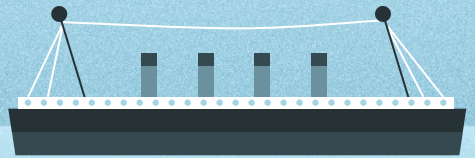
— KTH, ID2221, P1, Fall 2023—

TITANIC

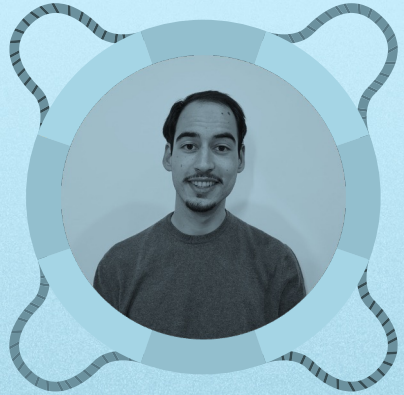
Spark and Machine Learning from disaster



— Our survivors —

Giovanni Manfredi

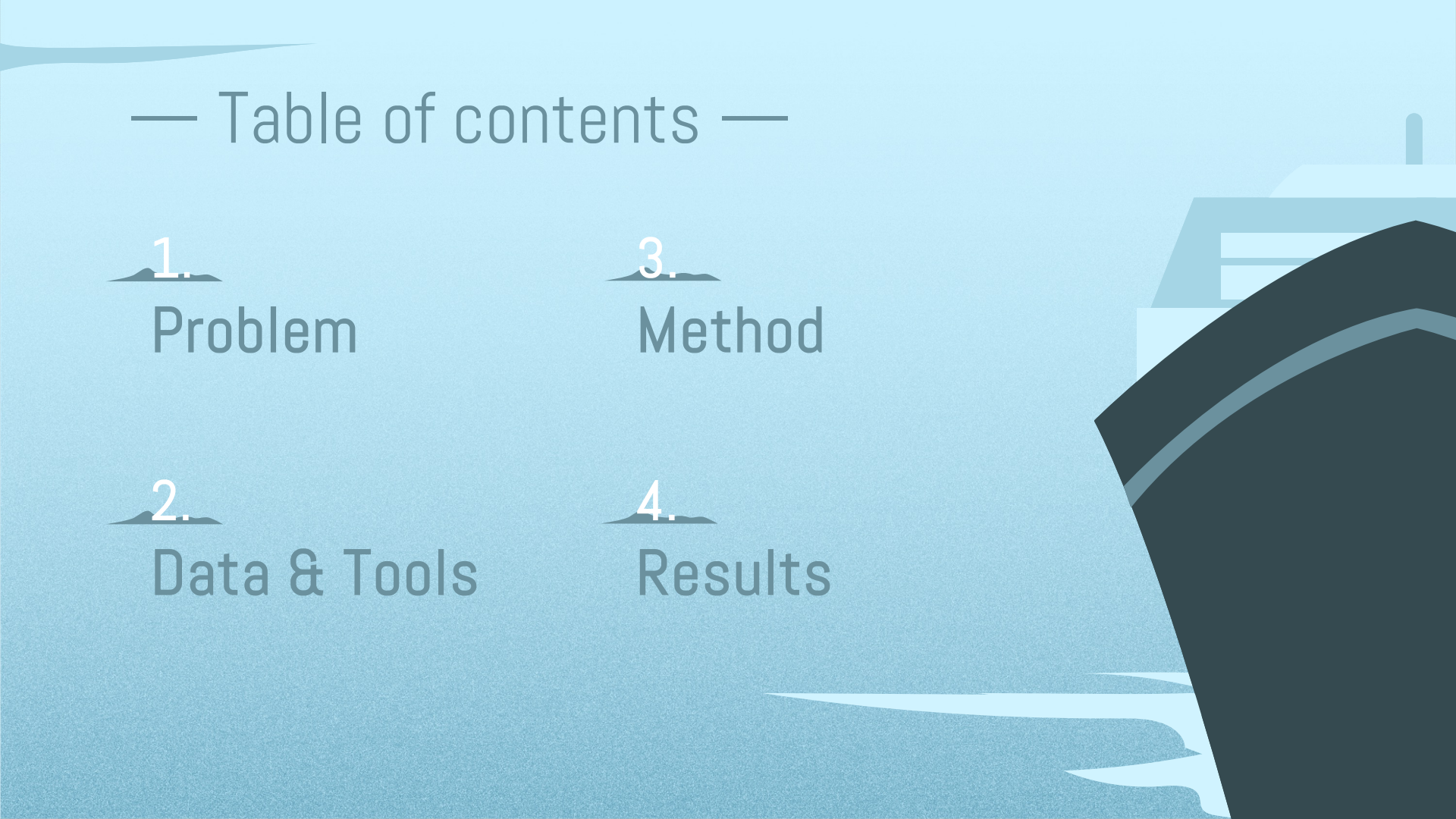
