

# 反垃圾邮件系统核心解密

讲师(yasaka)陈老师

# 机器学习

- 机器学习理论主要是设计和分析一些让计算机可以自动学习的算法。
- 机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。
- 监督学习
  - 分类（做出单一决策）
  - 推荐（选择许多可能，并对其进行排序）
- 无监督学习
  - 聚类

# 机器学习

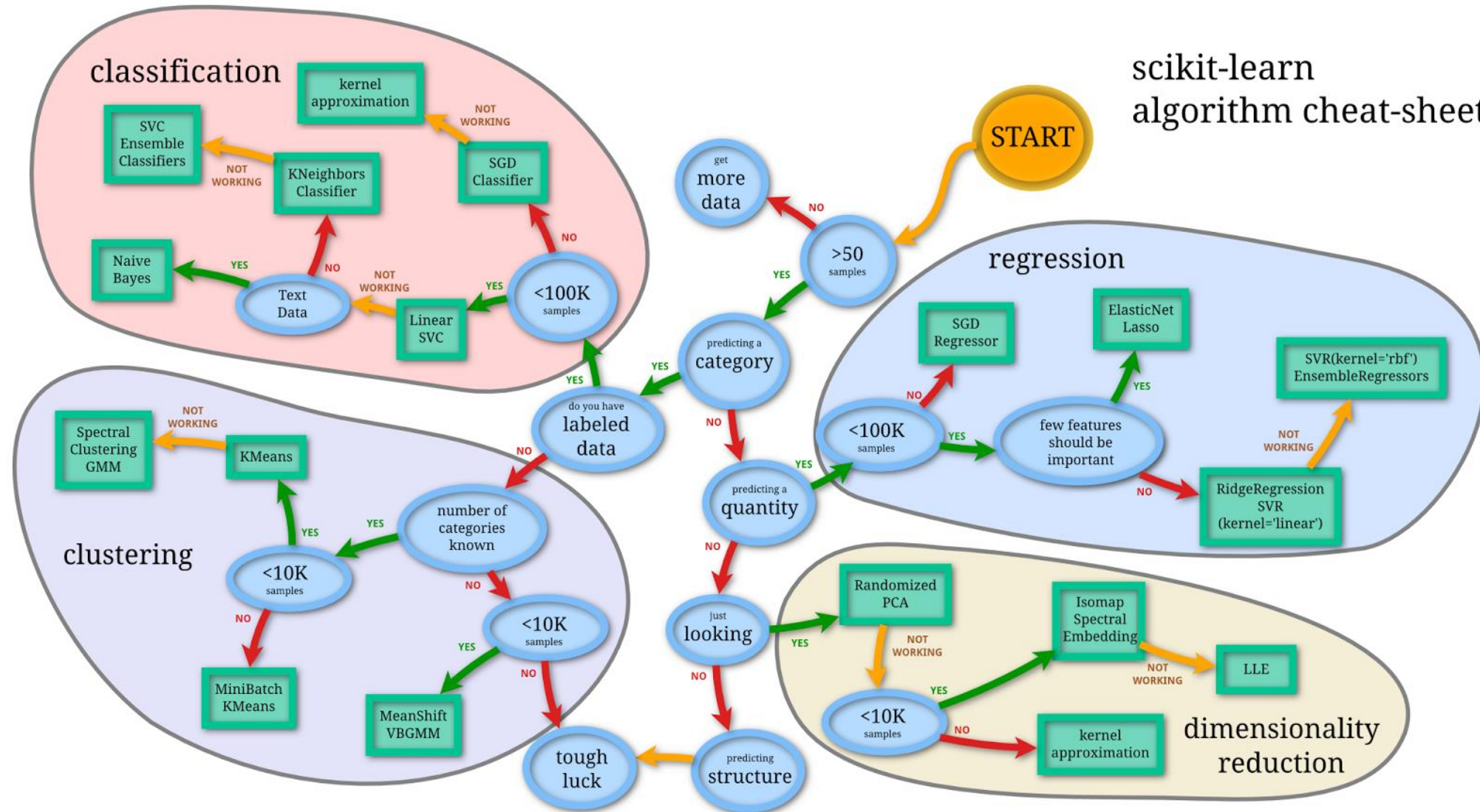
- R
- Python
- Mahout
- Spark MLlib

# Python机器学习

- Scikit-Learn是基于python的机器学习模块
- Scikit-Learn中的机器学习模型非常丰富，包括SVM，决策树，GBDT，KNN等等，可以根据问题的类型选择合适的模型
- 安装scikit\_learn

