分类的应用

- 故障识别

  北京高碑店华能电厂项目，6台机组，根据实时监控数据预测是否会出现故障。

- 贷款风险预防

  ZestFinance通过各种数据源获得用户行为数据，预测用户信用与还款行为。已与百度签订合作协定。

- 医疗图像识别

  IBM通过图像识别技术，判断痣与皮肤癌。

分类的应用

- 故障识别

北京高碑店华能电厂项目，6台机组，根据实时监控数据预测是否会出现故障。

- 贷款风险预防

ZestFinance通过各种数据源获得用户行为数据，预测用户信用与还款行为。已与百度签订合作协定。

- 医疗图像识别

IBM通过图像识别技术，判断痣与皮肤癌。

分类的应用

- 故障识别

北京高碑店华能电厂项目，6台机组，根据实时监控数据预测是否会出现故障。

- 贷款风险预防

ZestFinance通过各种数据源获得用户行为数据，预测用户信用与还款行为。已与百度签订合作协定。

- 医疗图像识别

IBM通过图像识别技术，判断痣与皮肤癌。

什么是好的分类

- 以交警抓酒后驾车司机为例

- 目标：抓出酒驾，不冤枉一个好人，不放过一个隐患

A. 实际：没喝酒，识别出来：没喝酒
B. 实际：没喝酒，识别出来：喝酒了
C. 实际：喝酒了，识别出来：没喝酒
D. 实际：喝酒了，识别出来：喝酒了

什么是好的分类

	识别结果：喝酒了	识别结果：没喝酒
实际：喝酒了		
实际：没喝酒		

目标：抓酒驾

测试仪若显示浓度超标：Positive，**阳性**

测试仪若显示浓度正常：Negative, **阴性**

什么是好的分类

Confusion Matrix混淆矩阵

	识别结果：喝酒了 (Positive)	识别结果：没喝酒 (Negative)
实际：喝酒了 (Positive)	True Positive	False Negative
实际：没喝酒 (Negative)	False Positive	True Negative

目标：抓酒驾

测试仪若显示浓度超标： Positive，**阳性**

测试仪若显示浓度正常： Negative, **阴性**

什么是好的分类

- 查酒驾结果，以测试仪显示结果是否>=20mg/100ml为标准。
- 一共查了200人，其中，170人显示超过20mg/100ml，其中163人证实喝酒，7人确实没喝酒。剩余30人显示低于20mg/100ml，但交警时候发现，其中有3人也喝过酒，只是采取了一些特殊方式蒙骗了测试仪，其余27人没喝过酒。
- 怎样填写上页的矩阵?

什么是好的分类

	识别结果：喝酒了 (Positive)	识别结果：没喝酒 (Negative)
实际：喝酒了 (Positive)	True Positive 163	False Negative 3
实际：没喝酒 (Negative)	False Positive 7	True Negative 27

目标：抓酒驾

测试仪若显示浓度超标： Positive，**阳性**

测试仪若显示浓度正常： Negative, **阴性**

什么是好的分类

召回率，覆盖率，Recall rate：实际有多少比例被抓到

	识别结果：喝酒了 (Positive)	识别结果：没喝酒 (Negative)
实际：喝酒了 (Positive)	True Positive 163	False Negative 3
实际：没喝酒 (Negative)	False Positive 7	True Negative 27

