Information Visualization on Titanic

Li, Xuan
Wang, Lee-Yin
Zhang, Ruonan
Zhou, Chaoran

Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

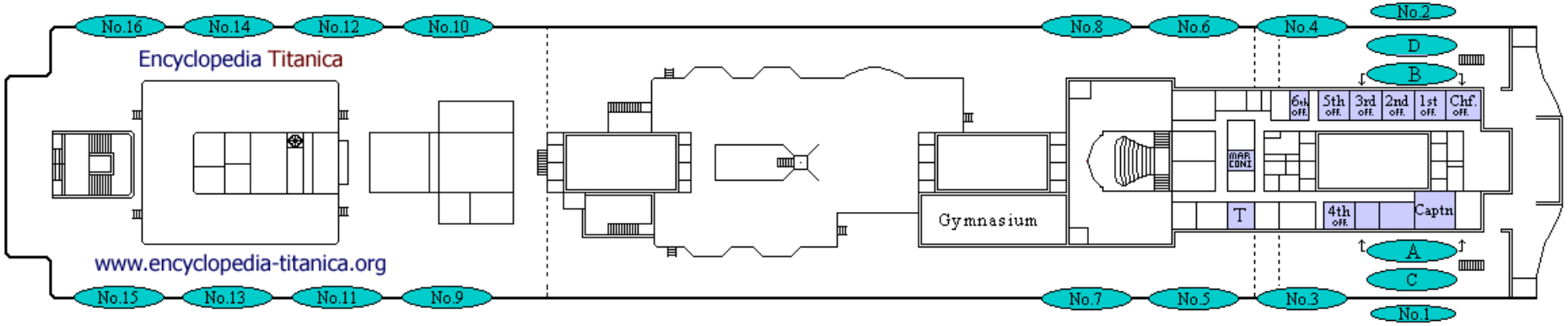
Introduction

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.

64: The number of lifeboats the Titanic was capable of carrying (a total well over the ships maximum capacity of 3547 people)

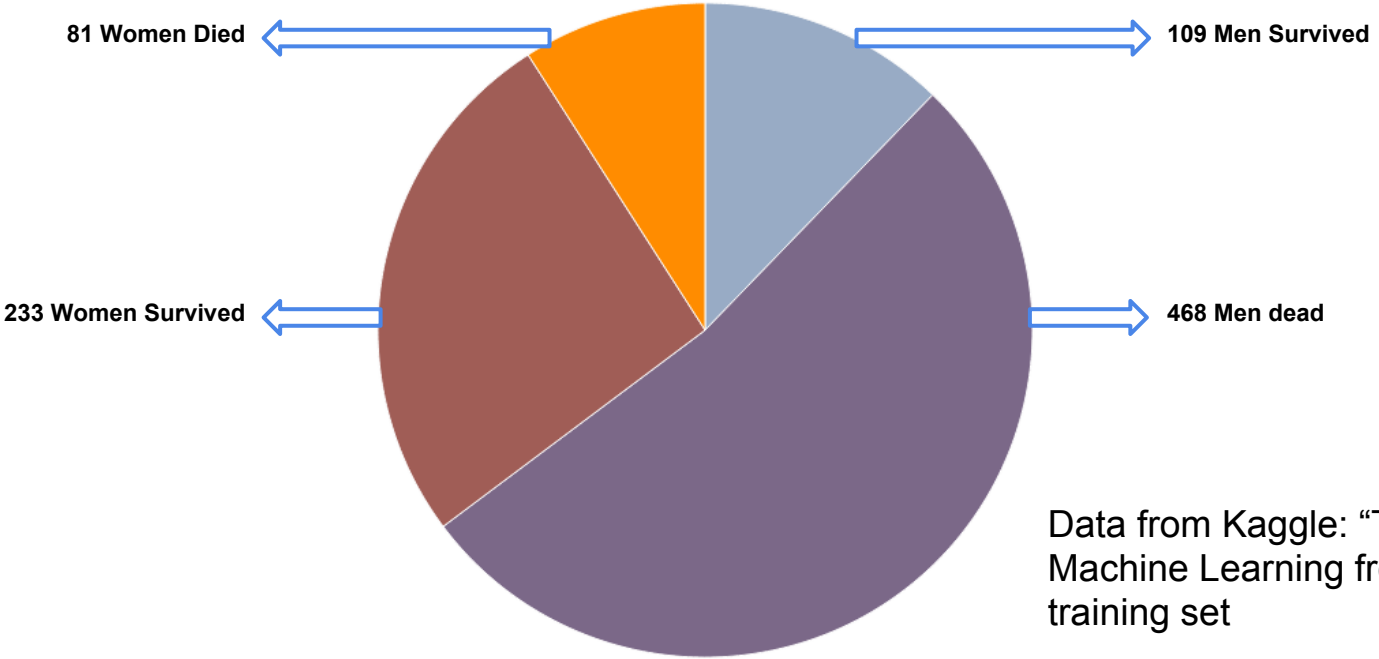
48: The number of lifeboats originally planned for Titanic by the chief designer Alexander