gioman@kth.se



Sebastiano Meneghin

meneghin@kth.se

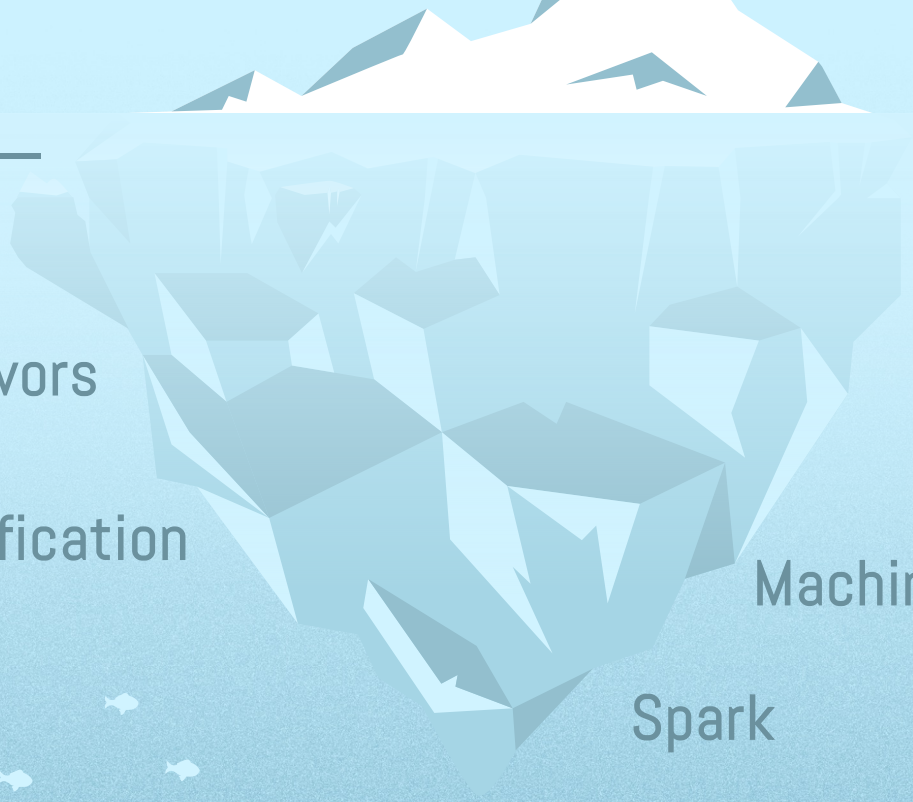
— Table of contents —



— Problem —

Predict the survivors

Classification



Machine Learning

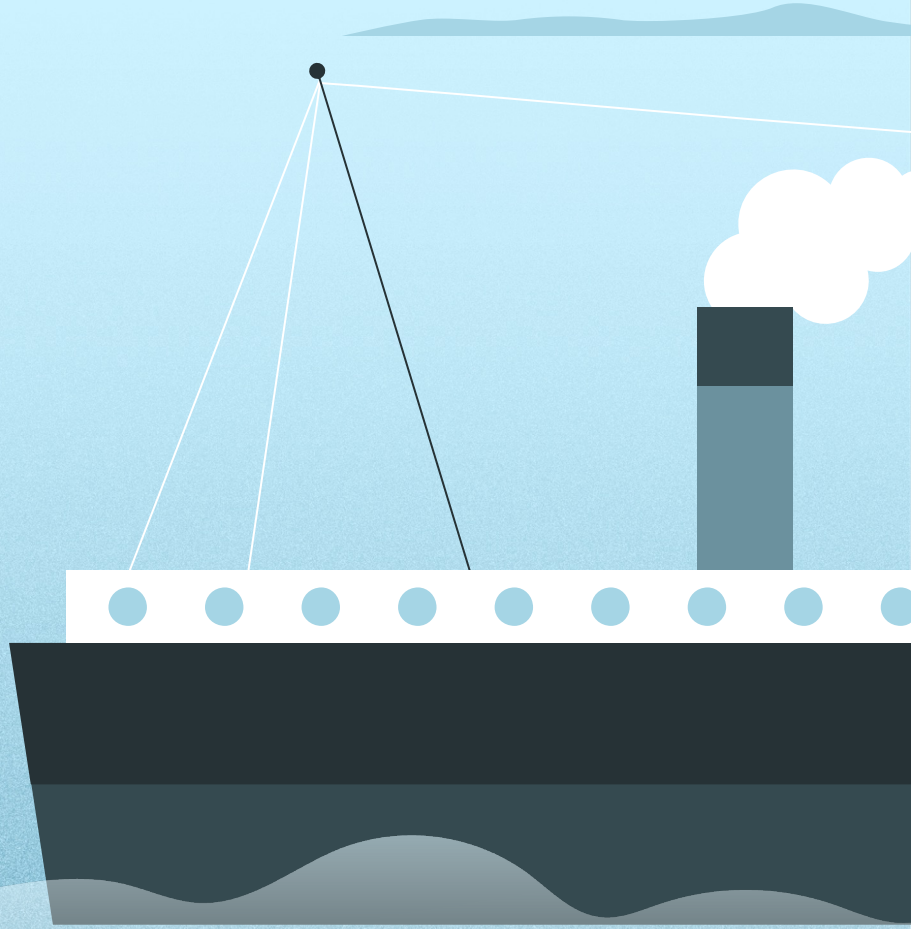