目标：抓酒驾

测试仪若显示浓度超标：Positive，**阳性**

测试仪若显示浓度正常：Negative, **阴性**

什么是好的分类

召回率，覆盖率，Recall rate：实际有多少比例被抓到

准确度，Precision：预测喝酒的有多少是真的喝了酒

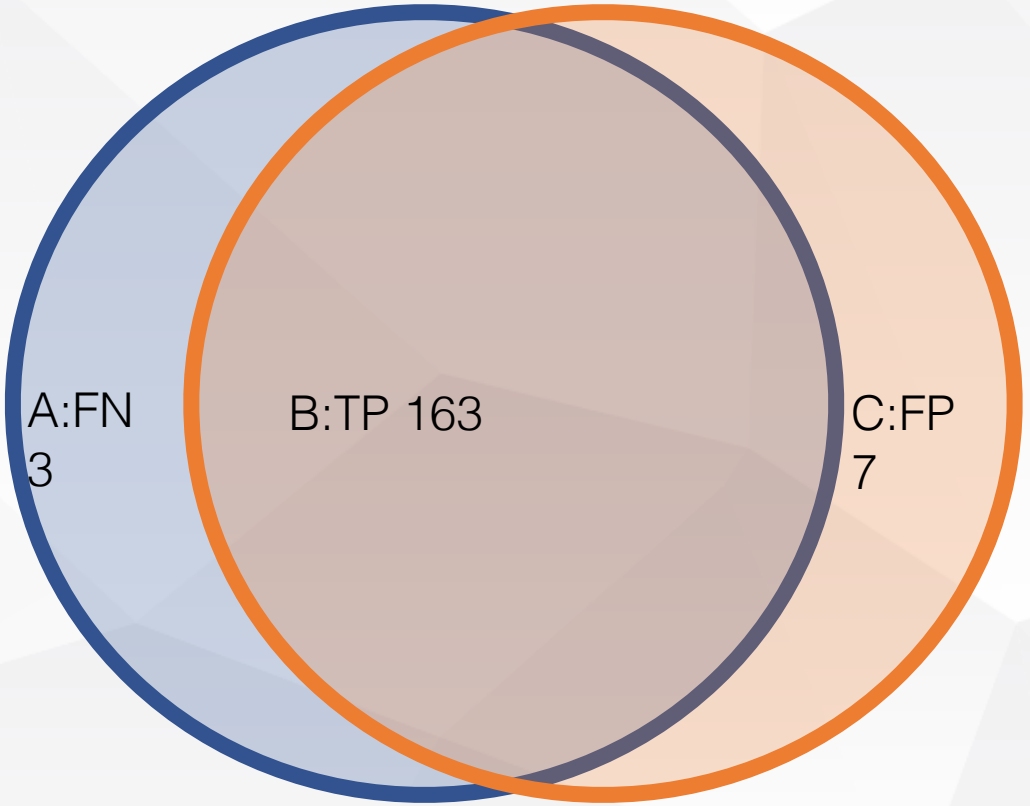
	识别结果：喝酒了 (Positive)	识别结果：没喝酒 (Negative)
实际：喝酒了 (Positive)	True Positive 163	False Negative 3
实际：没喝酒 (Negative)	False Positive 7	True Negative 27