20: The number of lifeboats actually carried aboard



Introduction

Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.



Data from Kaggle: “Titanic: Machine Learning from Disaster” training set

Introduction

This project, our purpose is to present you information that affect surviving the tragedy. We analyzed factors (such as gender, age, number of siblings, etc.), and used D3js to show the correlations between different factors and the survival. Moreover, we made use of information from the training data to design an interactive “prediction” system: Predict survival or not based on passenger’s properties (age, gender, etc.).

Data

The data is from Kaggle website “Titanic: Machine Learning from Disaster” competition. There are a total of **890** records in the training set. After deleting records with missing values, our data contains **714** records with columns:

PassengerID: Unique identifier to distinguish each passenger

Survival: survival of the passenger (0 = No; 1 = Yes)

Pclass: passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

Name: name of the passenger

Sex: gender of the passenger (0 = female; 1 = male)

Age: age of the passenger

Sibsp: number of siblings/spouses aboard

Parch: number of parents/children aboard

Ticket: ticket number

Fare: passenger fare

Cabin: cabin number

Embarked: port of embarkation

Data

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S

Data

We used JMP to generate histograms of variables, e.g.:

fig. 1: conditional distribution of Survival for Female

fig. 2: conditional distribution of Survival for Male

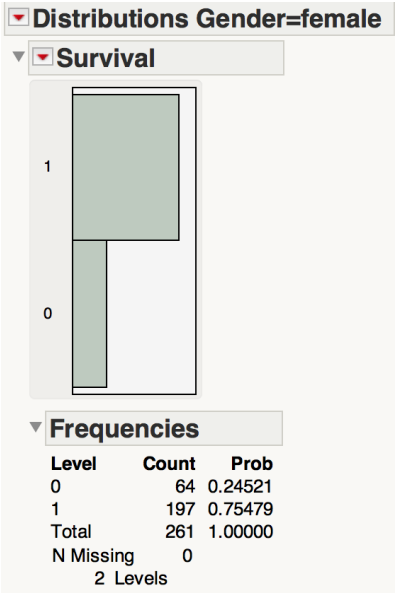


fig. 1

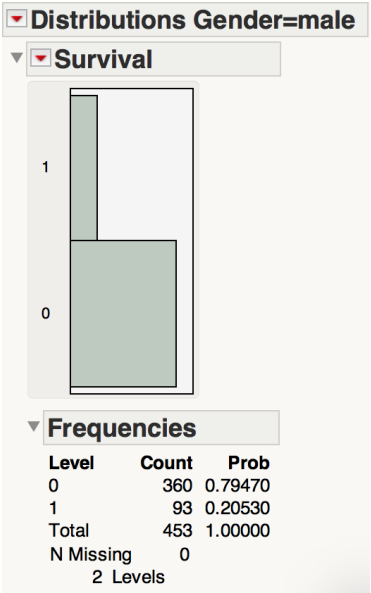


fig. 2

Data

After preprocessing, we kept columns (PID, Survived, Pclass, Sex, Age, Sibsp, Parch) for analysis and visualization. Data looks like as follows:

PID	Survived	Pclass	Sex	Age	Sibsp	Parch
1	0	3	male	22	1	0
2	1	1	female	38	1	0
3	1	3	female	26	0	0
4	1	1	female	35	1	0
5	0	3	male	35	0	0
7	0	1	male	54	0	0
8	0	3	male	2	3	1
9	1	3	female	27	0	2

Data

We applied Logistic Regression for the data. With feature selection, we kept: Pclass, Sex, Age, and Sibsp for our statistical model. And we used 0.5 as the cut-off point. The purpose for building the statistical model is for an interactive design.

