mahout机器学习研究红酒质量影响因素</li> <li>• 基于树的方法对白酒质量分类并行计算实现</li> <li>• .....</li> </ul>

# 实验教学评分标准

大项	子项	得分
案例是否有实际应用价值	是否为真实数据且有一定的新颖性	10
	是否是独自收集数据（非第三方数据）	20
案例实施过程是否完整	是否有完整易懂的案例操作过程	20
	是否有完整的结论和展示	20
案例是否具有复制和扩展性	案例代码是否具有可读性	10
	案例代码是否充分利用大数据分布式计算平台优势	20
	总分	100



# 3

## 案例教学展示

# 案例准备

回顾

分布式计算原理与基础

回顾

统计模型并行计算特点

掌握

线性回归的并行分解

回顾

Hadoop Mapreduce原理

掌握

基于R的并行计算

熟悉

大数据分布式计算平台的使用

# 实验过程

熟悉大数据分布式平台并完成基本操作  
分组完成基于特定主题的过程设计

第一步

在大数据分布式平台上对所获数据进行数据清洗与结构化处理

第三步

利用大数据平台根据数据特征建立机器学习模型与模型评价

第五步

利用大数据分布式平台独立获取与所研究问题相关的原始数据收集

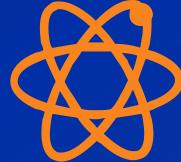
第二步

第四步

第六步

利用机器学习深度学习方法对所获复杂数据提取特征

总结并完成报告



# 典型实验案例

CASE STUDY

从海量文本数据挖掘  
新闻情绪与股票收益的关联

# 案例背景



新闻信息与股票波动有很强关系



487家中国公司在美国上市  
主营业务在中国



两地时差带来利用新闻预测股价的机会

# 案例目的

- 掌握传统回归方法在大数据和Hadoop分布式计算架构下的分布式计算方法。
- 根据所学大数据分布式计算知识以及大数据分布式平台，对数据的抓取和处理分析互联网新闻数据、并从中探寻新闻文本和股票收益的动态变化。

# 案例数据



数据来源为在美国上市的**487家**中国公司

- 为了实验叙述方便，以下内容以2014年在美国上市的阿里巴巴公司（NYSE:BABA）为例。
- 案例所附代码可以实现对所有487家公司的分析。



全部分析所涉及约**130万**条新闻正文

# 案例数据

- 所有在美上市中国公司名称及股票代码来自新浪财经
- 对应的股票日收益数据来自Yahoo财经
- 新闻文本数据来自财新网

**美股行情**

热门股票 实时行情

中国公司(488家)	科技类(100家)	金融类(34家)	医药/食品类(25家)	媒体类(8家)	汽车/能源类(27家)	制造/零售类(47家)	ETF(25家)
XNET	迅雷	15.43 +1.78 +13.040%	14.290 13.650 16.200	13.830 27.000	3.110 10	-- 0.00	
KNDI	康达车业	7.35 +0.55 +8.090%	6.850 6.800 7.400	6.800 9.800	3.500 4	-- 0.00	
OSN	震旦创新	2.45 +0.14 +6.060%	2.350 2.310 2.450	2.350 3.610	1.500 0	6.28 0.00	
DQ	大全新能源	56.95 +2.37 +4.340%	54.580 54.580 58.490	53.200 58.490	18.010 6	14.24 0.00	
RYB	红黄蓝	18.30 +0.60 +3.390%	17.250 17.700 19.200	17.000 31.800	15.560 5	-- 0.00	
CTEC	新兴佳	4.15 +0.12 +2.980%	4.000 4.030 4.190	3.960 4.250	2.700 0	-- 0.00	
CLDC	中海资源公司	2.84 +0.08 +2.890%	2.800 2.760 2.880	2.800 8.301	2.000 1	3.59 0.00	
IMOS	南茂科技	18.48 +0.43 +2.380%	18.320 18.050 18.690	18.320 21.990	13.940 158	369.60 0.00	
VIPS	唯品会	8.41 +0.18 +2.190%	8.170 8.230 8.660	8.150 15.490	7.790 50	17.16 0.00	
SINO	中环球船务	2.80 +0.06 +2.190%	2.740 2.740 2.800	2.700 5.500	2.230 0	28.00 0.00	
SMI	芯志国际	7.10 +0.14 +2.010%	7.200 6.960 7.210	7.040 9.142	4.490 60	18.21 0.00	
KGII	金墨珠宝	2.18 +0.03 +1.400%	2.180 2.150 2.190	2.160 2.500	1.000 1	2.56 0.00	
CAAS	中汽系统	5.09 +0.07 +1.390%	5.010 5.020 5.090	4.950 7.960	4.300 2	7.27 0.00	
CHT	华住酒店集团	107.53 +1.44 +1.360%	106.340 106.090 111.445	106.060 142.800	45.610 72	65.17 0.00	