# Scikit-learn

## scikit-learn algorithm cheat-sheet

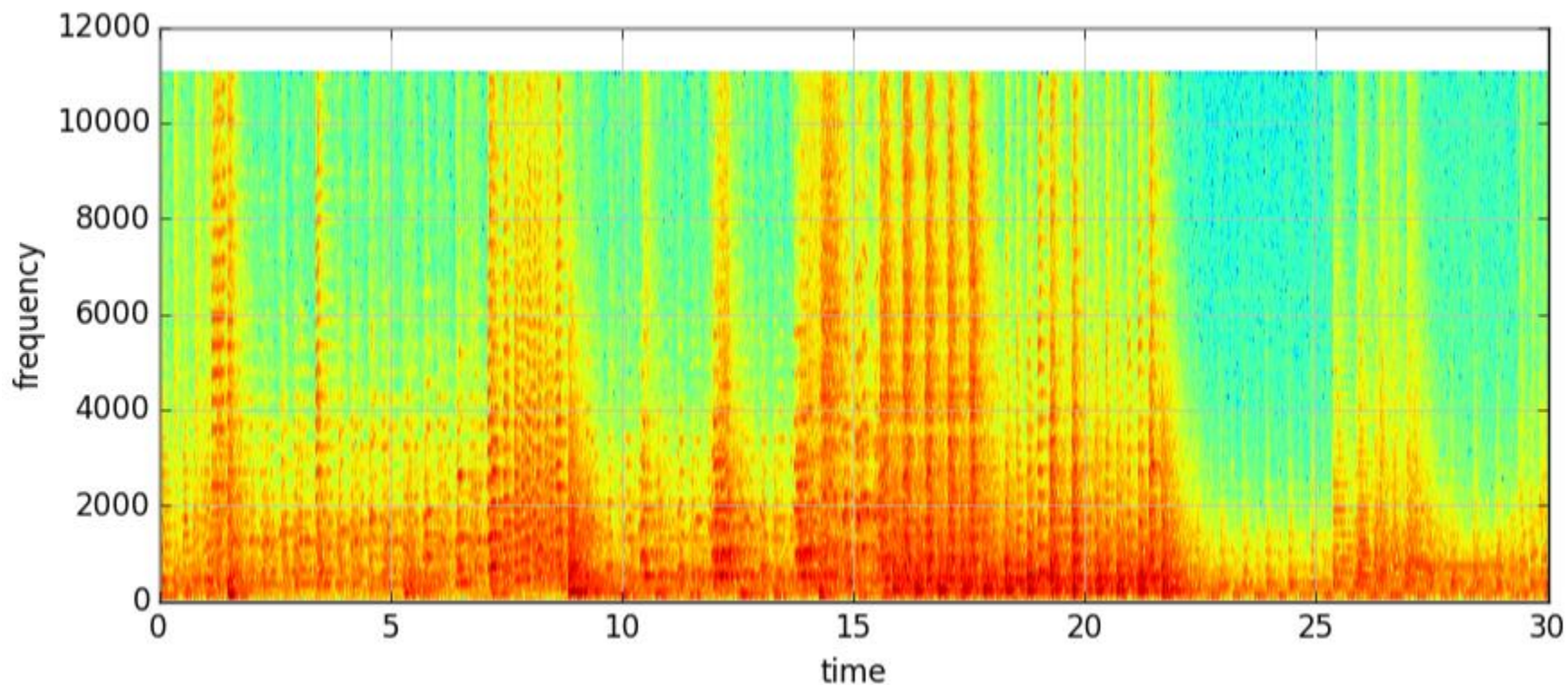


# 音乐分类

- 数据集(音乐数据)
- 算法使用(scikit-learn中的logistic regression)
- 期望结果(输入一首歌,可以对输入的歌曲进行分类)

# 音乐数据

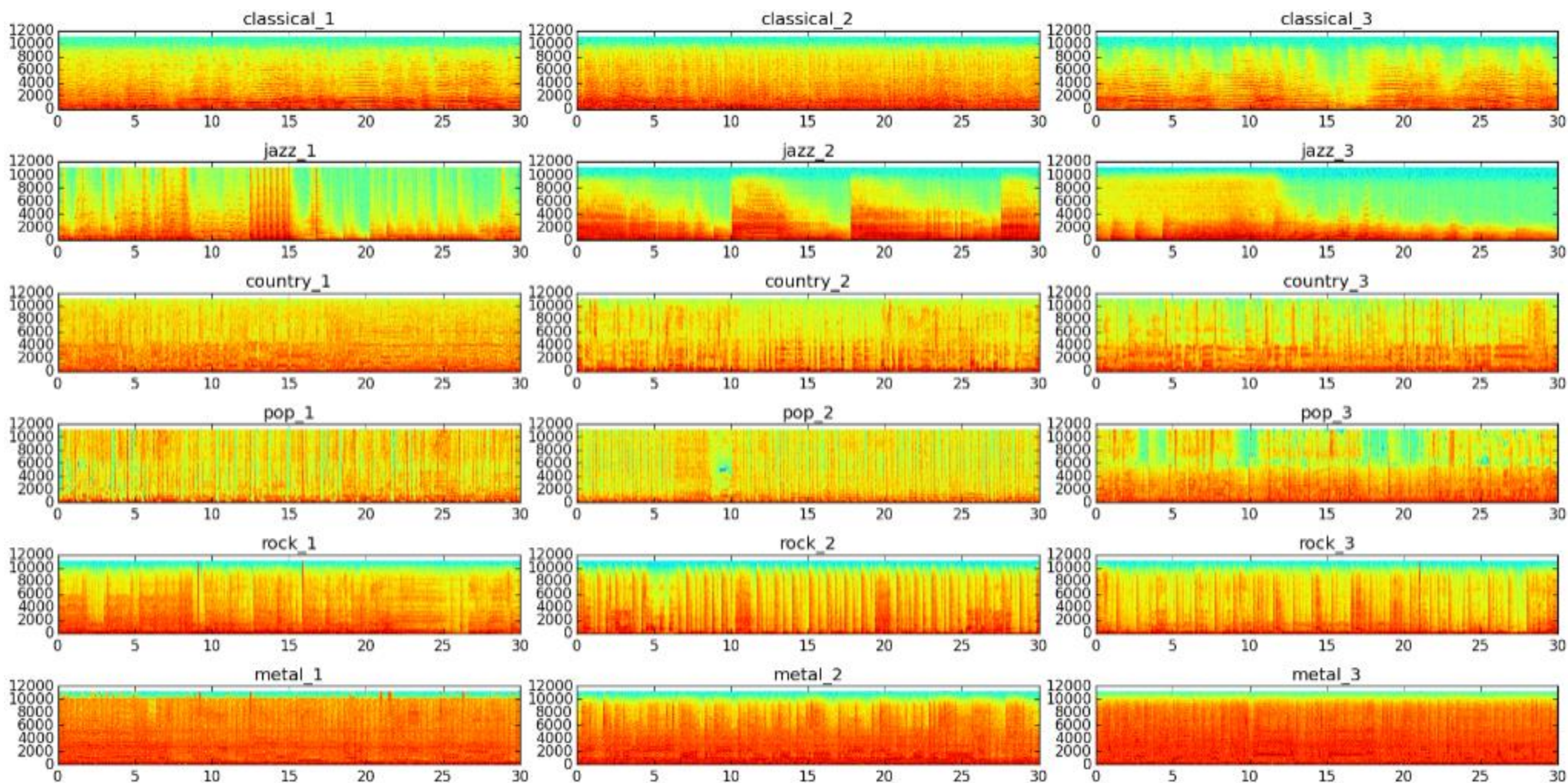
- 分类型存在文件夹中
- 以先把一个wma文件读入python,然后绘制它的频谱图(spectrogram)来看看是什么样的jazz





# 音乐数据

- 可以把每一种的音乐都抽一些出来打印频谱图以便比较,如下图:





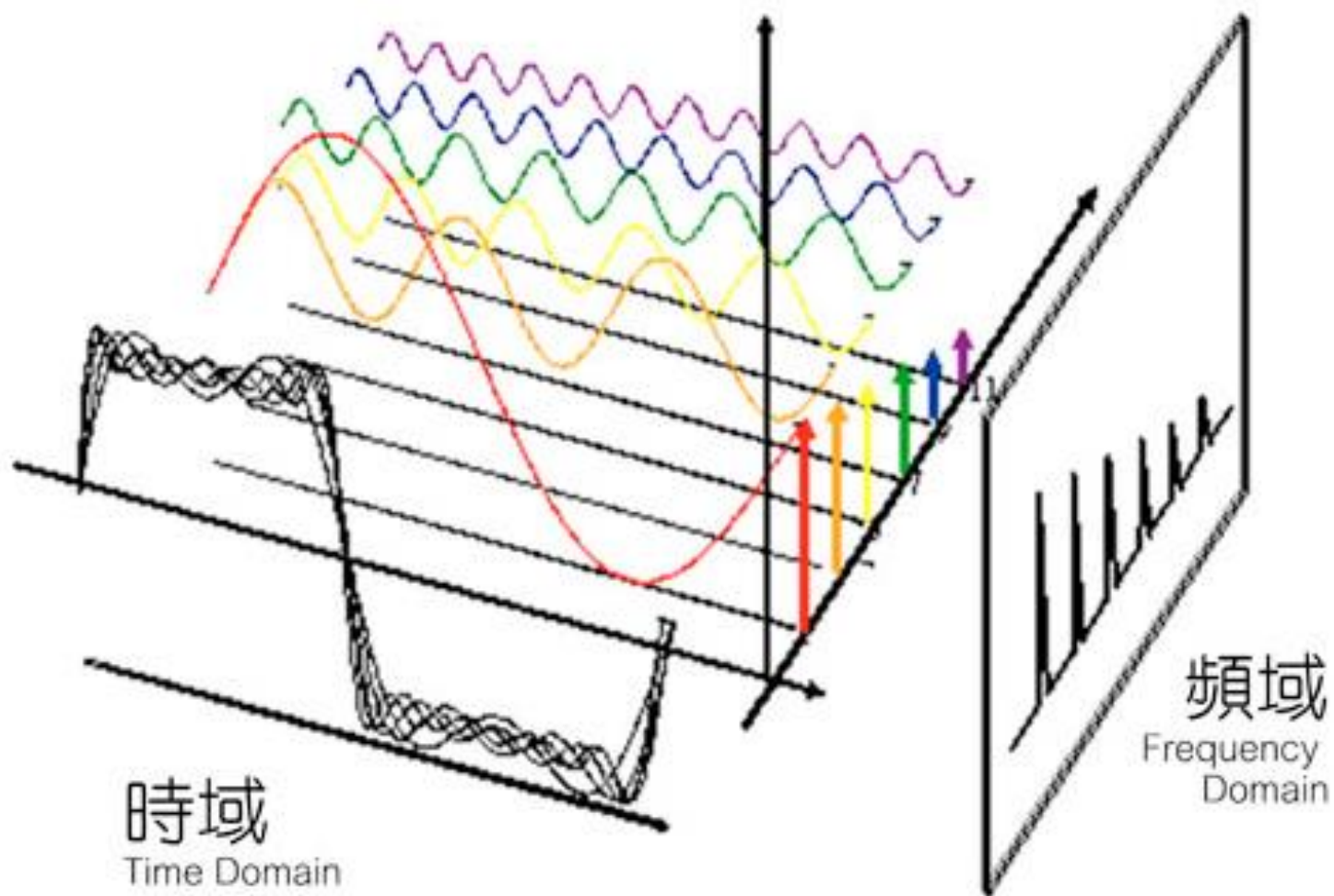
# 时域

- 什么是时域？？？
- 从我们出生，我们看到的世界都以时间贯穿。
- 股票的走势、人的身高、汽车的轨迹都会随着时间发生改变。
- 这种以时间作为参照来观察动态世界的方法我们称其为时域分析。
- 而我们也想当然的认为，世间万物都在随着时间不停的改变，并且永远不会静止下来

# 频域

- 什么是频域？？？
- 频域(**frequency domain**)是描述信号在频率方面特性时用到的一种坐标系。用线性代数的语言就是装着正弦函数的空间。
- 频域最重要的性质是：它不是真实的，而是一个数学构造。
- 正弦波是频域中唯一存在的波形，这是频域中最重要的规则，即正弦波是对频域的描述，因为时域中的任何波形都可用正弦波合成。

