中山大學
SUN YAT-SEN UNIVERSITY

HYPERCOMP
Hyperspectral Computing Laboratory

Multivariate Statistics Analysis:

多元统计分析

Lecture 1： Basic Concept

Jun Li（李军）
School of Geography and Planning
Sun Yat-Sen University, Guangzhou, China
Mobile: 13922375250; Office: D307
E-mail: lijun48@mail.sysu.edu.cn; Webpage: http://www.lx.it.pt/~jun/

Examples

1) Establish the relationship between salary and demographic variables in population survey data(工资和其他因素的关系)

2) Identify the numbers in a handwritten zip code（手写字体的识别）

3) Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements（心脏病突发的影响因素）

4) Customize an email spam detection system（垃圾邮件过滤器）

5)Analyze the stock market（股市分析）

6) Classify the pixels in a LANDSAT image, by usage（遥感图像分类）

Example 1

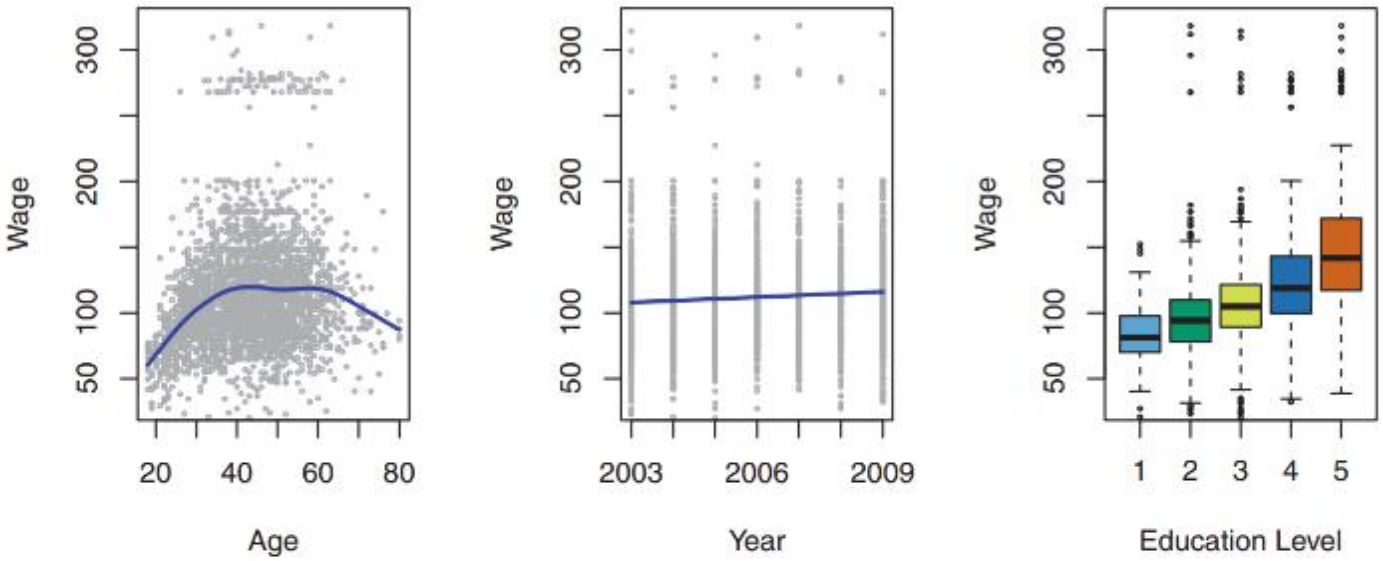


FIGURE 1.1. `Wage` *data, which contains income survey information for males from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009. Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

Examples

1) Establish the relationship between salary and demographic variables in population survey data（工资和其他因素的关系）