目标：抓酒驾

测试仪若显示浓度超标： Positive，**阳性**

测试仪若显示浓度正常： Negative, **阴性**

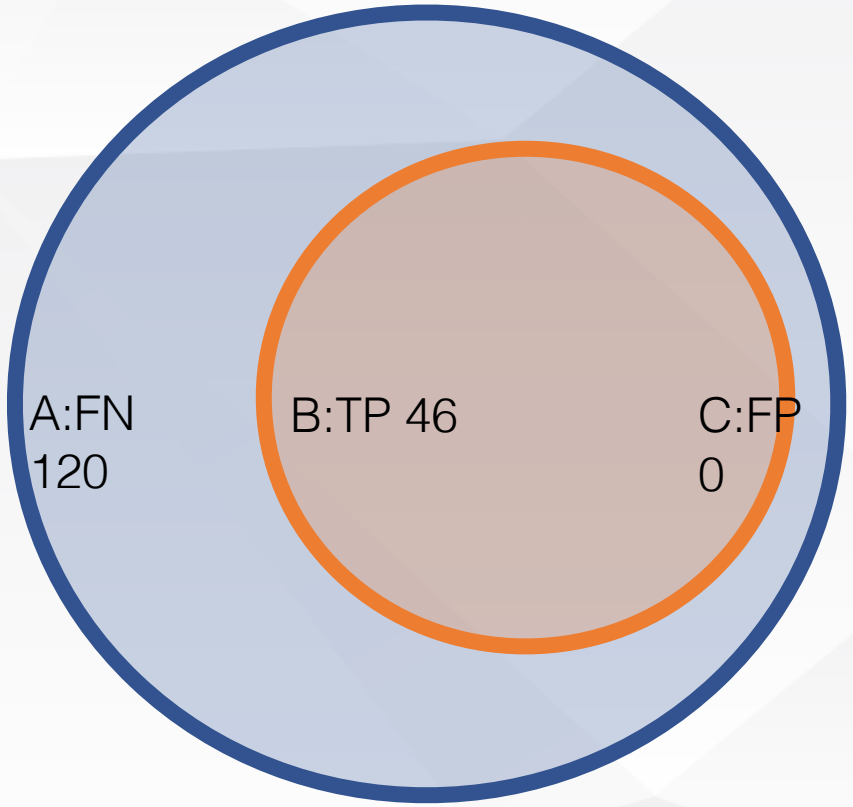
什么是好的分类



召回率: $\mathrm{R} = \frac{B}{A+B}$

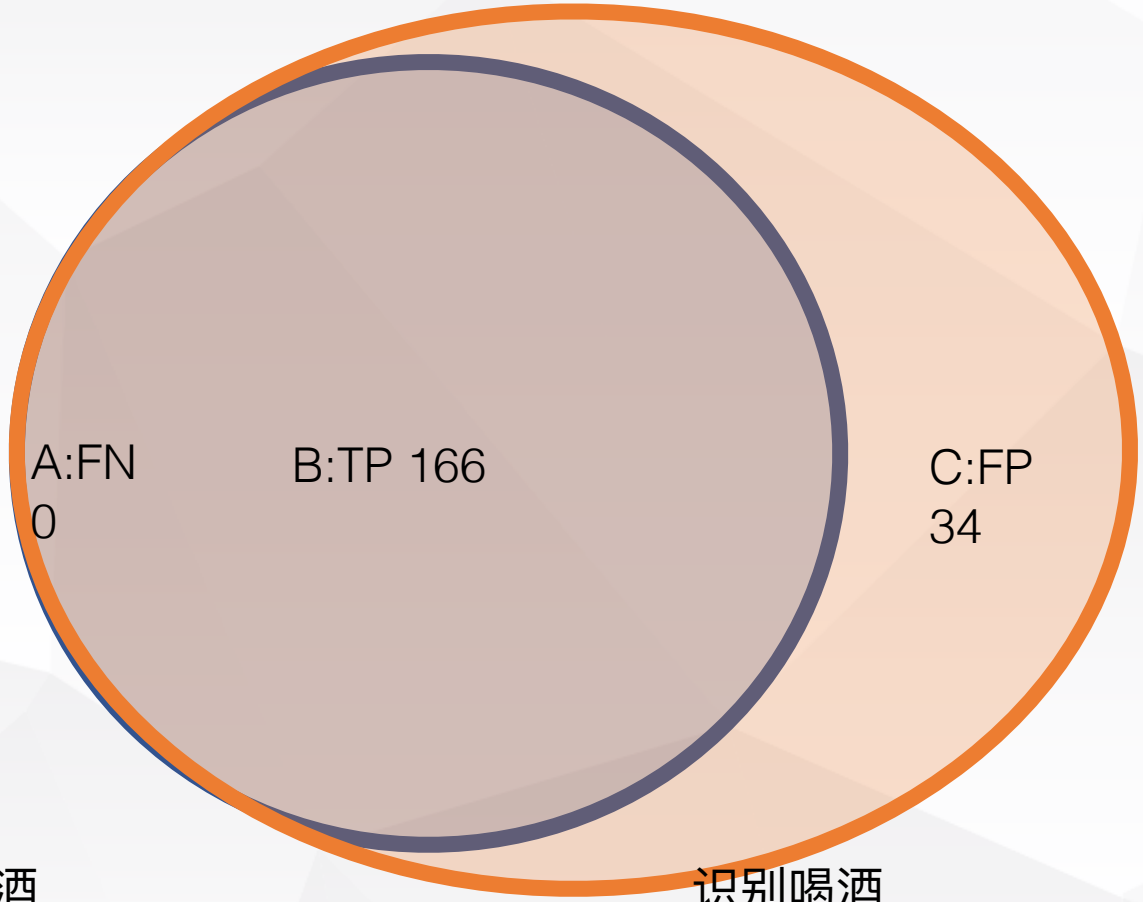
准确率: $\mathrm{P} = \frac{B}{C+B}$