Spark

— Data & Tools —

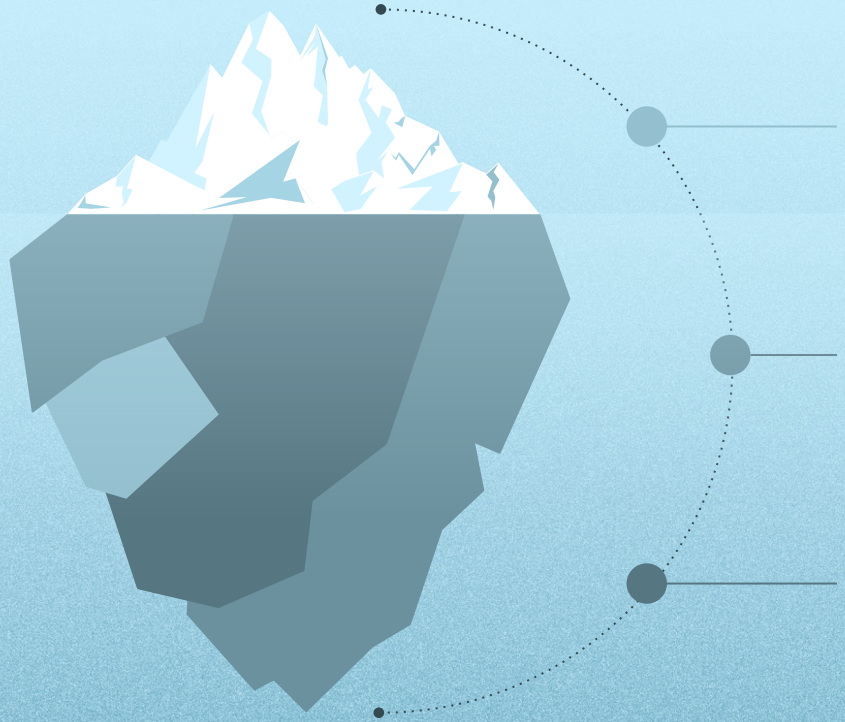
databricks

APACHE Spark™

kaggle



— Method —



Data analysis

Understanding the data set and types, discovering correlations or their lack.