2) Identify the numbers in a handwritten zip code（手写字体的识别）

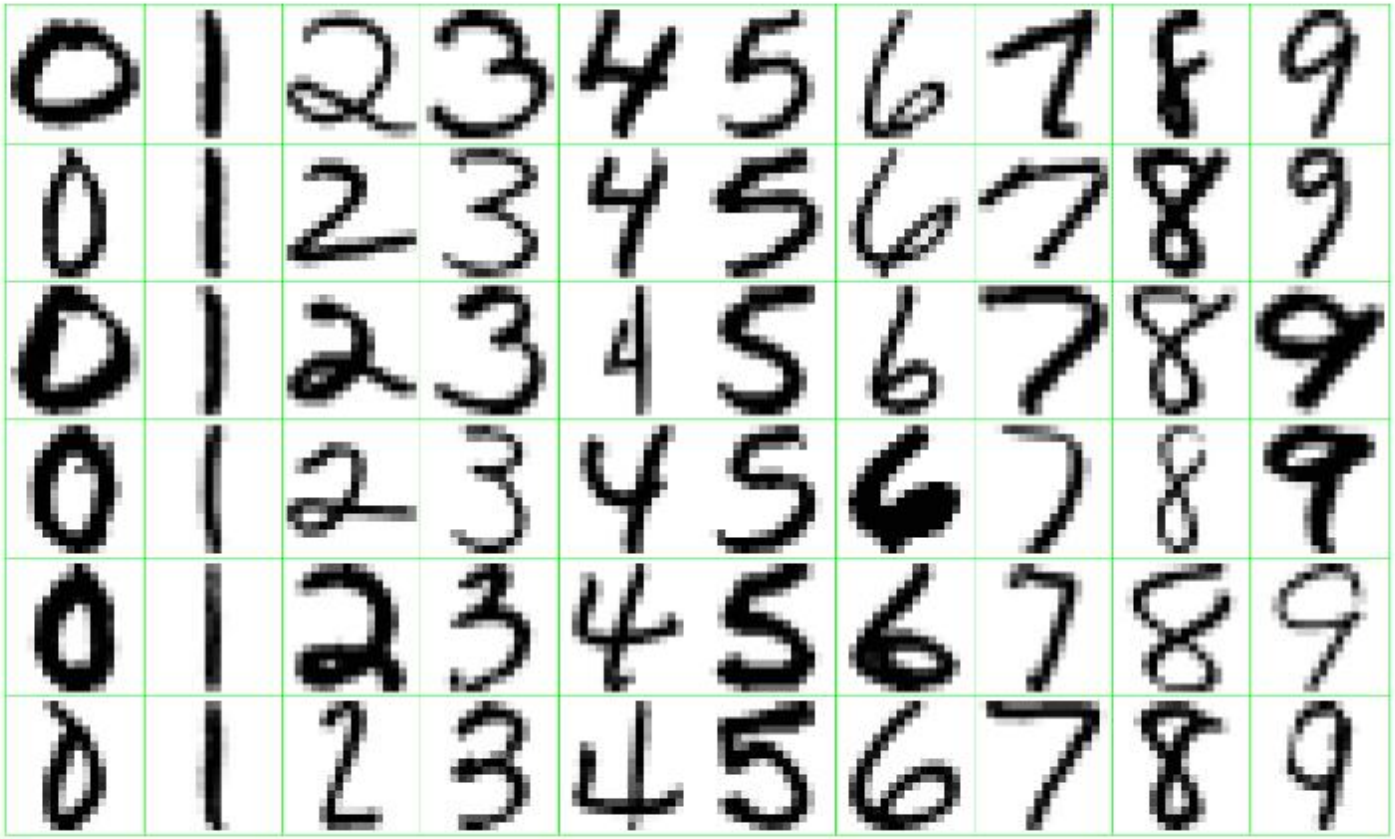
3) Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements（心脏病突发的影响因素）

4) Customize an email spam detection system（垃圾邮件过滤器）

5)Analyze the stock market（股市分析）

6) Classify the pixels in a LANDSAT image, by usage（遥感图像分类）

Example 2



Examples

1) Establish the relationship between salary and demographic variables in population survey data（工资和其他因素的关系）