什么是好的分类



检测仪1结果

召回率: $R = \frac{B}{A+B}$ =46/166　　准确率: $P = \frac{B}{C+B}$=46/46



检测仪2结果

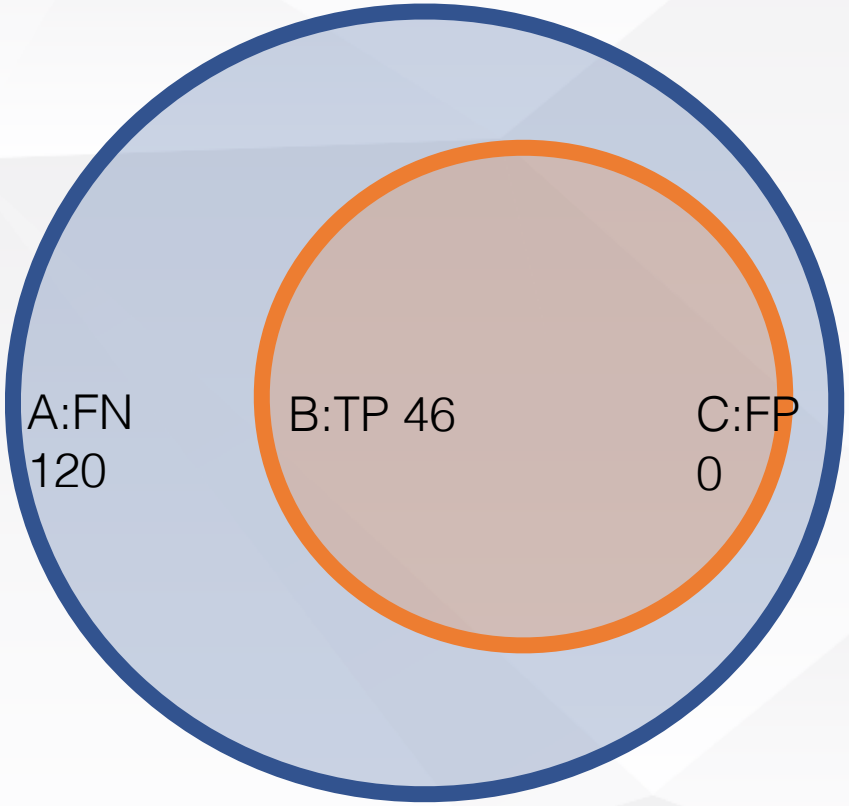
召回率: $R = \frac{B}{A+B}$ = 166/166　准确率: $P = \frac{B}{C+B}$=166/200

什么是好的分类

哪台检测仪好?