**YAHOO FINANCE**

Search for news, symbols or companies

Finance Home Explore My Portfolio My Screensers Markets Industries Originals Events Personal Finance Technology

**Security**

**Fastest VPN for China**

Zero Logs, No Trace. Access Content from Anywhere - on Any Device. 24/7 Support.

S&P 500: 2,642.22 +5.38 (+0.20%) Dow 30: 24,231.59 -40.76 (-0.17%) Nasdaq: 6,847.59 -26.39 (-0.39%) Crude Oil: 57.94 -0.42 (-0.72%) Gold: 1,273.50 -5.39 (-0.41%) Silver: 16.31 +0.01 (+0.00%)

0 S. Markets open in 9 hrs 13 mins

Submit Your Questions! Live chat: Q&A on ACA 2018 open enrollment Our experts will answer your questions LIVE on Tuesday, December 5th, at 2 p.m. EST

What to watch in the markets: Monday, December 4 Read More >

Quote Lookup My Portfolio & Markets Customize Recently Viewed > Your list is empty.

# Step1 获得数据

**Hadoop Mapper:** 对于给定的关键字，按照以下步骤获取新闻全文

输入：关键字x；

输出：含有关键字x的

初始化：令页码数

步骤1：拼接链接

链接url.search；

步骤2：通过步骤

url.news和总页数

articles.content.

步骤3：通过步骤

容放入列表article

步骤4：令 $k \leftarrow k + 1$ ；

步骤5：重复步骤1到步骤4，直至 $k > pages.n$ ，跳出循环，输出结果。articles.content.list.