2) Identify the numbers in a handwritten zip code（手写字体的识别）

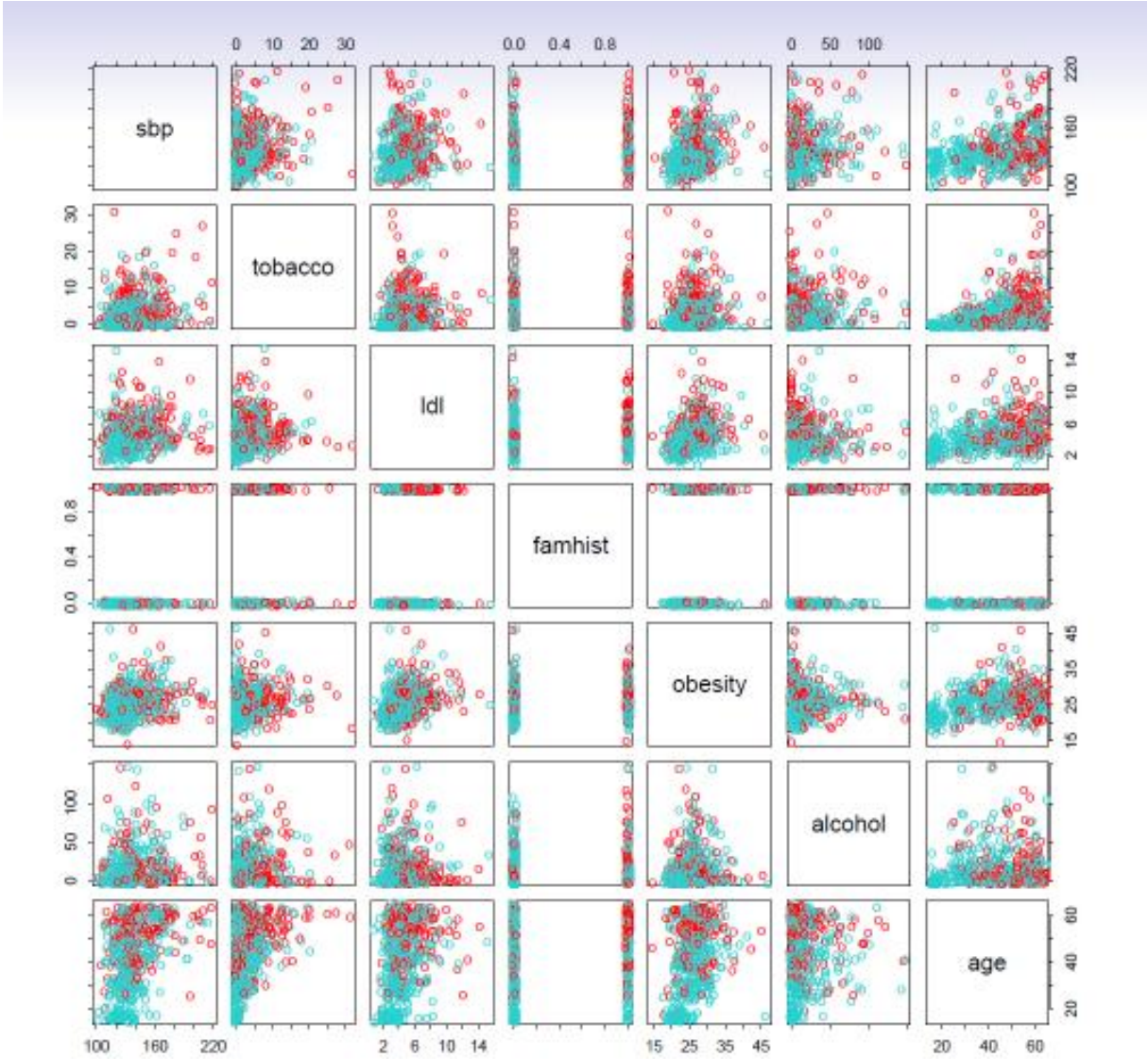
3) Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements（心脏病突发的影响因素）

4) Customize an email spam detection system（垃圾邮件过滤器）

5)Analyze the stock market（股市分析）

6) Classify the pixels in a LANDSAT image, by usage（遥感图像分类）

Example 3



Examples

1) Establish the relationship between salary and demographic variables in population survey data（工资与其他因素的关系）

2) Identify the numbers in a handwritten zip code（手写字体的识别）

3) Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements（心脏病突发的影响因素）

4) Customize an email spam detection system（垃圾邮件过滤器）

5)Analyze the stock market（股市分析）

6) Classify the pixels in a LANDSAT image, by usage（遥感图像分类）

Example 4

Spam Detection

- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam *and* email.

