## Business understanding

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  I'm sure you've all seen the sad and touching love story. On 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1500 out of 2000 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although in such huge disaster, a group of people still survived. The goal of this project is to apply the tools of machine learning to predict which passengers survived the tragedy. And which attributes will lead them have higher survival rate. Most importantly, it also can reveal the nature of society.

## Data Preprocessing

The old is weak compared with young.

Reveal the nature of human.

Something is above the morality under such huge disaster. Data won’t tell lie.

Maybe it’s outlier on the model, but it shows real society.

Outlier:

Tukey

Data don’t tell lie

Just report the positive news, how crew fulfill their duty and stick to the principal.

Missing value:

Money is not everything, but without money, everything is nothing. In some extreme case, money can save life.

Model Building:

# 1. bias an estimator's average error for different training sets.

# 2. variance an estimator indicates how sensitive it is to varying training sets

# 3. noise a property of the data

# first picture-high bias, second-high variance

# underfitting and overfitting

# http://scikit-learn.org/stable/auto\_examples/model\_selection/plot\_underfitting\_overfitting.html#example-model-selection-plot-underfitting-overfitting-py

一般采取判断某函数是高方差还是高偏差，简单的判断是看训练误差与测试误差的差距，差距大说明是高方差的，过拟合。差距小说明是高偏差的。欠拟合

当训练集和测试集的误差收敛但却很高时，为高偏差。

左上角的偏差很高，训练集和验证集的准确率都很低，很可能是欠拟合。

我们可以增加模型参数，比如，构建更多的特征，减小正则项。

此时通过增加数据量是不起作用的。

当训练集和测试集的误差之间有大的差距时，为高方差。

当训练集的准确率比其他独立数据集上的测试结果的准确率要高时，一般都是过拟合。

右上角方差很高，训练集和验证集的准确率相差太多，应该是过拟合。

我们可以增大训练集，降低模型复杂度，增大正则项，或者通过特征选择减少特征数。

理想情况是是找到偏差和方差都很小的情况，即收敛且误差较小。