

# Get the Materials!

- GitHub Repo:
  - [https://github.com/datasciencedojo/meetup/tree/master/intro to ml with r and caret](https://github.com/datasciencedojo/meetup/tree/master/intro%20to%20ml%20with%20r%20and%20caret)
- Kaggle Titanic Competition:
  - <https://www.kaggle.com/c/titanic>

# An Intro to Machine Learning with R and caret



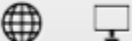





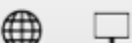

June 7th, 2017

# Who Am I?

- Dave Langer, VP of Data Science – Data Science Dojo
- 20+ years in technology:
  - Roles in development, architecture, & BI/DW/analytics.
  - Last job – Sr. Director, BI & Analytics @ Microsoft.
- Hooked on Data Science 5 years ago:
  - Extensive background in data and analytics.
  - Learned Machine Learning from 2<sup>nd</sup> place Netflix Prize winner.
- Joined Data Science Dojo to democratize Data Science:
  - More tutorials on our YouTube channels!

# Motivation

## The IEEE's 2016 Ranking of most Popular Programming Languages

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

R has experienced rapid YoY increases in popularity.

This is remarkable as R is a specialized language for data and analytics!

<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>

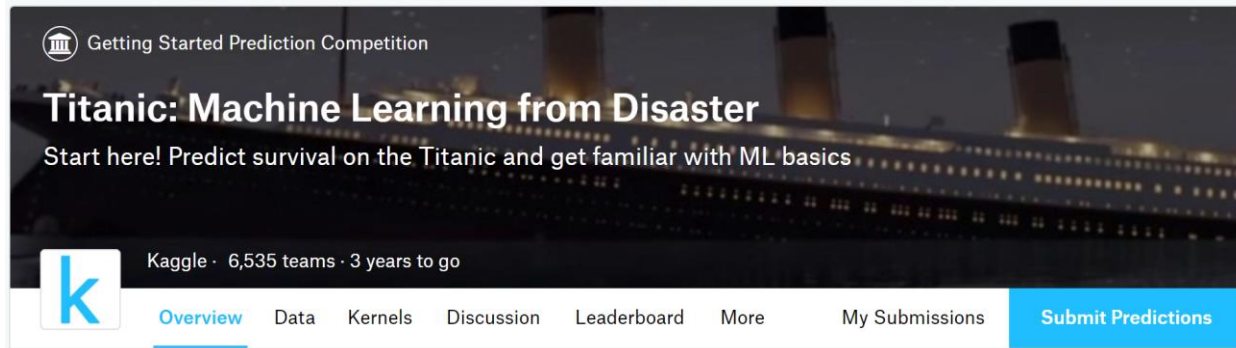
# Expectation Setting

- I am assuming the following:
  - You are experienced with R coding.
  - You are experienced with machine learning concepts.
  - You are interested in learning more about how caret can accelerate your ML work.
- This is a quick intro to ML with R and caret:
  - I will gloss over a lot of things (e.g., feature engineering).
  - I will illustrate some “art of the possible”.
  - More in-depth coverage is available via resources I will mention later.
- My goal is to make you excited about doing ML with caret!

# Prerequisites

- To follow along you will need the following:
  - R
  - Rstudio
- The following R packages are required to follow along:
  - e1071, caret, doSNOW, ipred, and xgboost
- The GitHub repo has source, data, and slide files.

# The Data



Why use  
this  
dataset?

1. Everyone is familiar with the problem domain.
2. It is a good proxy for common business data – for example, customer profile data.

# The Data

## Data Dictionary

Variable	Definition
survival	Survival
pclass	Ticket class
sex	Sex
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation

### Key

0 = No, 1 = Yes

1 = 1st, 2 = 2nd, 3 = 3rd

C = Cherbourg, Q = Queenstown, S = Southampton



# caret

- Caret is an acronym for **[C]**lassification **[A]**nd **[RE]**gression **[T]**raining.
- Written and maintained by Max Kuhn, caret has become a go-to package for ML in R.
- Caret accelerates your ML work in R by providing functions for:
  - Working with hundreds of ML algorithms using a common interface.
  - Data splitting/sampling.
  - Feature selection.
  - Model tuning.
  - Etc.

**LET'S LEARN SOME CARET!**

# QUESTIONS

# APPENDIX

# Get the Materials!

- GitHub Repo:
  - [https://github.com/datasciencedojo/meetup/tree/master/intro to ml with r and caret](https://github.com/datasciencedojo/meetup/tree/master/intro%20to%20ml%20with%20r%20and%20caret)
- Kaggle Titanic Competition:
  - <https://www.kaggle.com/c/titanic>

# Resources

- caret web site:
  - <http://topepo.github.io/caret/index.html>
- A most excellent book (highly recommended)
  - <https://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/>