Examples

1) Establish the relationship between salary and demographic variables in population survey data（工资与其他因素的关系）

2) Identify the numbers in a handwritten zip code（手写字体的识别）

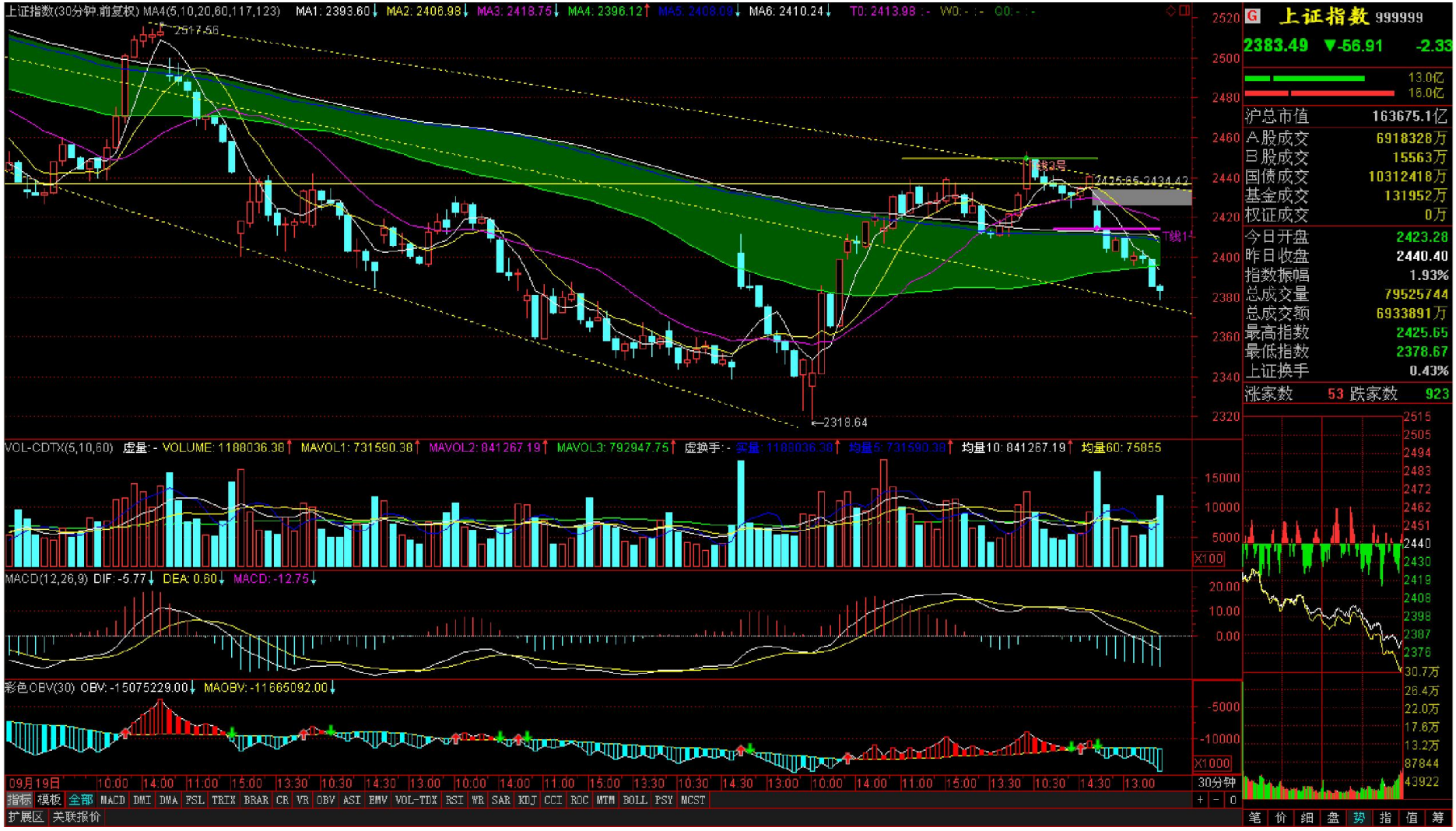
3) Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements（心脏病突发的影响因素）