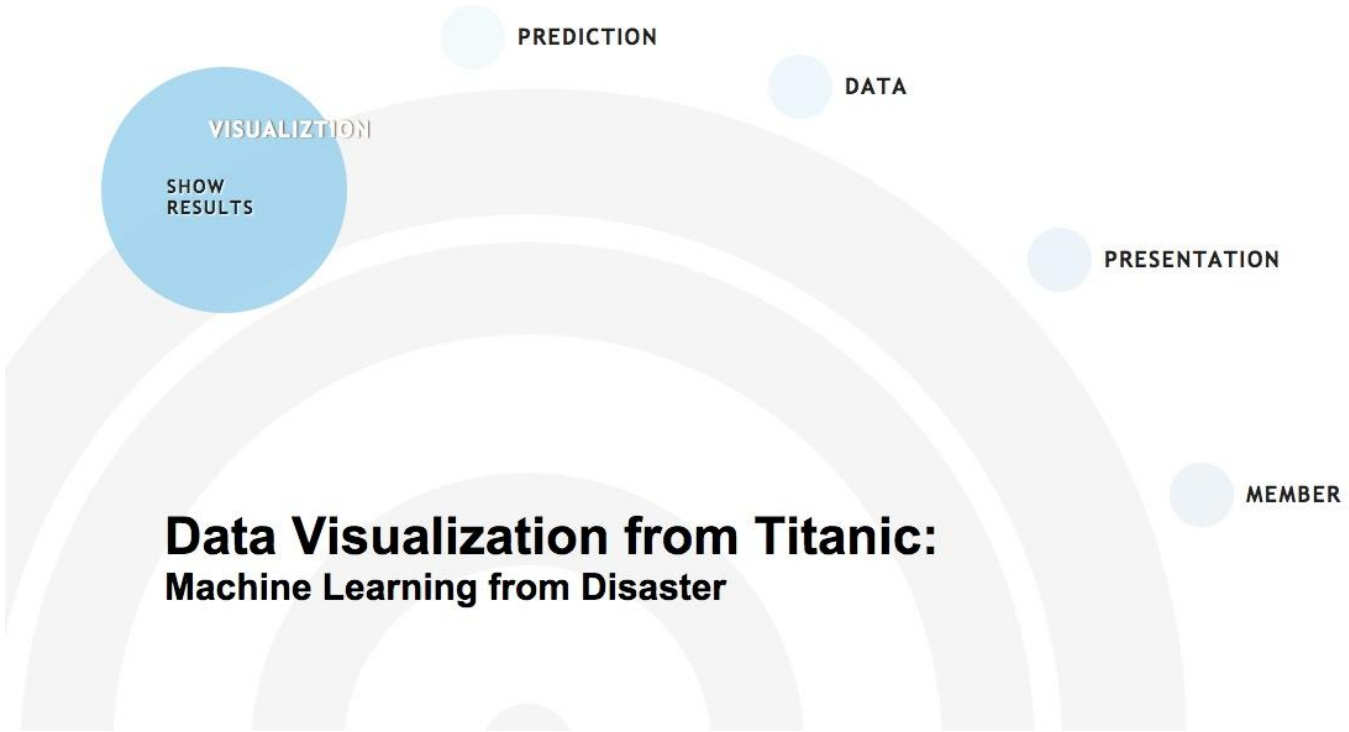
Visualization

Our design is comprised of two parts. The first part is data representation. We show passengers as nodes, and use the ship as the background. Nodes are divided above and under the sea (Survival), and are grouped based on their ticket class (1st, 2nd, 3rd). And are colored based on their gender.