## 学生能力提升

- 掌握数据获取技术
- 结合Hadoop分布式计算平台实现大规模数据同步抓取

**Hadoop Reducer:** 收集所有输出并按照stdout输出到HDFS

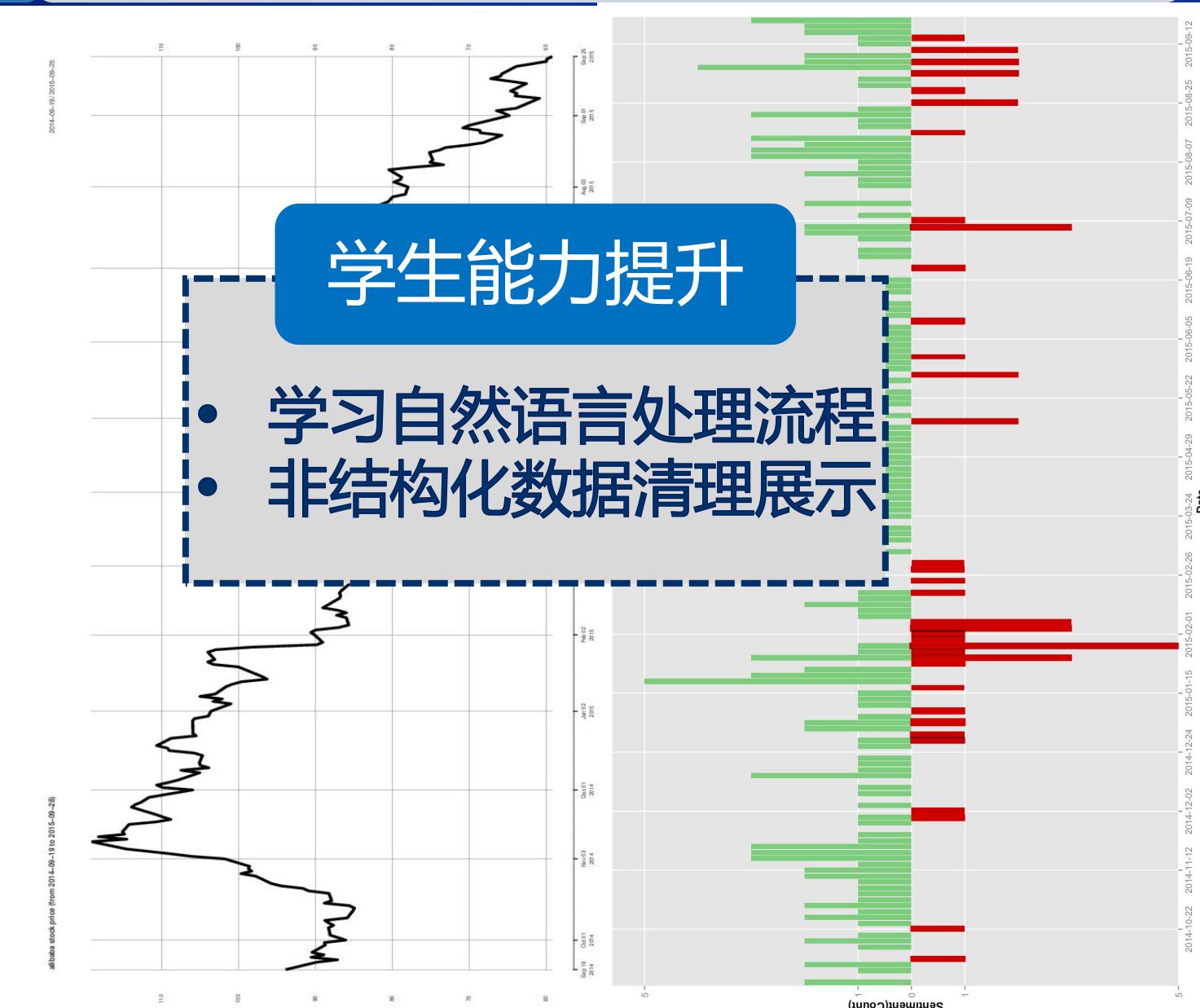
输出会以(key, value)的格式匹配

## Step2

## 海量文本数据新闻情绪量化

学生能力提升

- 学习自然语言处理流程
- 非结构化数据清理展示



# Step3 → 发掘新闻情绪与股票收益的关联

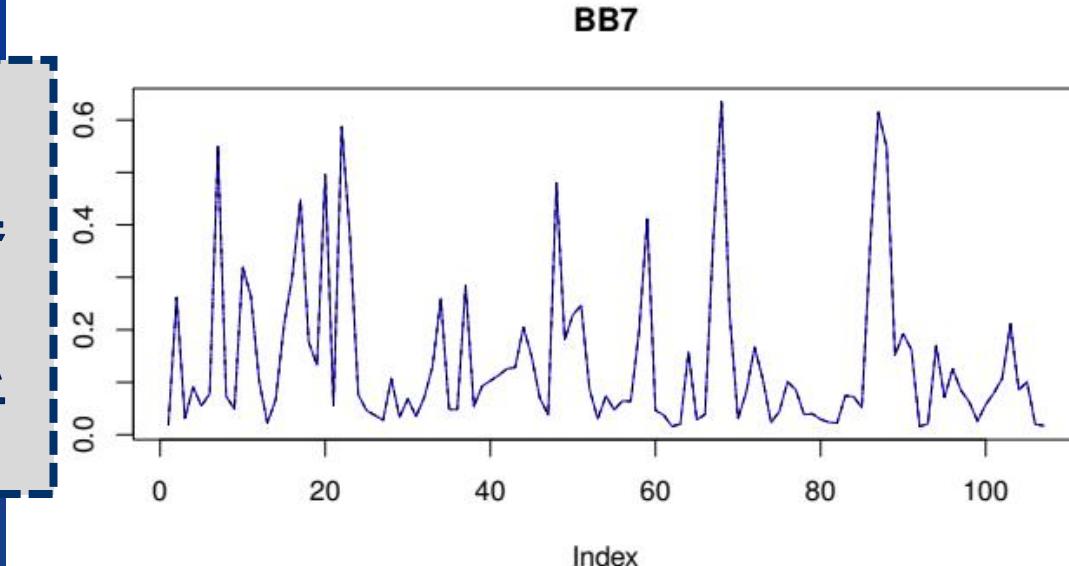
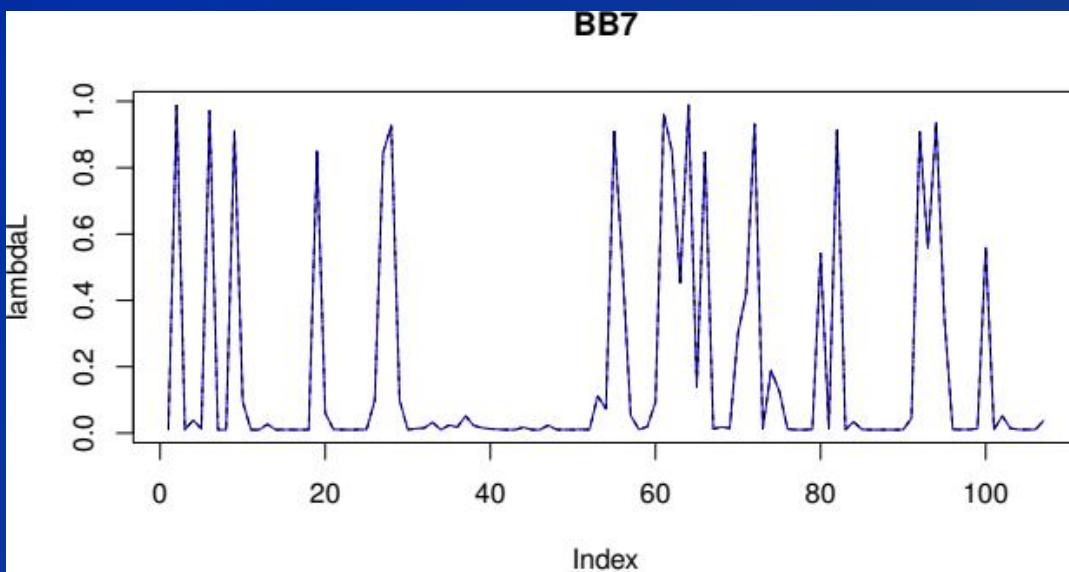
Poisson模型对487家新闻数据做正负情绪建模

$$P(P_i = p_i) = \frac{\lambda_i^{Pp_i}}{p_i!} \exp(-\lambda_i^P), p_i = 0, 1, 2 \dots; i = 1, 2, \dots 252.$$

贝叶斯变量选择识别影响股票发展的新闻指标

学生能力提升

- 大数据平台下的机器学习算法与模型
- 海量数据的关联分析



# 部分计算机代码

```
#!/usr/bin/env Rscript

#定义sigmoid函数
sigmoid = function(Retio)  ##系数数组
{
    return (1/(1 + exp(-Retio)))
}

#定义logistics回归函数，返回系数向量
logistic=function(data,a,b) ##a步数, b迭代次数
{
    n = dim(data)[1] ##行数
    p = dim(data)[2]-1 ##变量数量
    y = data[,p+1]
    x = data[,-(p+1)]
    ratio = runif(p,0,0.01) ##生成p个随机系数
    for(i in 1:b)
    {
        output = sigmoid(x%*%ratio)
        error = y - output
        ratio = ratio + a*t(x)%*%error
    }
    return(ratio)
}

##通过数据集训练系数
input=file("stdin","r") ##建立连接
data=readLines(input,n=1,warning=F)
data1=unlist(strsplit(data,","))
##以,分割, 转换为向量
p=length(data1)

while(length(data)<-readLines(input,n=100,warning=F))>0 ) ##每次读取100行
{
    fields=matrix(as.numeric(unlist(strsplit(data,""))),ncol=p,byrow=T)

    y=fields[,p]
    fields=cbind(1,fields[,-p],y)

    temp=logistic(fields,0.2,100)

    cat(temp, '\n')
}
```