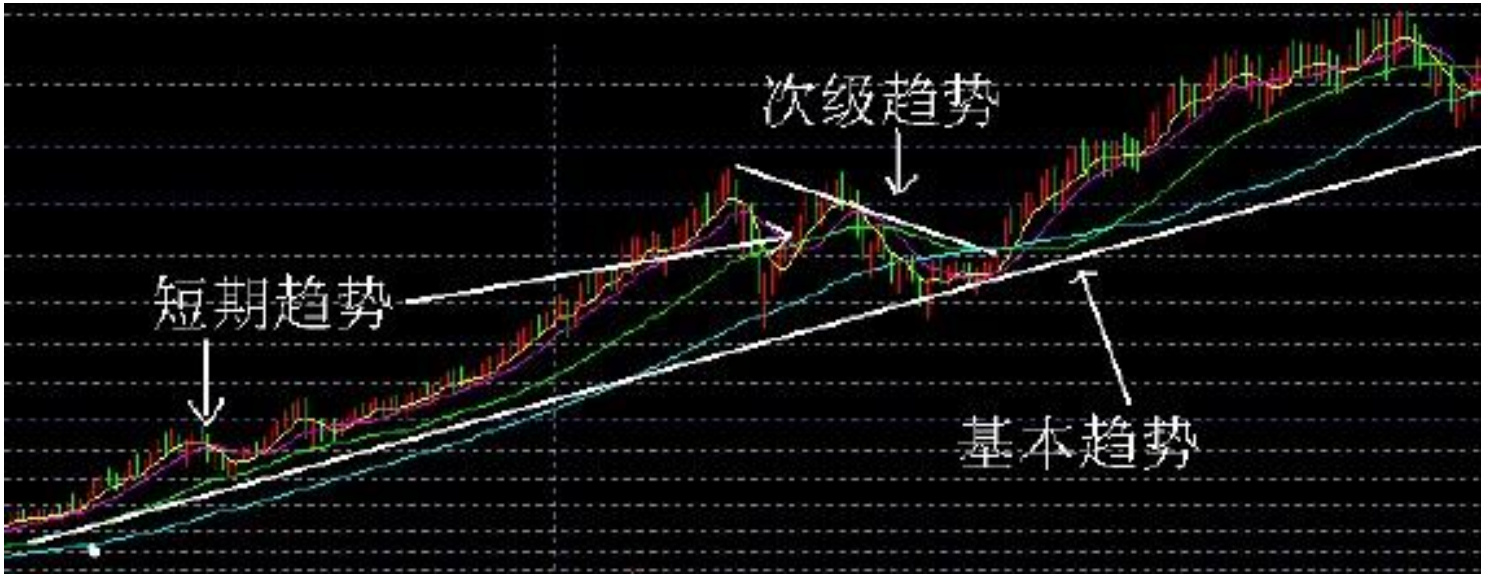
4) Customize an email spam detection system（垃圾邮件过滤器）