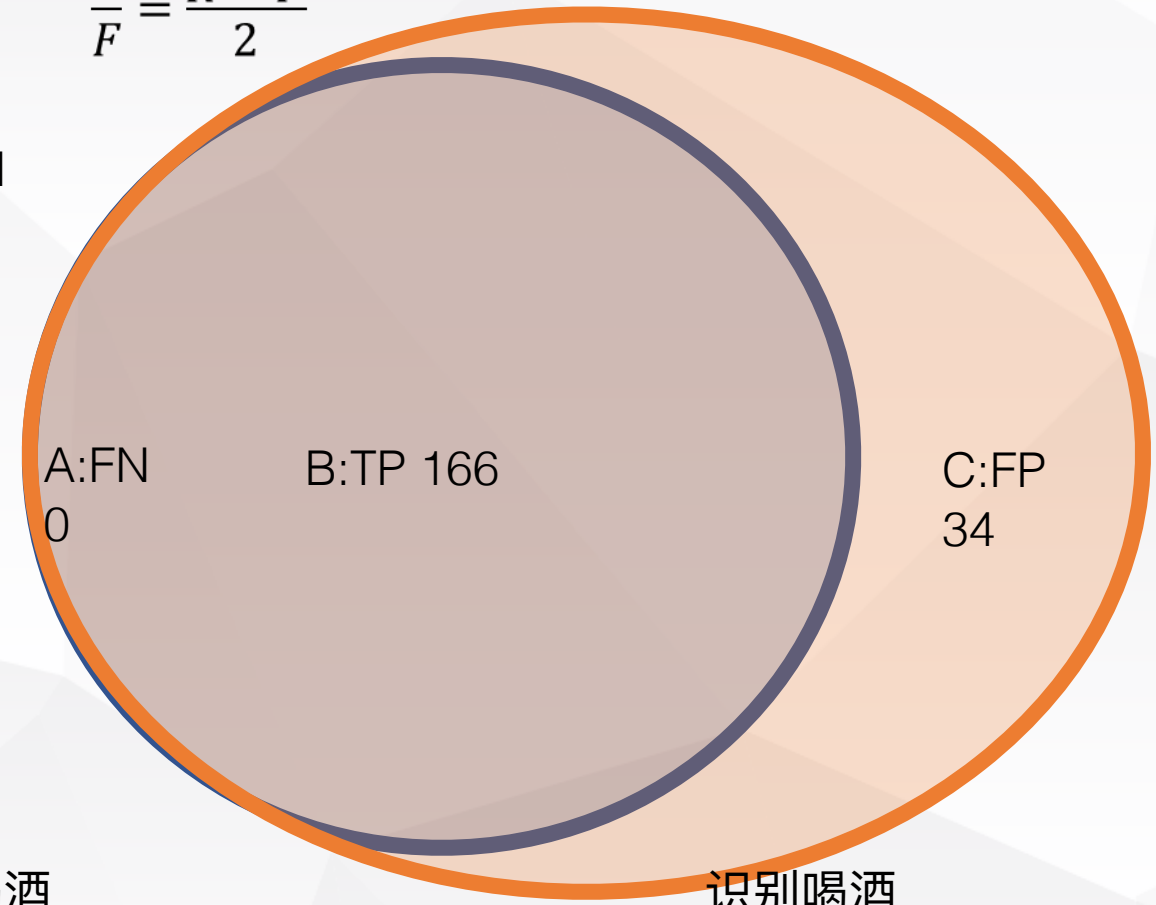
$$F = \frac{2\ R^{*}P}{R + P}$$

$$\frac{1}{F} = \frac{\frac{1}{R} + \frac{1}{P}}{2}$$



检测仪1结果

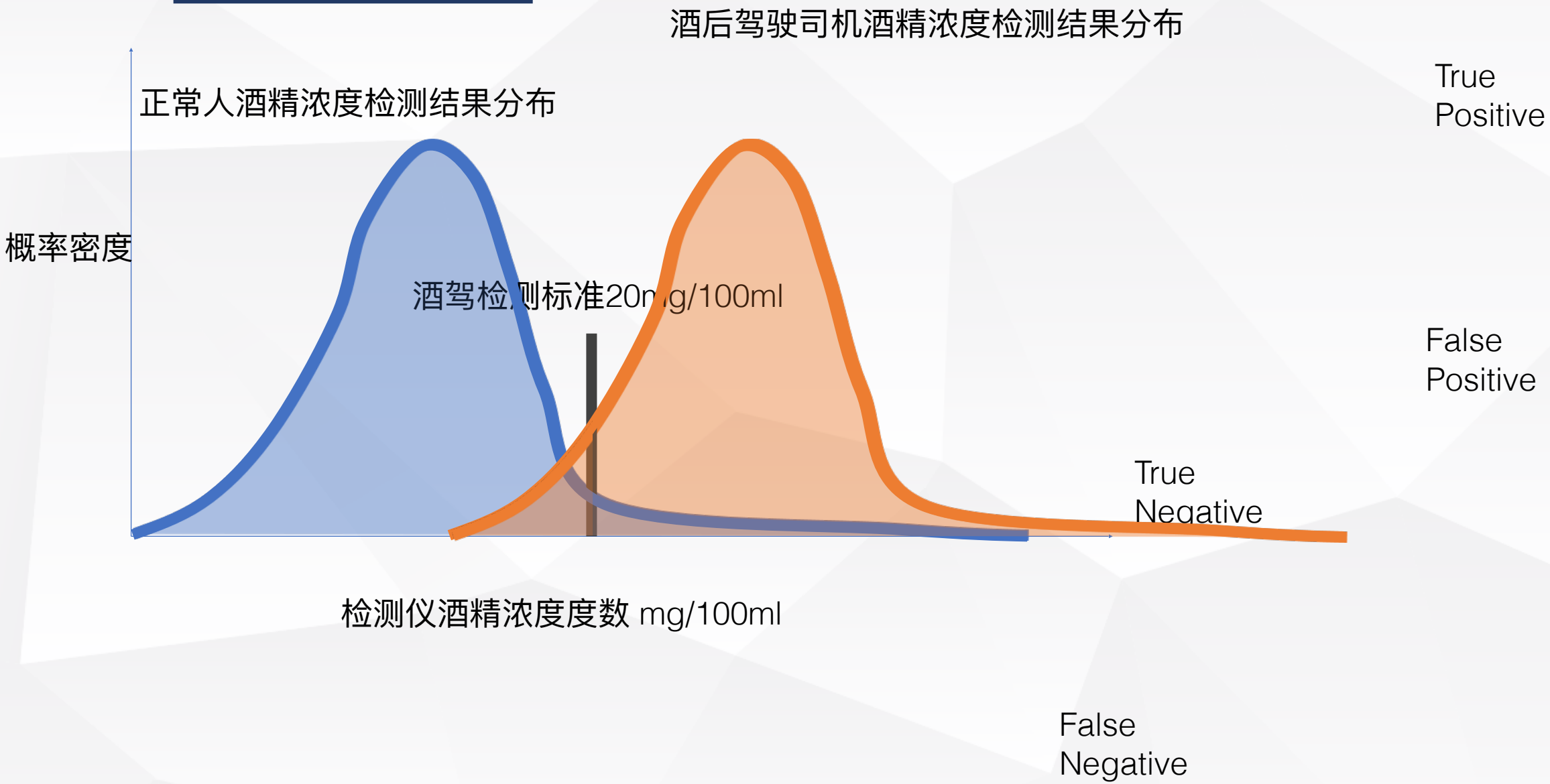
召回率: $\mathrm{R} = \frac{B}{A+B}$ =46/166　　准确率: $\mathrm{P} = \frac{B}{C+B}$=46/46