Data pre-processing

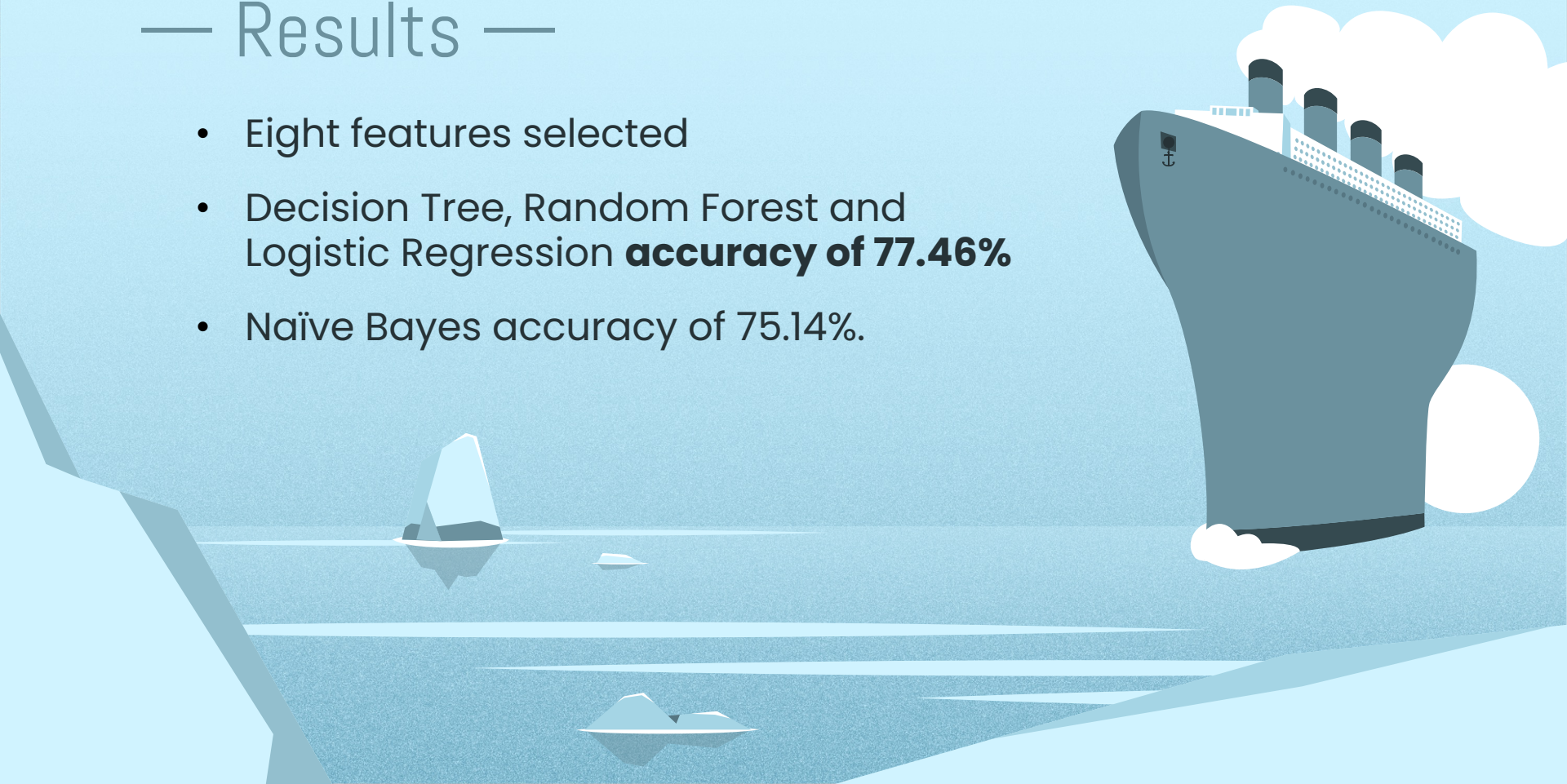
Feature selection, data completion, data correction, data conversion.

Model

Model training, model evaluation and model selection.

— Results —

- Eight features selected
- Decision Tree, Random Forest and Logistic Regression **accuracy of 77.46%**
- Naïve Bayes accuracy of 75.14%.



— Resources —

Project soon published on the repo:
https://github.com/Silemo/dic-2023-manfredi-meneghin

Find out more about Titanic's history on:
https://en.wikipedia.org/wiki/Titanic