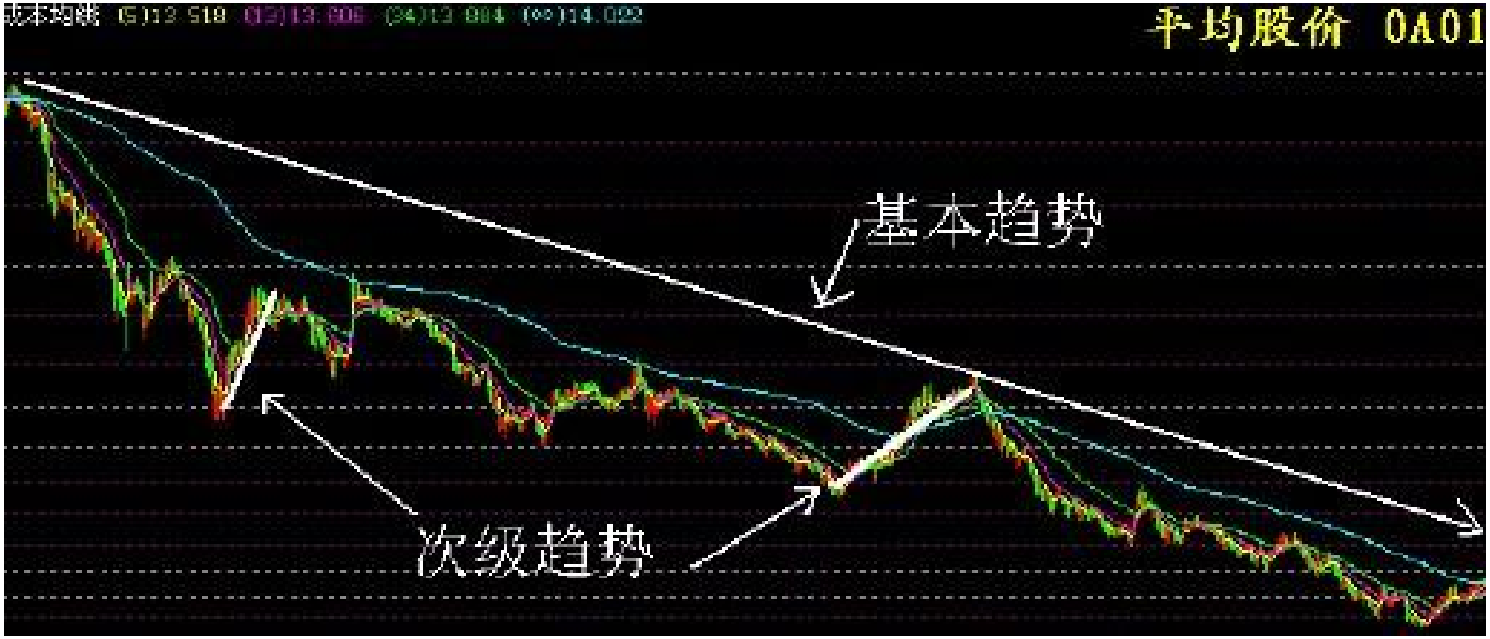
5)Analyze the stock market（股市分析）

6) Classify the pixels in a LANDSAT image, by usage（遥感图像分类）

Example 5


上证指数(30分钟,前复权) MA4(5,10,20,60,117,123) MA1: 2393.60 MA2: 2408.98 MA3: 2418.75 MA4: 2396.12 MA5: 2408.09 MA6: 2410.24 T0: 2413.98 W0: - Q0: -
2517.56
2425.85 2434.42
2318.64
VOL-CDTX(5,10,60) 虚量: - VOLUME: 1188036.38 MAVOL1: 731590.38 MAVOL2: 841267.19 MAVOL3: 792947.75 虚换手: - 买量: 1188036.38 均量5: 731590.38 均量10: 841267.19 均量60: 75855
MACD(12,26,9) DIF: -5.77 DEA: 0.80 MACD: -12.75
彩色OBV(30) OBV: -15075229.00 MAOBV: -11665092.00
09月19日
指标 模板 全部 MACD DMI DMA FSL TRIX BRAR CR VR OBV ASI EMV VOL-TDX RSI WR SAR KDJ CCI ROC MTM BOLL PSY MCST
扩展区 关联报价
上证指数 999999
2383.49 ▼-56.91 -2.33
13.0亿
16.0亿
沪总市值 163675.1亿
A股成交 6918328万
B股成交 15563万
国债成交 10312418万
基金成交 131952万
权证成交 0万
今日开盘 2423.28
昨日收盘 2440.40
指数振幅 1.93%
总成交量 79525744
总成交额 6933891万
最高指数 2425.65
最低指数 2378.67
上证换手 0.43%
涨家数 53 跌家数 923
笔 价 细 盘 势 指 值 筹


次级趋势
短期趋势
基本趋势


平均股价 0A01
基本趋势
次级趋势

Examples

1) Establish the relationship between salary and demographic variables in population survey data（工资与其他因素的关系）