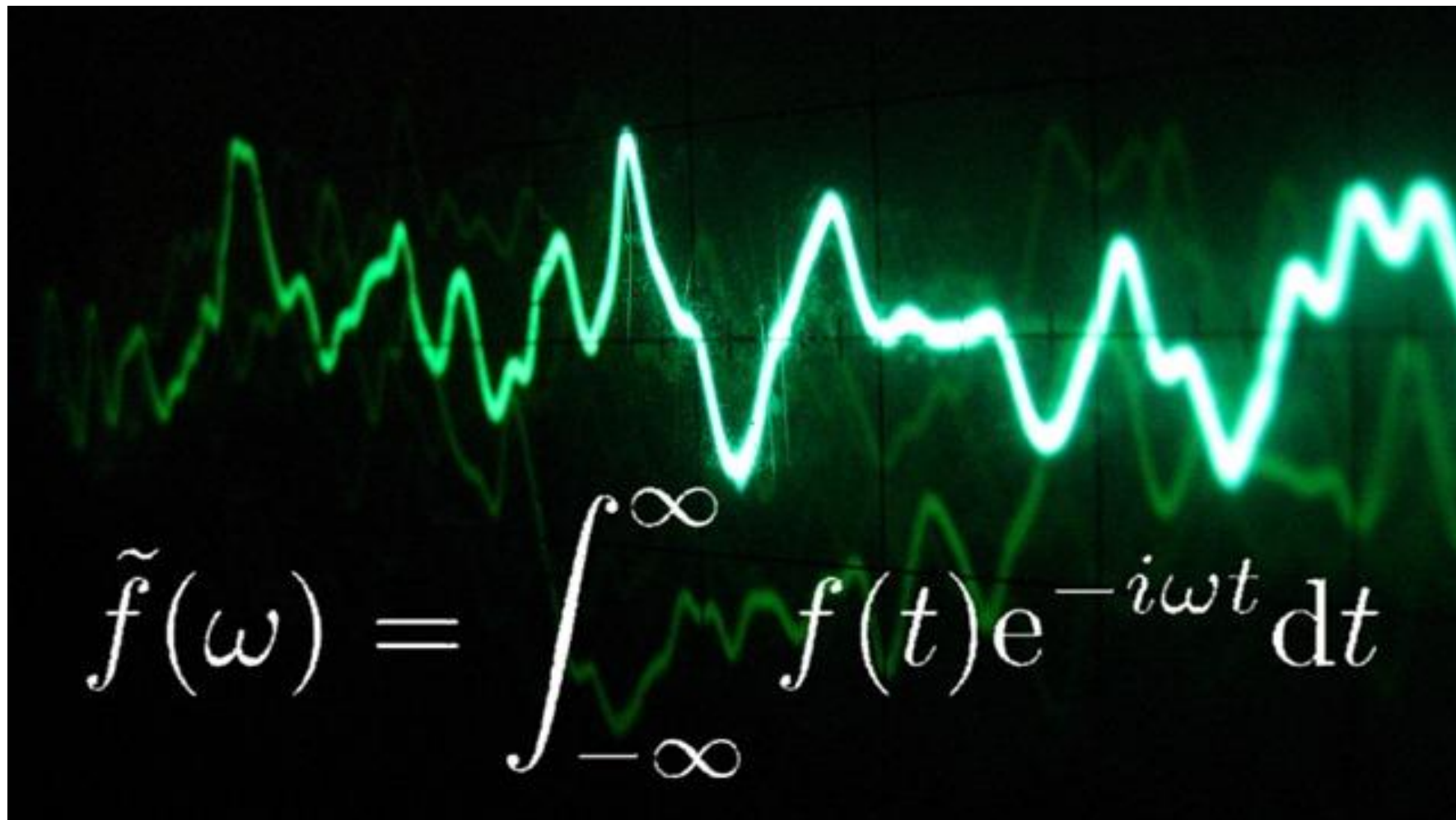
# Time Domain vs Frequency Domain



# 傅里叶变换

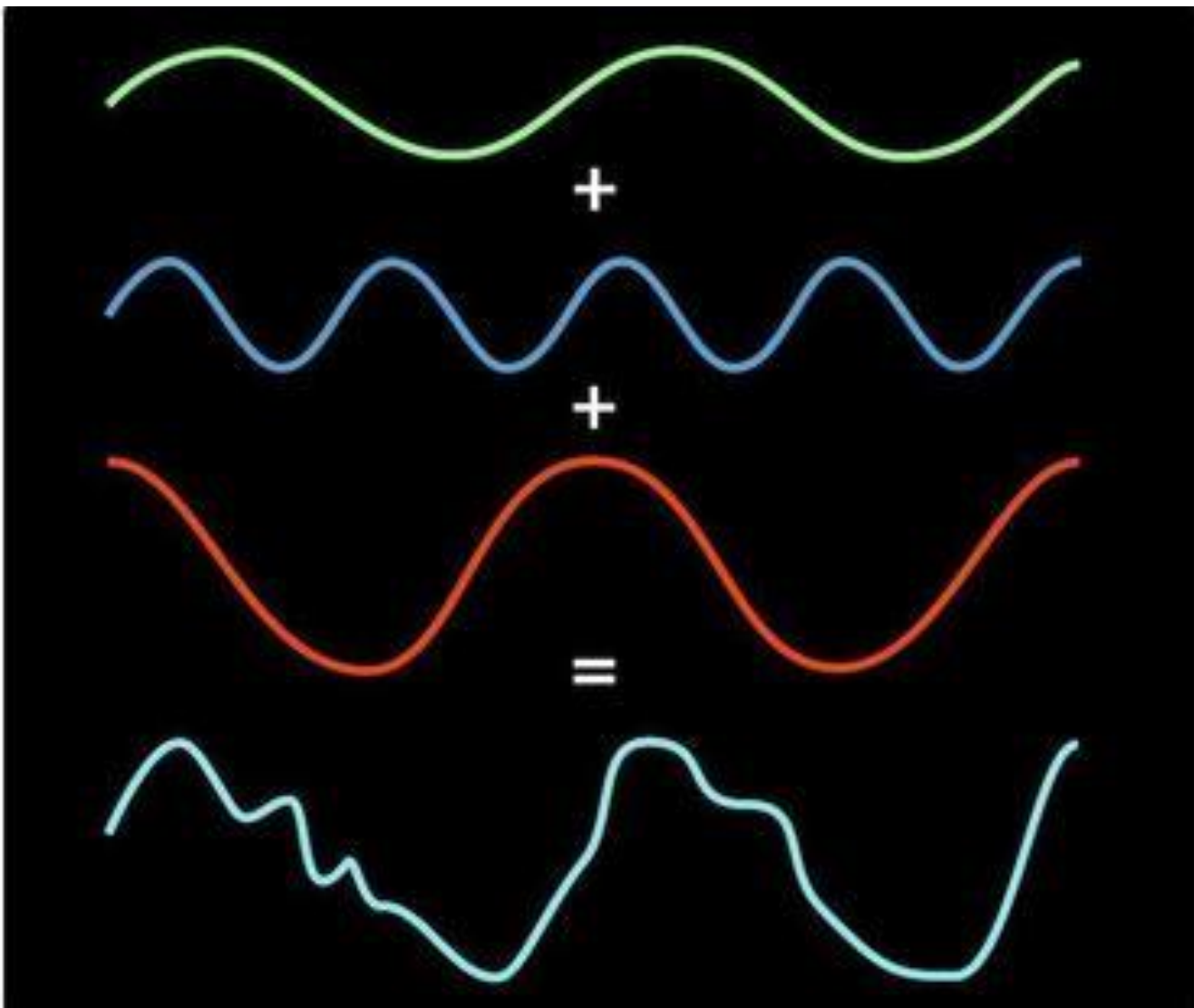
- 对于一个信号来说，信号强度随时间的变化规律就是时域特性，信号是由哪些单一频率的信号合成的就是频域特性。
- 时域分析与频域分析是对信号的两个观察面。
- 时域分析是以时间轴为坐标表示动态信号的关系；频域分析是把信号变为以频率轴为坐标表示出来。
- 一般来说，时域的表达较为形象与直观，频域分析则更为简练，剖析问题更为深刻和方便。
- 贯穿时域与频域的方法之一，就是传说中的叶变换(Fourier Transformation)。
- 傅里叶原理表明：任何连续测量的时序或信号，都可以表示为不同频率的正弦波信号的无限叠加。