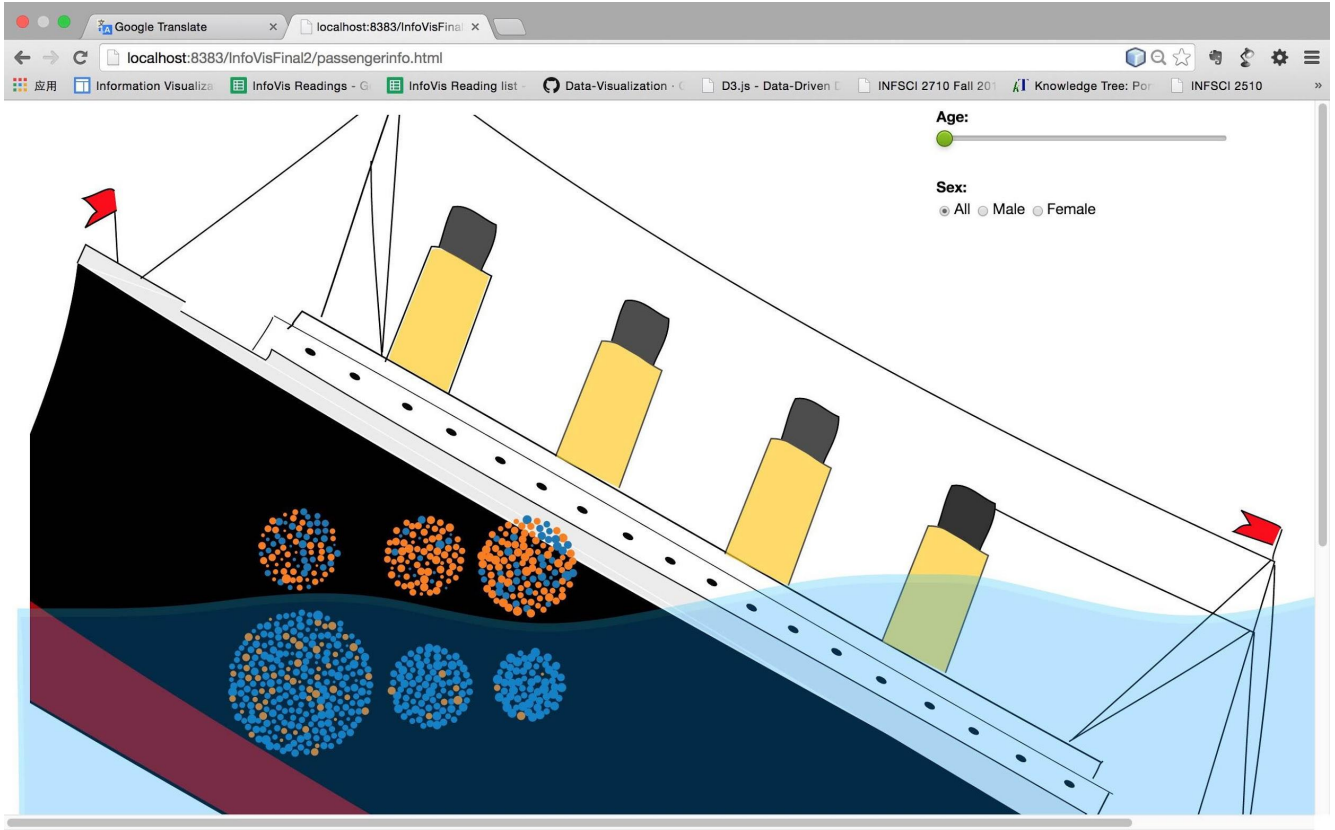
Visualization

The second part is an interactive design. We use the logistic regression model as the classifier. We can put values for these features, and web will return you the predicted result.