2) Identify the numbers in a handwritten zip code（手写字体的识别）

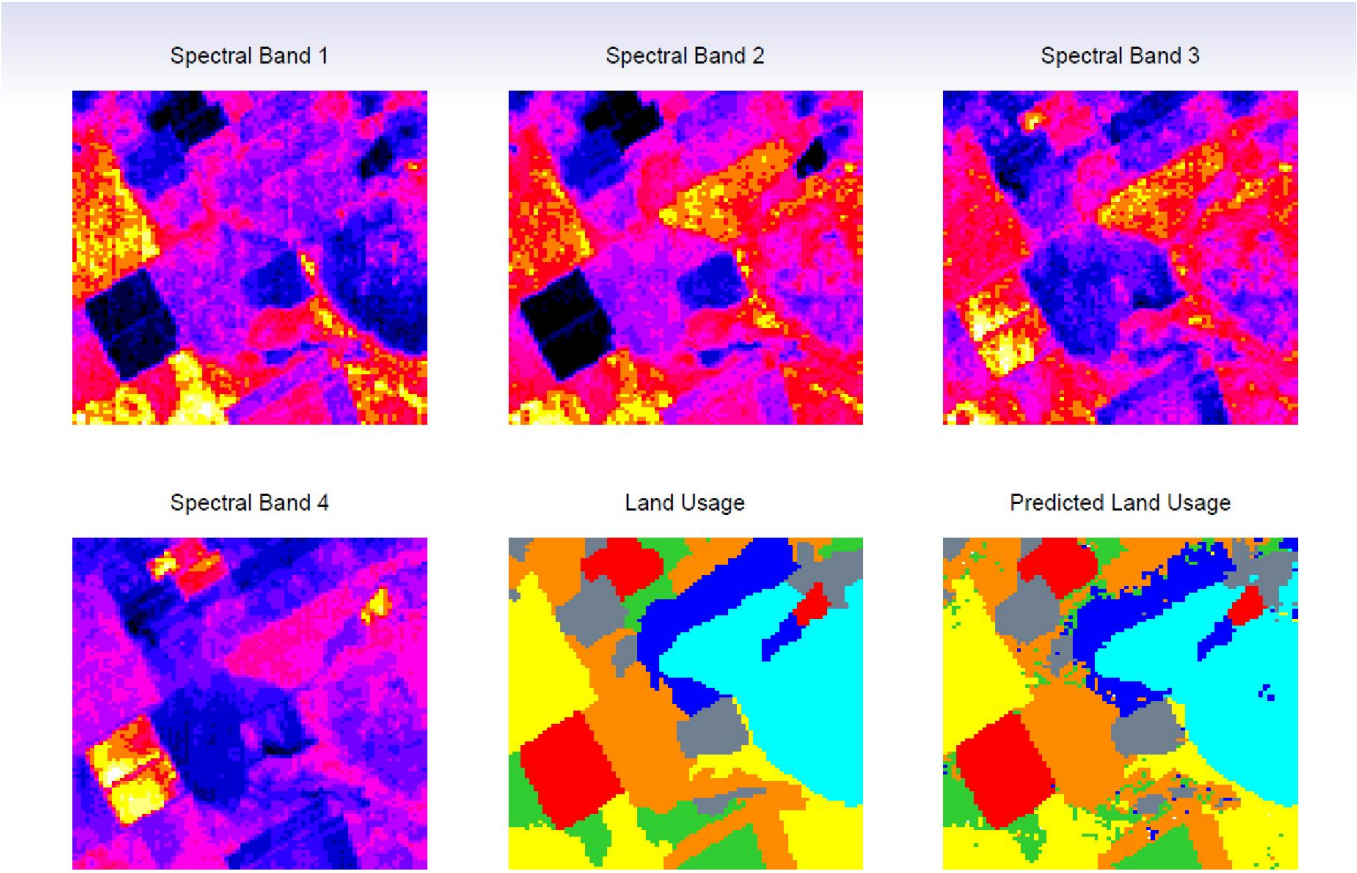
3) Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements（心脏病突发的影响因素）