# 傅里叶变换


$$\tilde{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$$

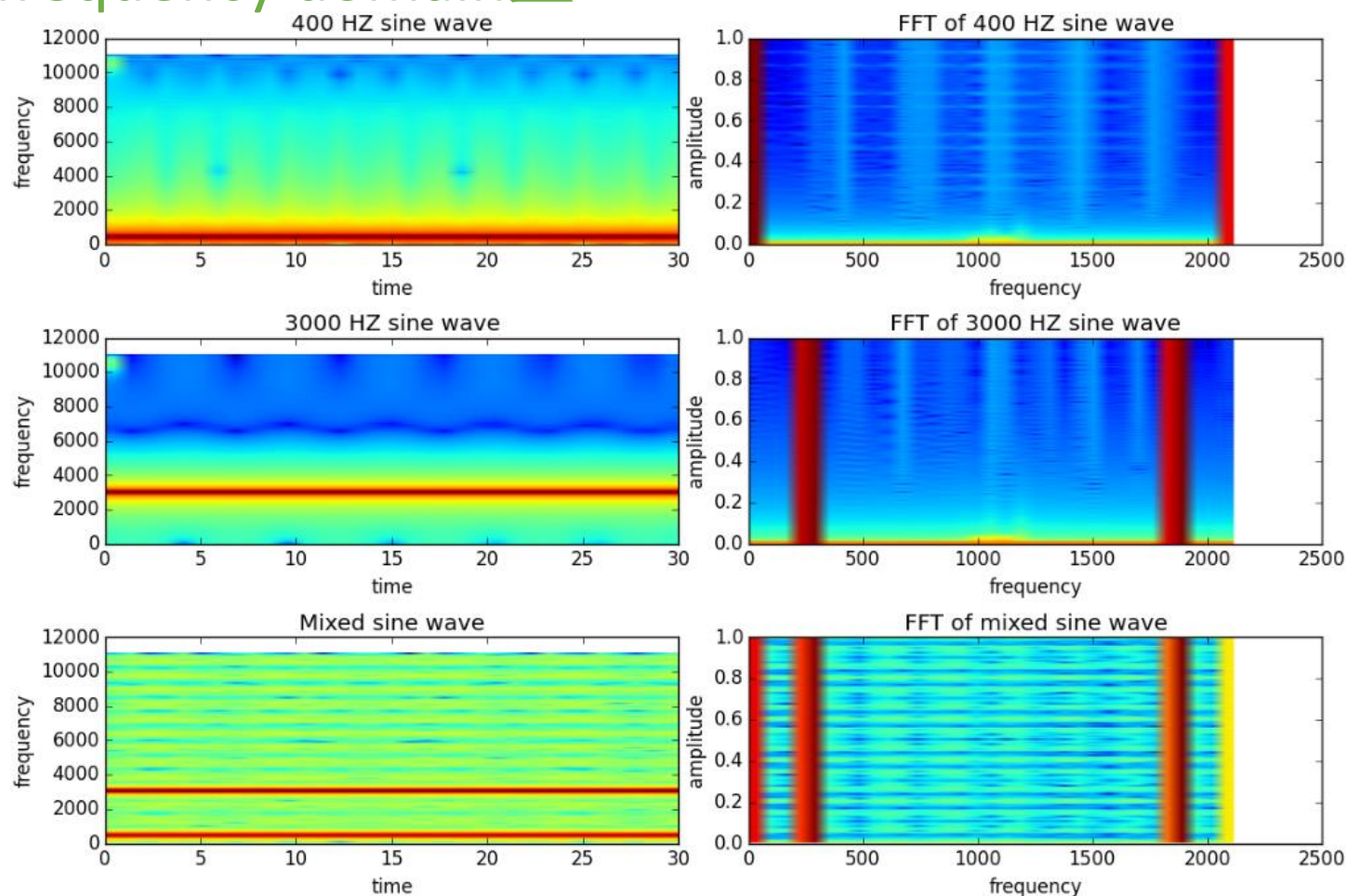


# 音乐由声波合成



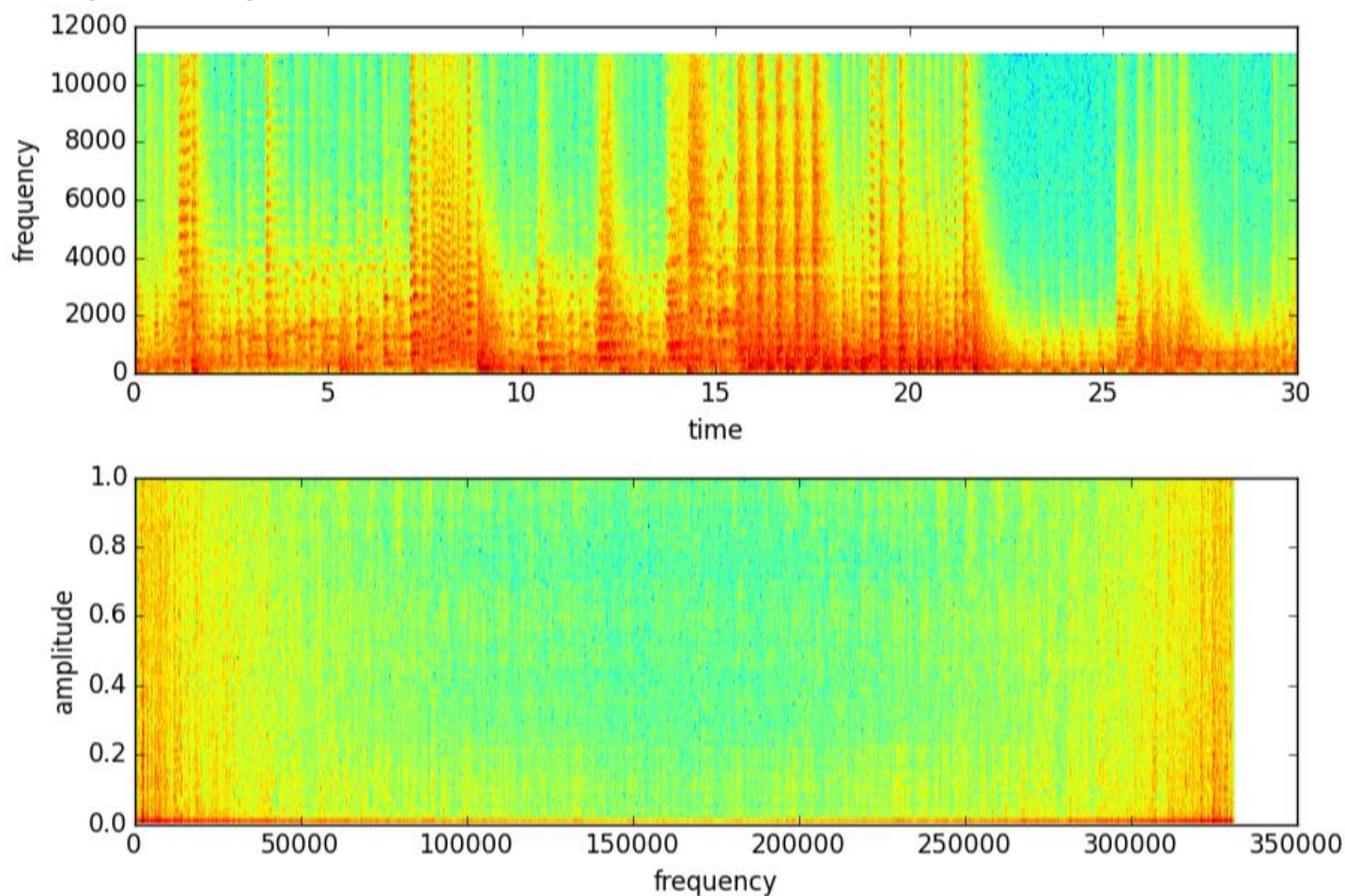