# 案例结果

## 实验结果

- 探究了487家在美国上市的中国公司财经新闻的新闻情绪和股票收益的相依关系
- 以阿里巴巴为例解释了正面新闻与股票收益存在可度量的正向相关性，负面新闻与股票收益存在负向相关性
- 中国企业的负面新闻对其在美国上市股票具有很强的预测性

## 学生能力提升

- 通过大数据分布式平台对海量数据的获取、分析，了解机器学习算法与分布式平台结合解决实际问题的价值
- 真的“大数据”应用
- “生产—加工—价值”完整大数据工具链实现

# 4

## 实验案例教学效果

# 大数据实验教学反馈

“觉得世界好大，非常感谢老师！”

“让我对大数据的处理方法有了新的理解，为我提供了新的学习方向。”

“对Python和大数据有了初步的了解，希望以后可以多多开设这种课程。”

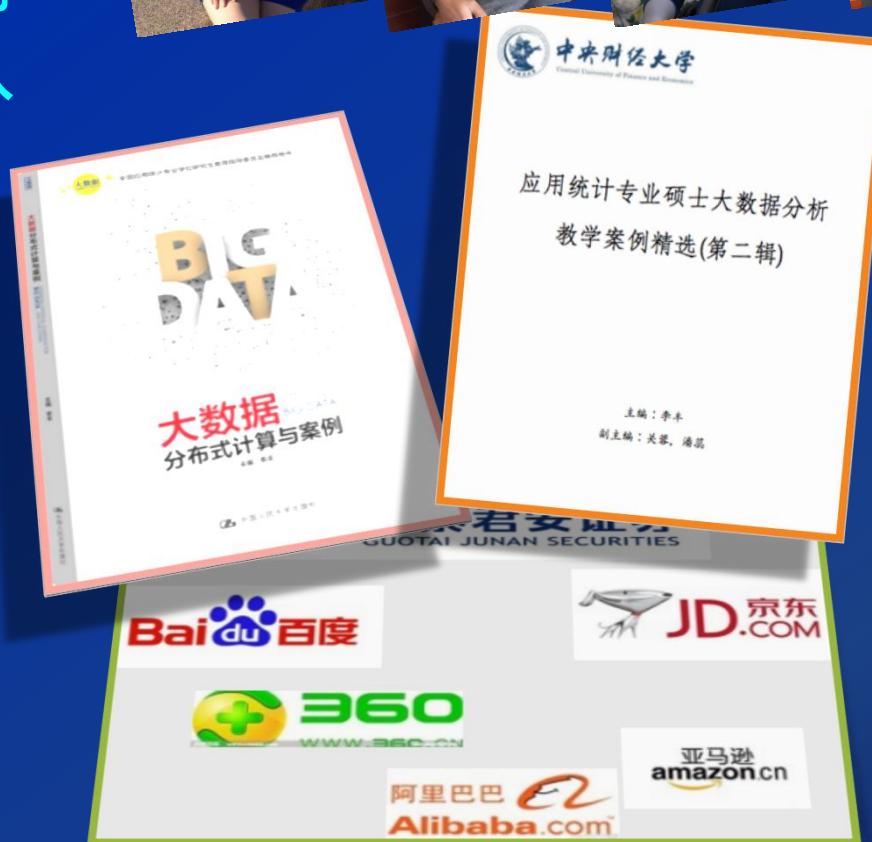
“感觉在学校能有这样的实践机会真的很惊喜，一开始自己对Python和数据分析只是比较好奇，现在越来越感兴趣了，即使课程结束应该自己也会继续学习相关方面的内容。”

“实验的选题非常热门，很有吸引力。课程时长为6周，这种短期课程的形式适合入门，今后有更深一层需要时，可以再自学。多媒体形式丰富、很炫酷，很有吸引力。”

# 大数据实验教学带来的成果

逐步形成中央财经大学大数据分析方向特色教学团队

《大数据分布式计算与案例》



《中央财经大学应用统计专硕大数据分析案例集》

150名大数据分析方向专业硕士

授之以渔而胜于鱼





---

请各位老师指正！

---

课程主页

<https://feng.li/distcomp/>

课程讲义与教学案例

<https://github.com/feng-li/Distributed-Statistical-Computing>