4) Customize an email spam detection system（垃圾邮件过滤器）

5)Analyze the stock market（股市分析）

6) Classify the pixels in a LANDSAT image, by usage（遥感图像分类）

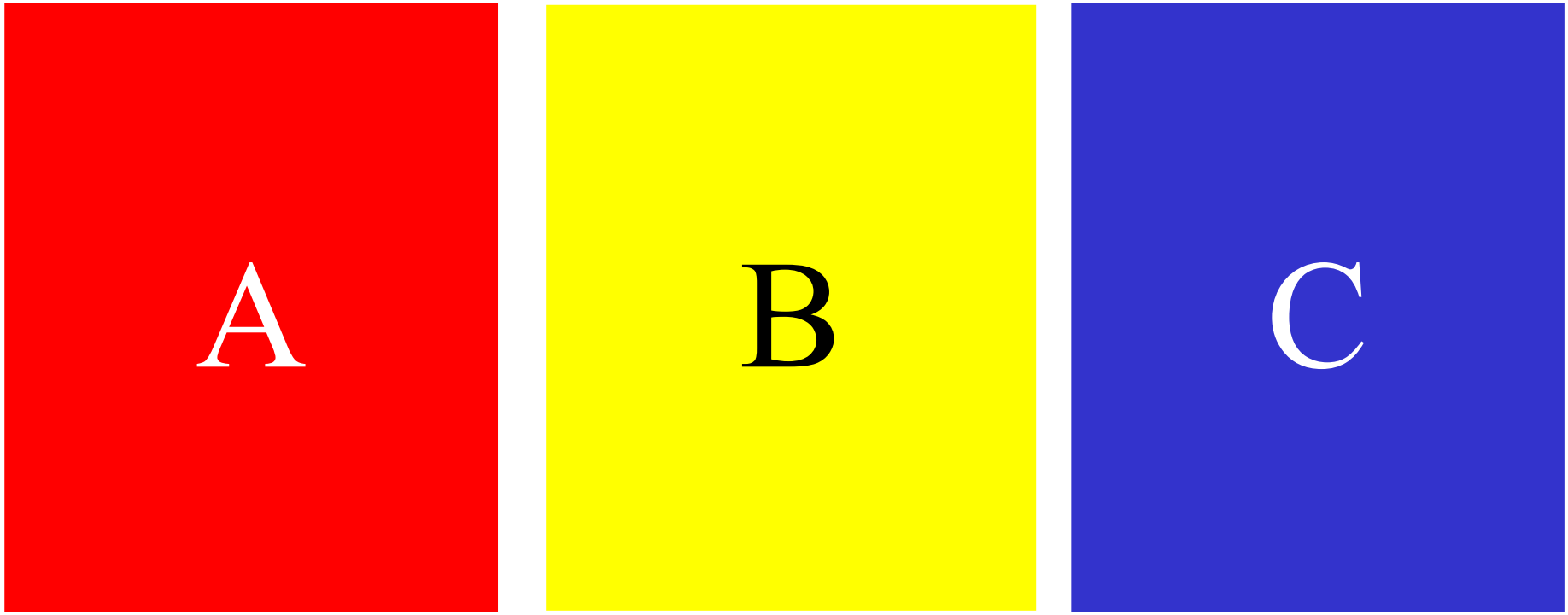
Example 6



$Usage \in \{red\ soil,\ cotton,\ vegetation\ stubble,\ mixture,\ gray\ soil,\ damp\ gray\ soil\}$

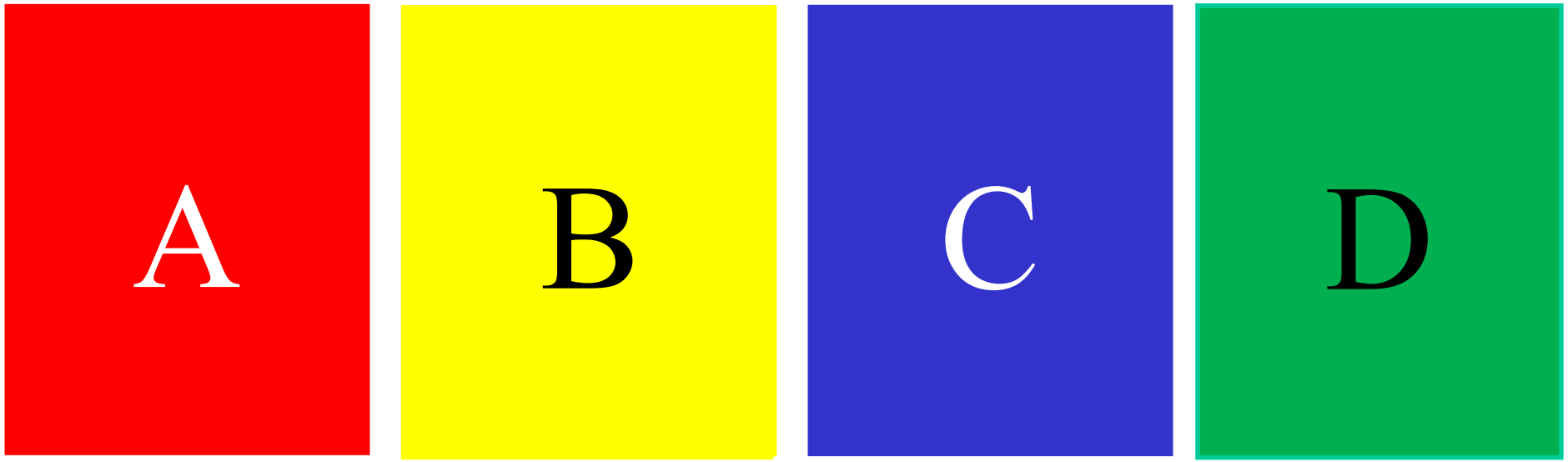
Discussion

其中一张后面有‘Good Luck’，请随机选一张，去掉一张空白的，请问是否更换选择？



Discussion

其中一张后面有‘Good Luck’，请随机选一张，去掉两张空白的，请问是否更换选择？



Discussion: More cards?

Discusssion !

Jun Li (李军)
School of Geography and Planning
Sun Yat-Sen University, Guangzhou, China
Mobile: 13922375250; Office: D307
E-mail: lijun48@mail.sysu.edu.cn; Webpage: http://www.lx.it.pt/~jun/