# 傅里叶变换

- 可以把time domain上的数据,例如一个音频,拆成一堆基准频率,然后投射到frequency domain上

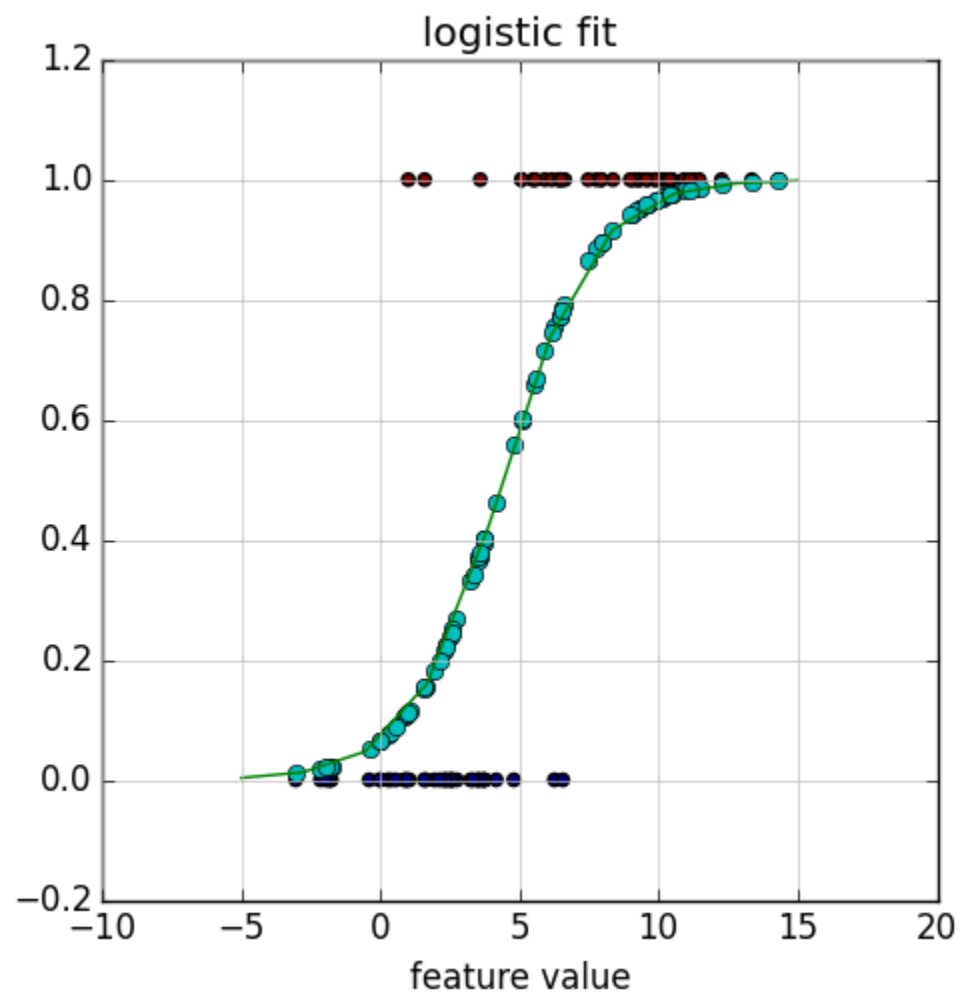
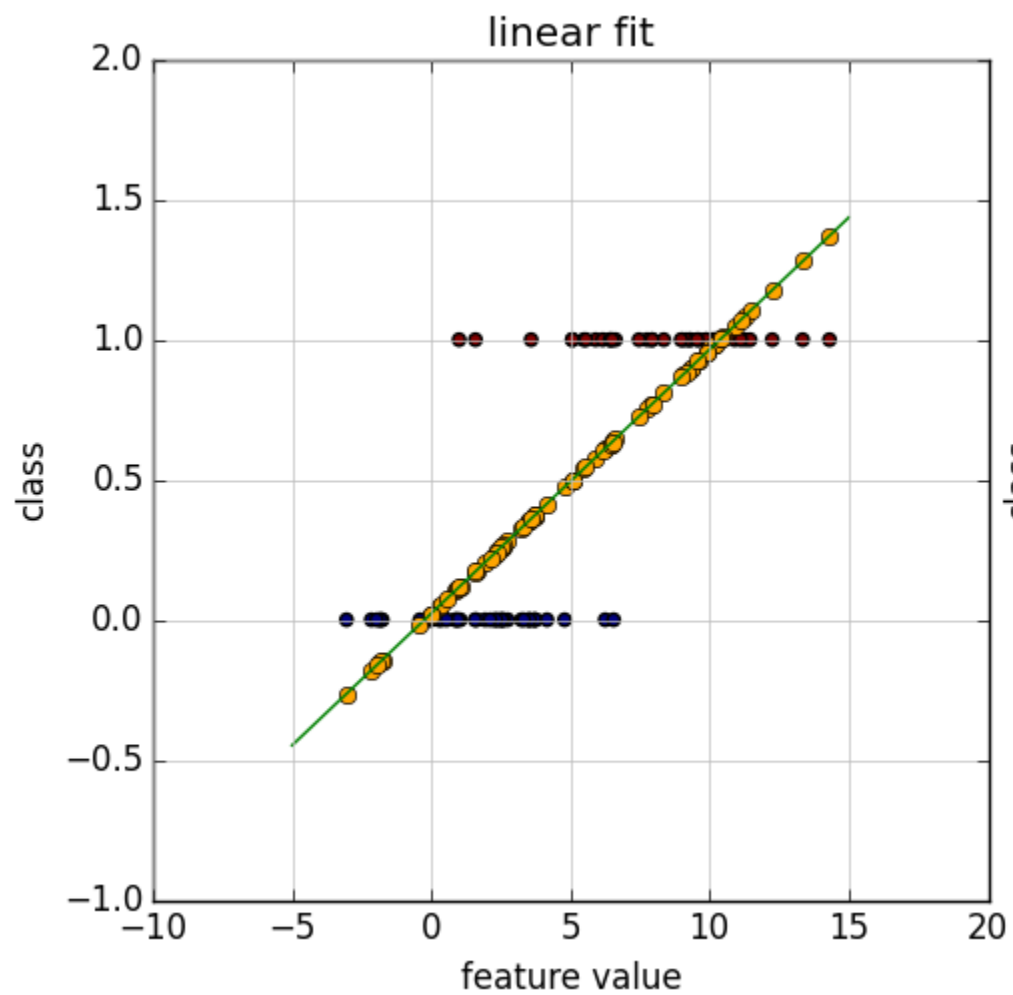