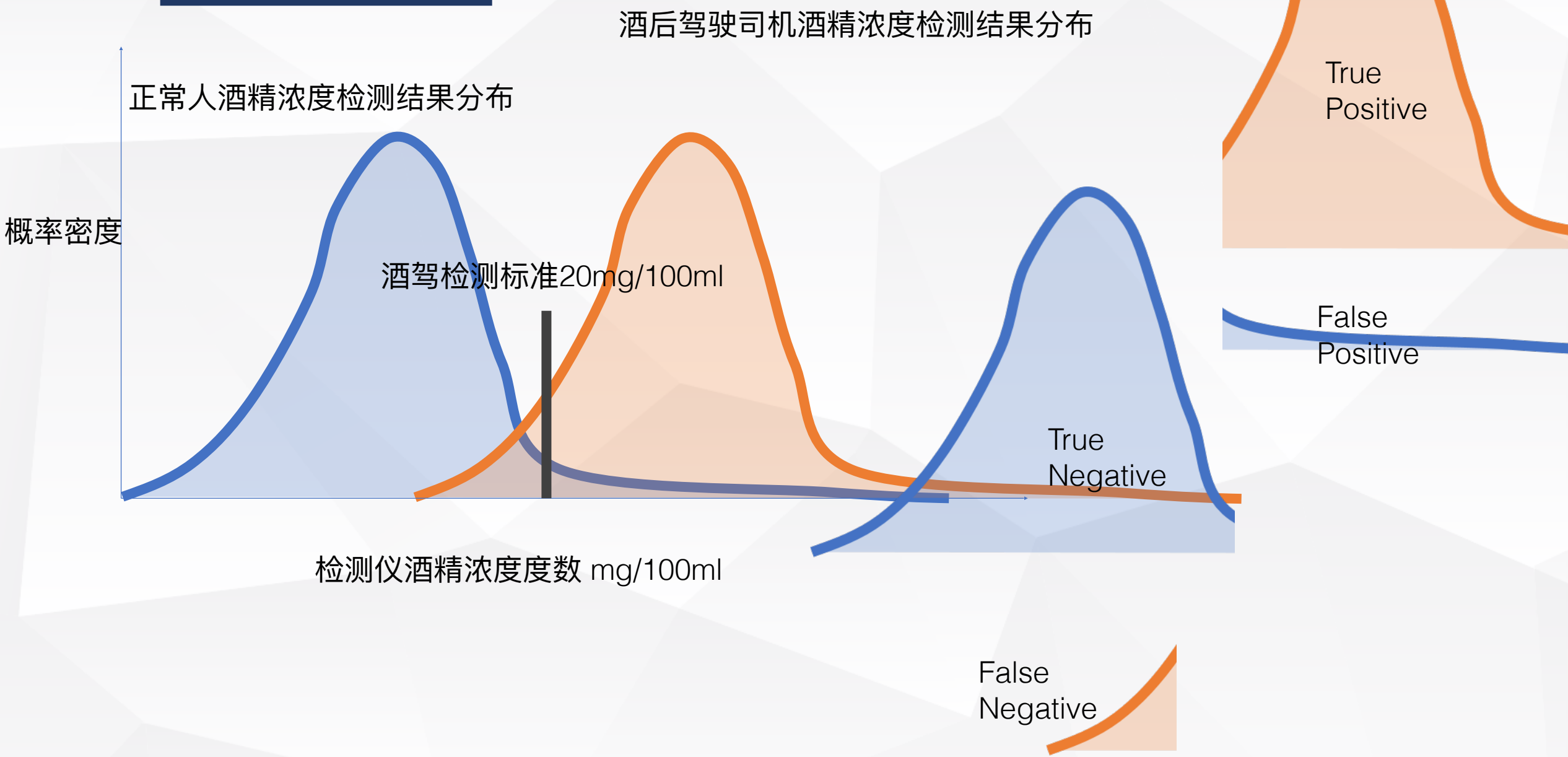
检测仪2结果

召回率: $\mathrm{R} = \frac{B}{A+B}$ = 166/166　准确率: $\mathrm{P} = \frac{B}{C+B}$=166/200