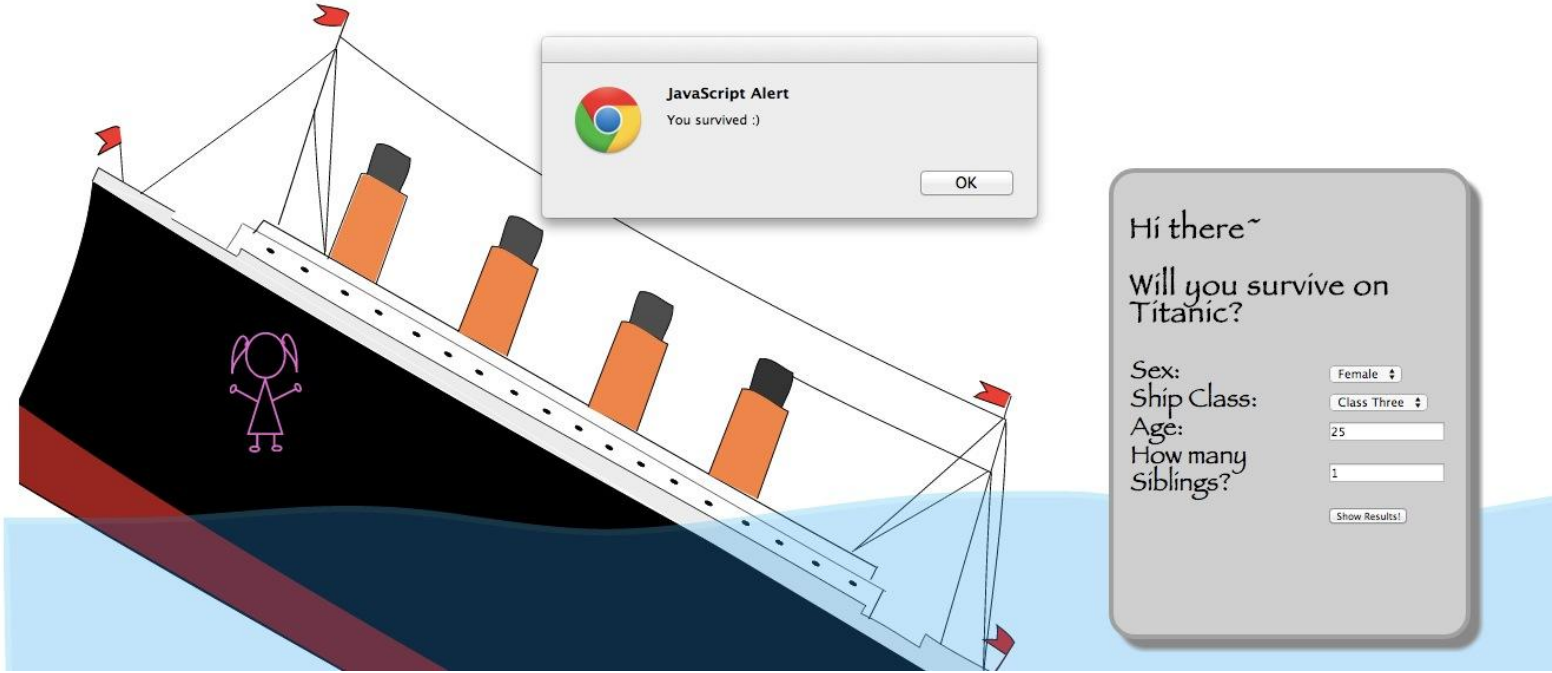
Visualization



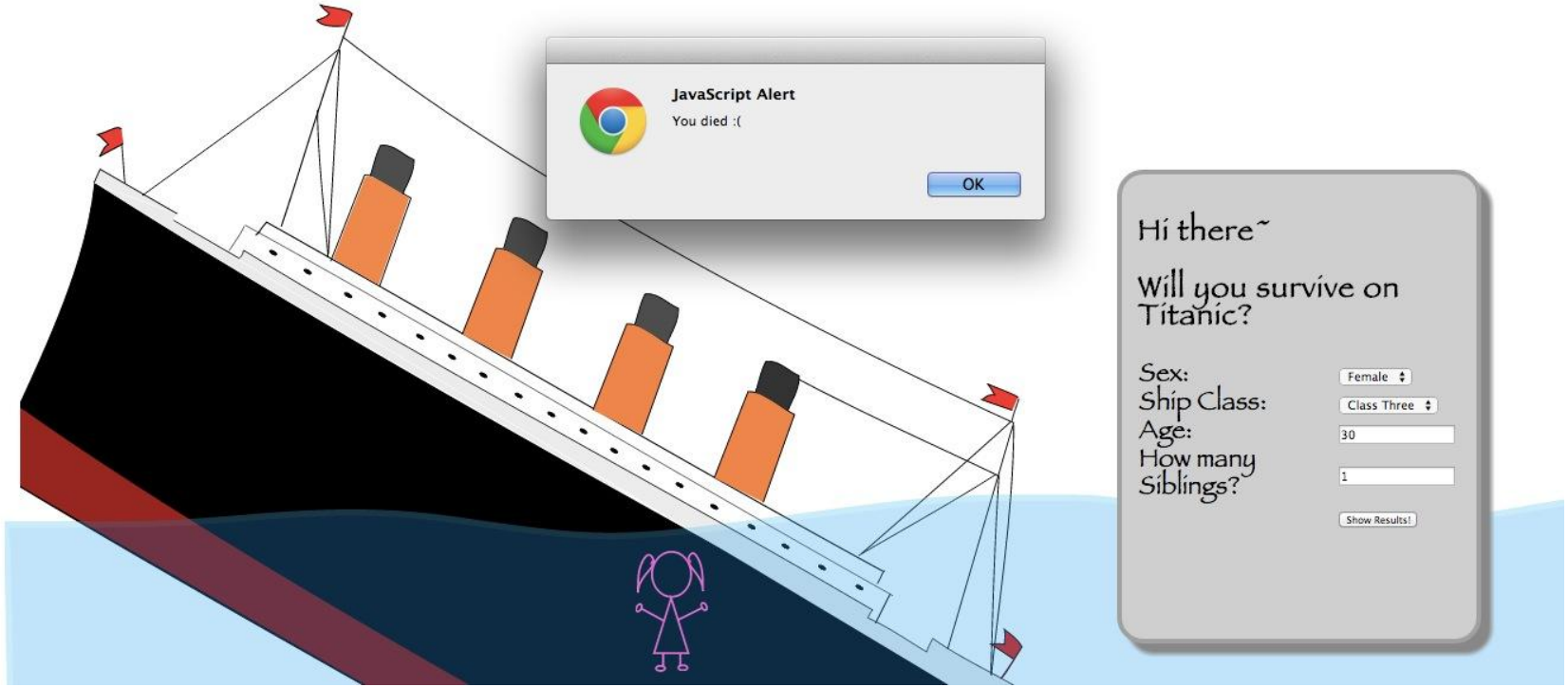
Visualization



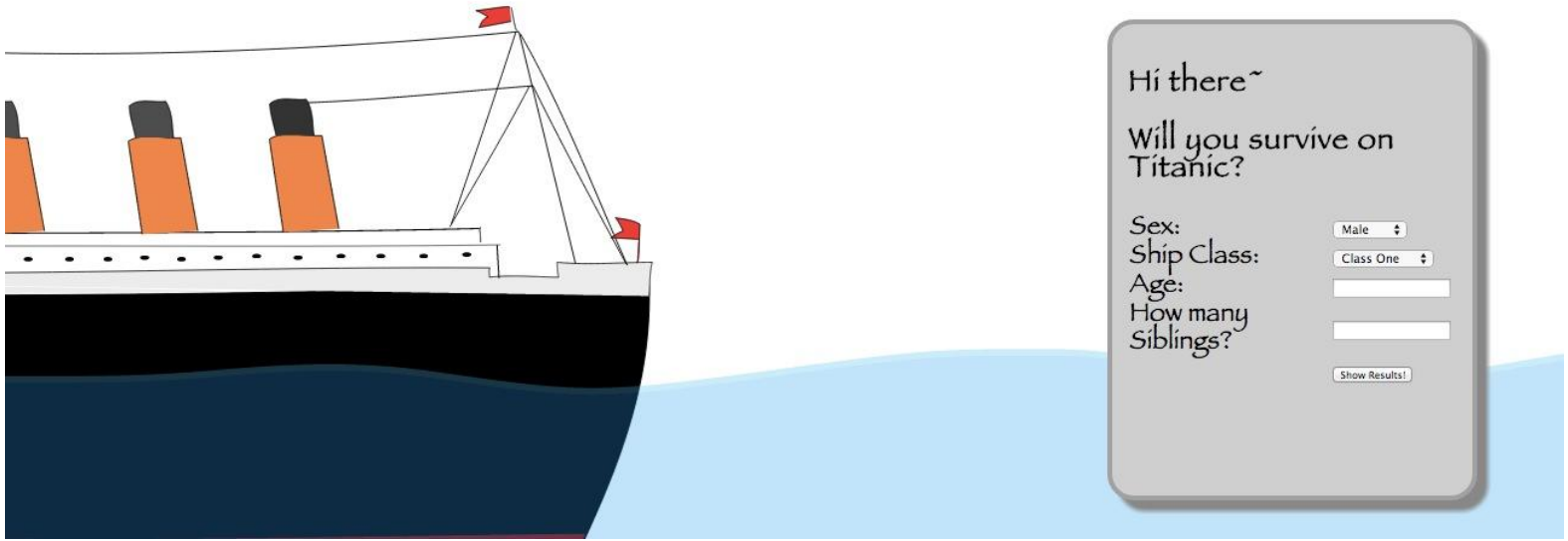
Visualization



Visualization



Visualization



Visualization

See Demo

Future work

- Model testing: We could test the performance of our model by applying it to the testing set, and submit the result to Kaggle website.
- Backend Data Modeling: We could also try some other classification methods, such as Naive Bayesians, SVM, etc.

Questions

Thank you !