# 傅里叶变换

- 可以把time domain上的数据,例如一个音频,拆成一堆基准频率,然后投射到frequency domain上



# 逻辑回归



# 案例流程

- `["classical", "jazz", "country", "pop", "rock", "metal"]`
- 通过傅里叶变换将以上6类里面所有原始wav格式音乐文件转换为特征,并取前1000个特征,存入文件以便后续训练使用
- 读入以上6类特征向量数据作为训练集
- 使用sklearn包中LogisticRegression的fit方法计算出分类模型
- 读入黑豹乐队歌曲“无地自容”并进行傅里叶变换同样取前1000维作为特征向量
- 调用模型的predict方法对音乐进行分类,结果分为rock即摇滚类



# confusion matrix

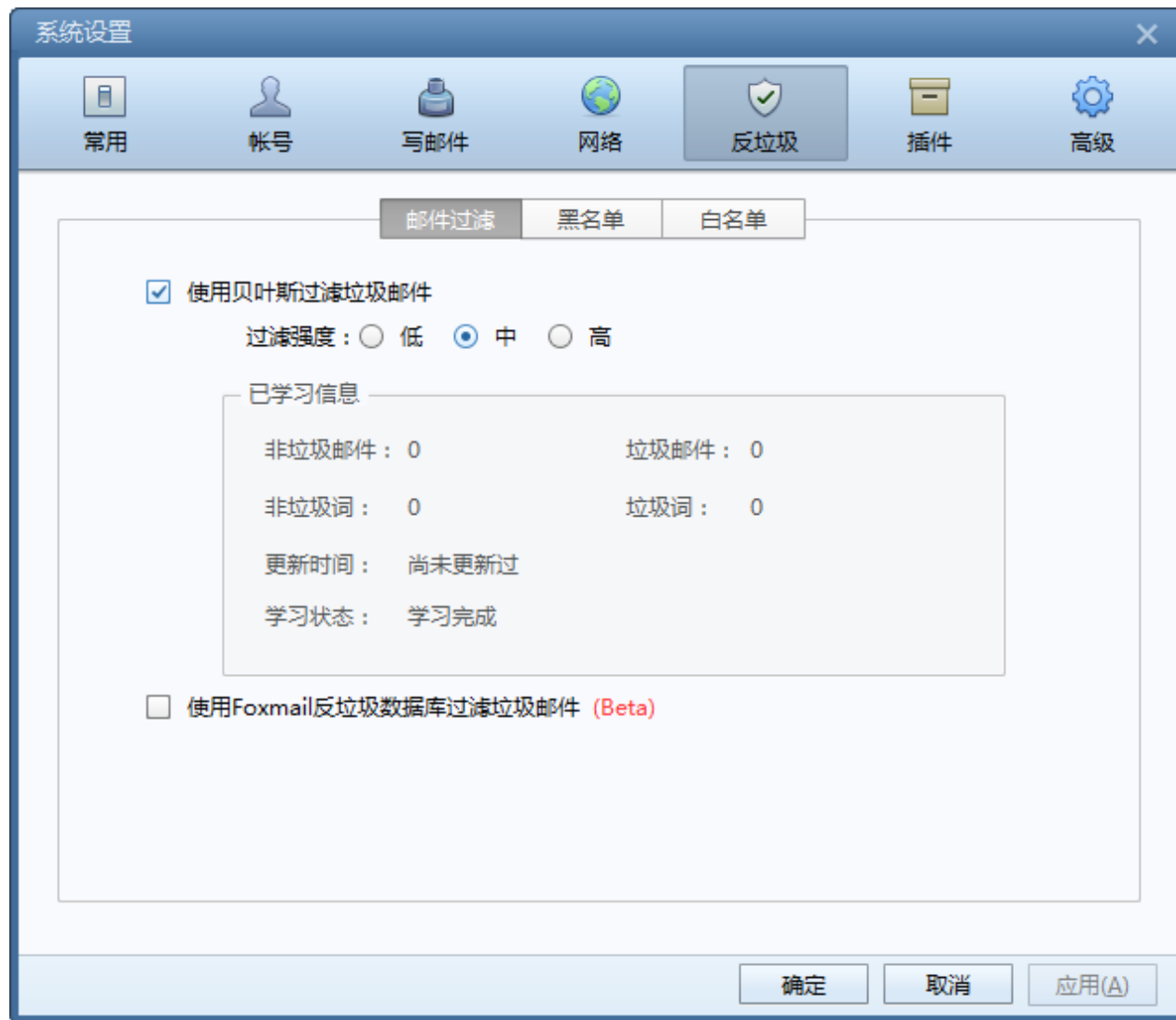
confusion matrix: FFT based logistic classifier

True class	classical	18	2	0	0	0	1
	jazz	2	10	4	1	1	3
	country	2	0	9	2	5	2
	pop	1	1	1	7	1	1
	rock	2	1	9	3	9	10
	metal	0	2	2	2	4	2
		classical	jazz	country	pop	rock	metal
		Predicted class					

confusion matrix: FFT based KNN classifier

True class	classical	22	0	0	0	0	0
	jazz	3	12	2	0	4	1
	country	0	3	12	5	3	3
	pop	0	0	1	7	0	2
	rock	0	1	8	1	11	2
	metal	0	0	2	2	2	11
		classical	jazz	country	pop	rock	metal
		Predicted class					