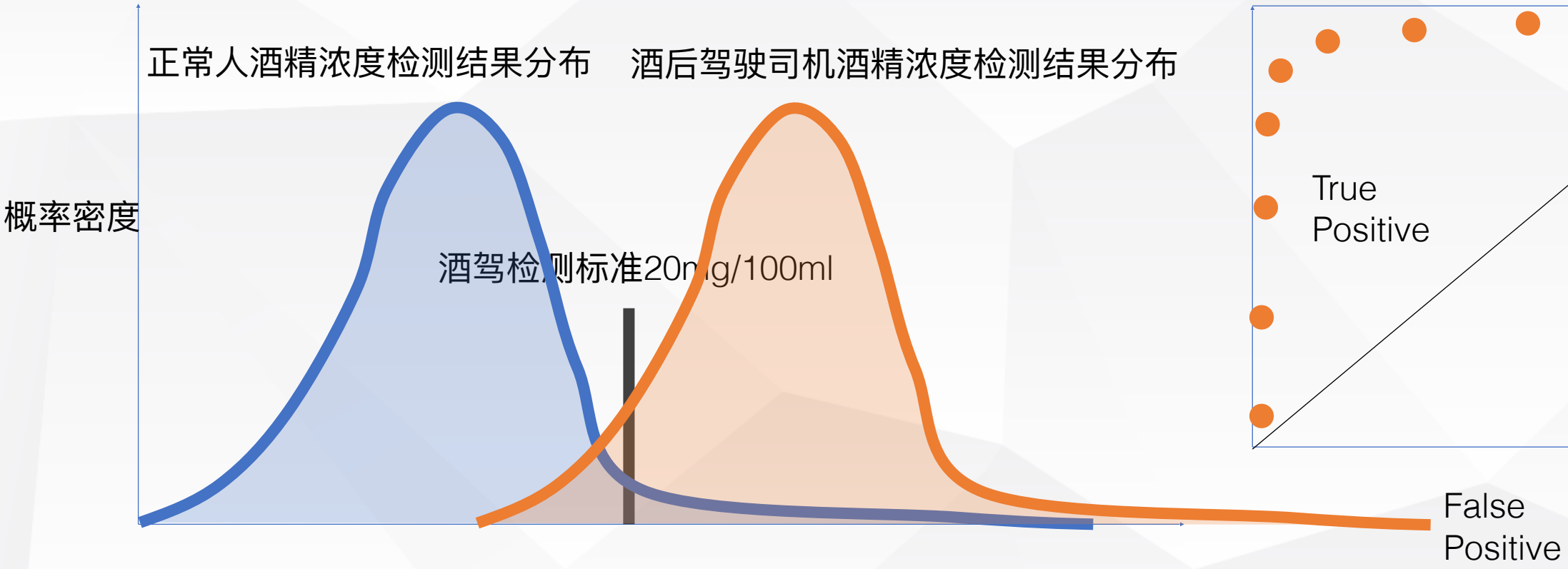
什么是好的分类-ROC曲线



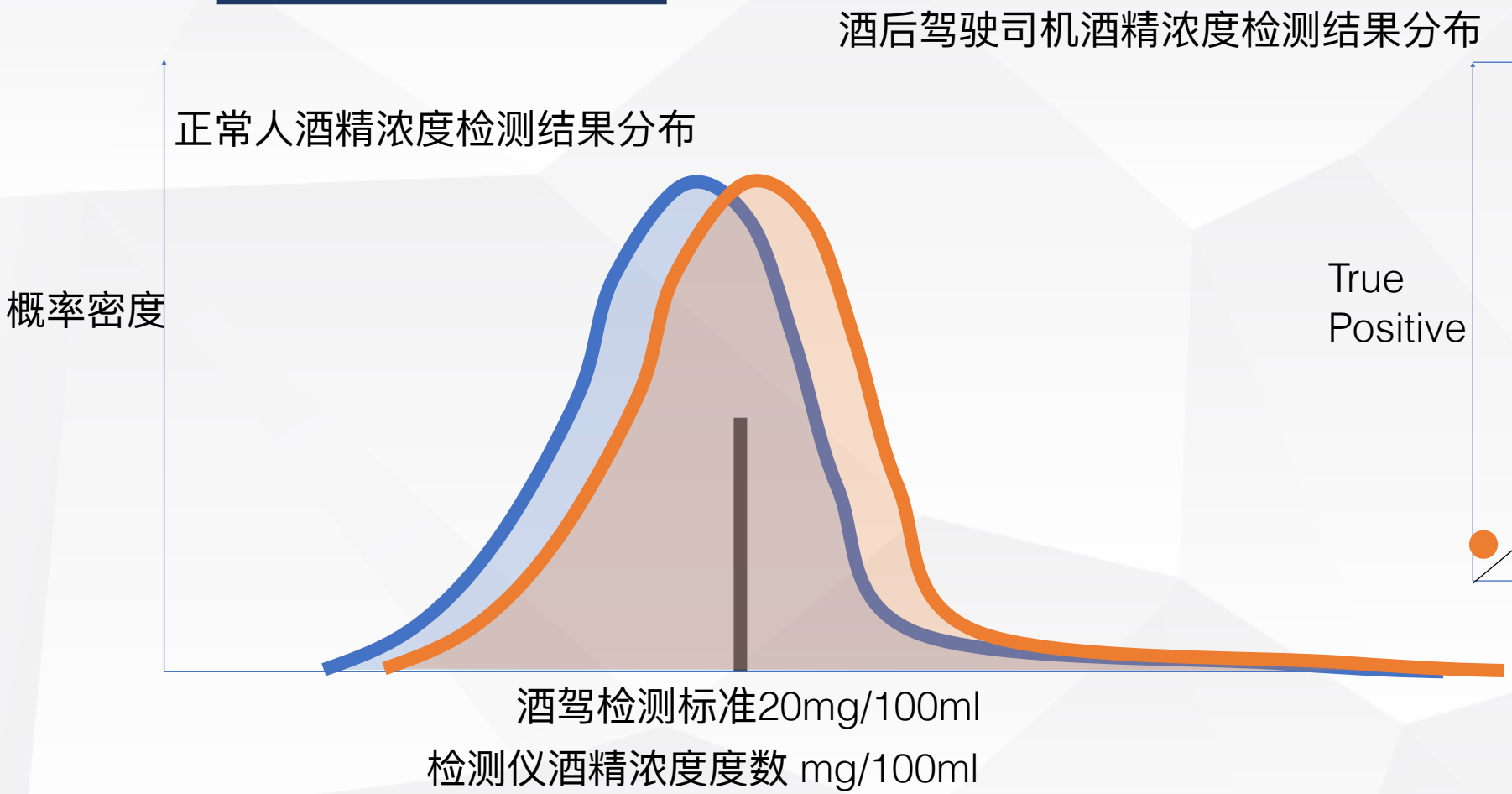
什么是好的分类-ROC曲线

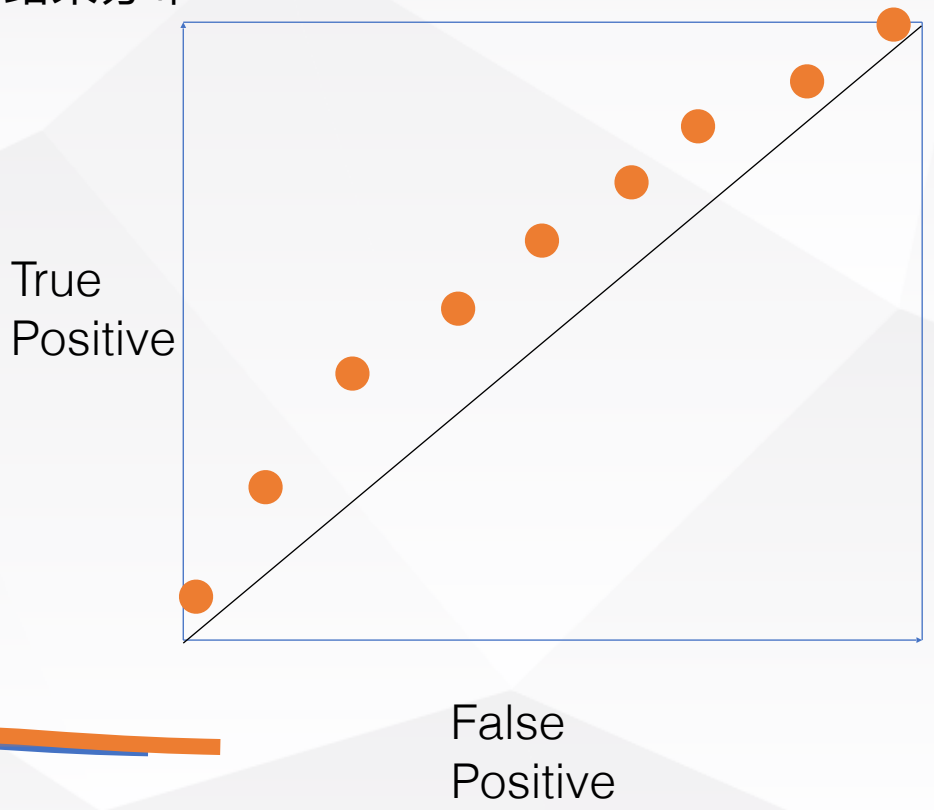


什么是好的分类-ROC曲线



什么是好的分类-ROC曲线





1. **曲线单调增**
2. **每个点都在45度线上方**
3. **这条曲线离45度线越远，表示分类效果越好**
4. **曲线下方的面积，就是** AUC (area under curve)

什么是好的分类-ROC曲线

- 最早用于二战时期衡量雷达识别效果(Receiver Operating Curve)
- 动动手