# 反垃圾邮件

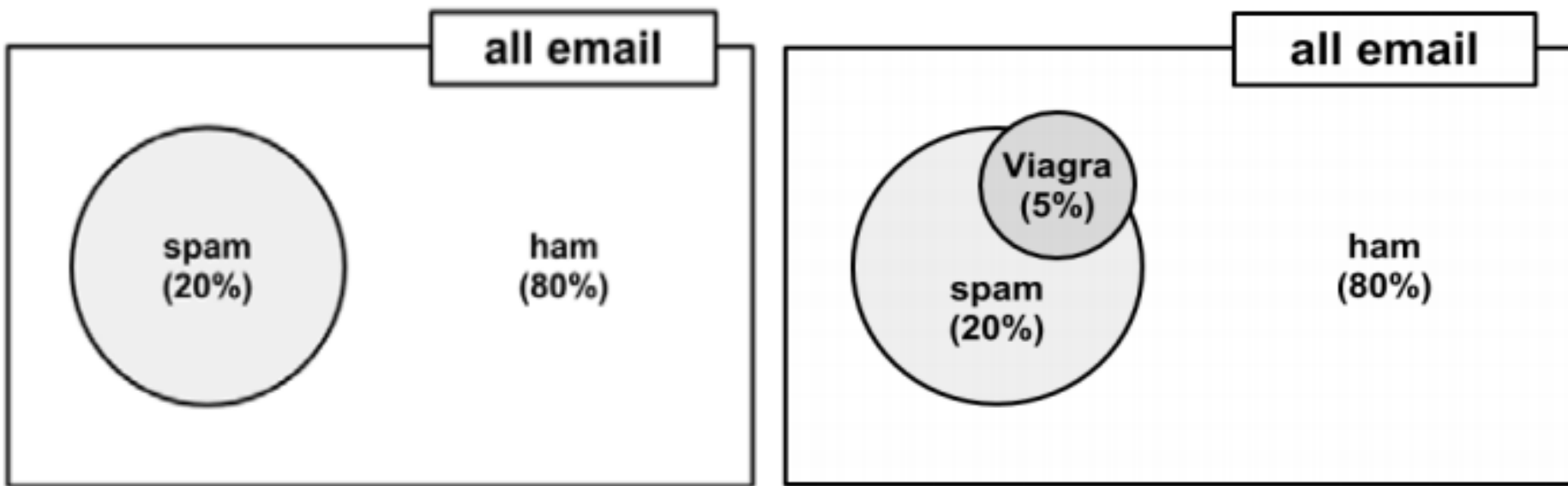


# 贝叶斯分类器

- 70%降水概率
- 机器学习算法中，有种依据概率原则进行分类的朴素贝叶斯算法，正如气象学家预测天气一样，朴素贝叶斯算法就是应用先前事件的有关数据来估计未来事件发生的概率

# 理解朴素贝叶斯

- 如果我们知道P（垃圾邮件）和P（Viagra）是相互独立的，则容易计算P（垃圾邮件&Viagra），即这两个事件同时发生的概率。  
 $20\% * 5\% = 1\%$



# 贝叶斯公式

- 独立事件我们可以简单的应用这个方法计算，但是在显示中，P（垃圾邮件）和P（Viagra）更可能是高度相关的，因此上述计算是不正确的，我们需要一个精确的公式来描述这两个事件之间的关系。
- 基于贝叶斯定理的条件概率

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$



# 基于贝叶斯定理的条件概率

- 对于我们垃圾邮件来说

$$P(\text{spam} | \text{Viagra}) = \frac{P(\text{Viagra} | \text{spam}) P(\text{spam})}{P(\text{Viagra})}$$

Diagram illustrating the components of Bayes' Theorem for spam classification:

- posterior probability**:  $P(\text{spam} | \text{Viagra})$
- likelihood**:  $P(\text{Viagra} | \text{spam})$
- prior probability**:  $P(\text{spam})$
- marginal likelihood**:  $P(\text{Viagra})$

# 理解贝叶斯分类

- 计算贝叶斯定理中每一个组成部分的概率，我们必须构造一个频率表

	Viagra		
Frequency	Yes	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100

	Viagra		
Likelihood	Yes	No	Total
spam	$4 / 20$	$16 / 20$	20
ham	$1 / 80$	$79 / 80$	80
Total	$5 / 100$	$95 / 100$	100

# 理解贝叶斯分类

- 计算贝叶斯公式
- $P(\text{垃圾邮件} | \text{Viagra}) = P(\text{Viagra} | \text{垃圾邮件}) * P(\text{垃圾邮件}) / P(\text{Viagra}) = (4/20) * (20/100) / (5/100) = 0.8$
- 因此，如果电子邮件含有单词Viagra，那么该电子邮件是垃圾邮件的概率为80%。所以，任何含有单词Viagra的消息都需要被过滤掉。

# 理解贝叶斯分类

- 当有额外更多的特征是，这一概念如何被使用

	Viagra ( $W_1$ )		Money ( $W_2$ )		Groceries ( $W_3$ )		Unsubscribe ( $W_4$ )		
Likelihood	Yes	No	Yes	No	Yes	No	Yes	No	Total
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

- 利用贝叶斯公式，我们得到概率如下：

$$P(\text{Spam} | W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 | \text{spam}) P(\text{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

# 理解贝叶斯分类

$$P(\text{Spam} | W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 | \text{spam}) P(\neg W_2 | \text{spam}) P(\neg W_3 | \text{spam}) P(W_4 | \text{spam}) P(\text{spam})}{P(W_1) P(\neg W_2) P(\neg W_3) P(W_4)}$$

$$P(\text{ham} | W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 | \text{ham}) P(\neg W_2 | \text{ham}) P(\neg W_3 | \text{ham}) P(W_4 | \text{ham}) P(\text{ham})}{P(W_1) P(\neg W_2) P(\neg W_3) P(W_4)}$$

- 分母可以先忽略它，垃圾邮件的总似然为：
- $(4/20) * (10/20) * (20/20) * (12/20) * (20/100) = 0.012$
- 非垃圾邮件的总似然为：
- $(1/80) * (66/80) * (71/80) * (23/80) * (80/100) = 0.002$
- 将这些值转换成概率，我们只需要一步得到垃圾邮件概率为85.7%



# 问题

- 另一个例子包含了4个单词的邮件呢？
- 我们可以计算垃圾邮件的似然如下：
- $(4/20)*(10/20)*(0/20)*(12/20)*(20/100)=0$
- 非垃圾邮件的似然为：
- $(1/80)*(14/80)*(8/80)*(23/80)*(80/100)=0.00005$
- 因此该消息是垃圾邮件的概率为  $0/(0+0.00005)=0$
- 该消息是非垃圾邮件的概率为  $0.00005/(0+0.00005)=1$
- 问题出在Groceries这个单词，所有单词Groceries有效抵消或否决了所有其他的证据

# 拉普拉斯估计

- 拉普拉斯估计本质上是给频率表中的每个计数加上一个较小的数，这样就保证了每一类中每个特征发生概率非零。
- 通常情况下，拉普拉斯估计中加上的数值设定为1，这样就保证每一类特征的组合至少在数据中出现一次。
- 然后，我们得到垃圾邮件的似然为：
  - $(5/24)*(11/24)*(1/24)*(13/24)*(20/100)=0.0004$
  - 非垃圾邮件的似然为：
    - $(2/84)*(15/84)*(9/84)*(24/84)*(80/100)=0.0001$
  - 这表明该消息是垃圾邮件的概率为80%，是非垃圾邮件的概率为20%。

# 联系我们

- 北京尚学堂官网: <http://www.bjsxt.com/html/cloud/>
- QQ讨论群: 172599077
- 咨询老师:
- 何老师: 1926106490
- 贾老师: 1786418286
- 詹老师: 2805048645
- 张老师: 3254755158