

How did I become a Data Scientist in Portugal?



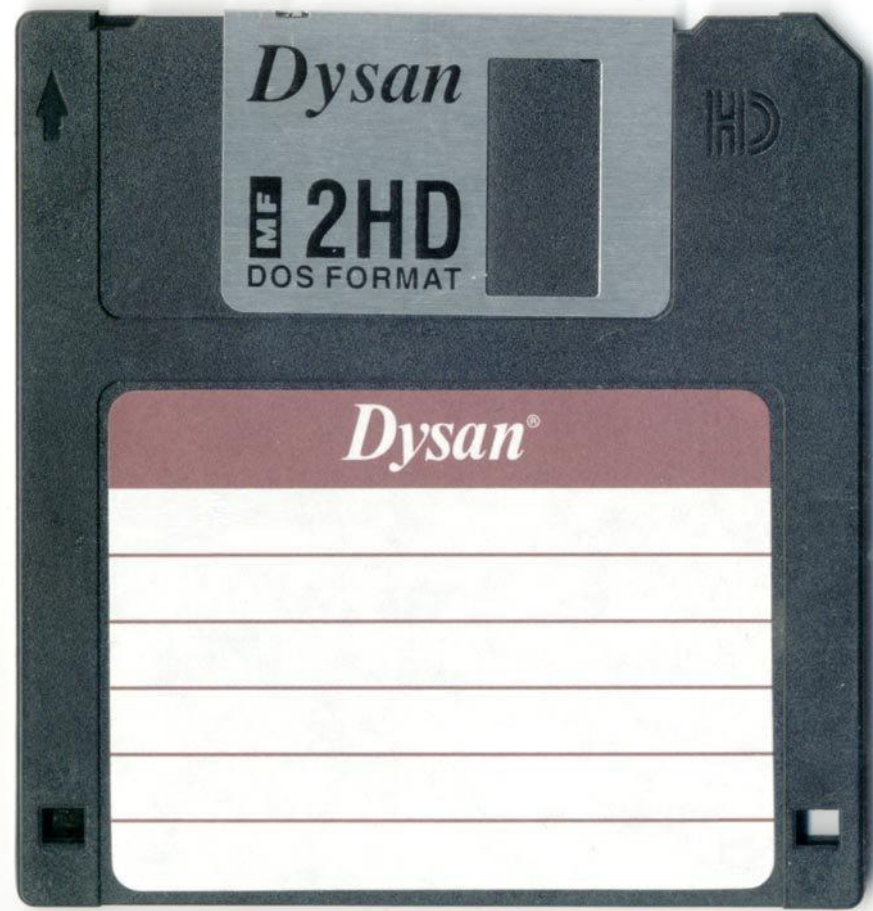


ABOUT ME

Data Scientist
@ Singularity Digital Enterprise

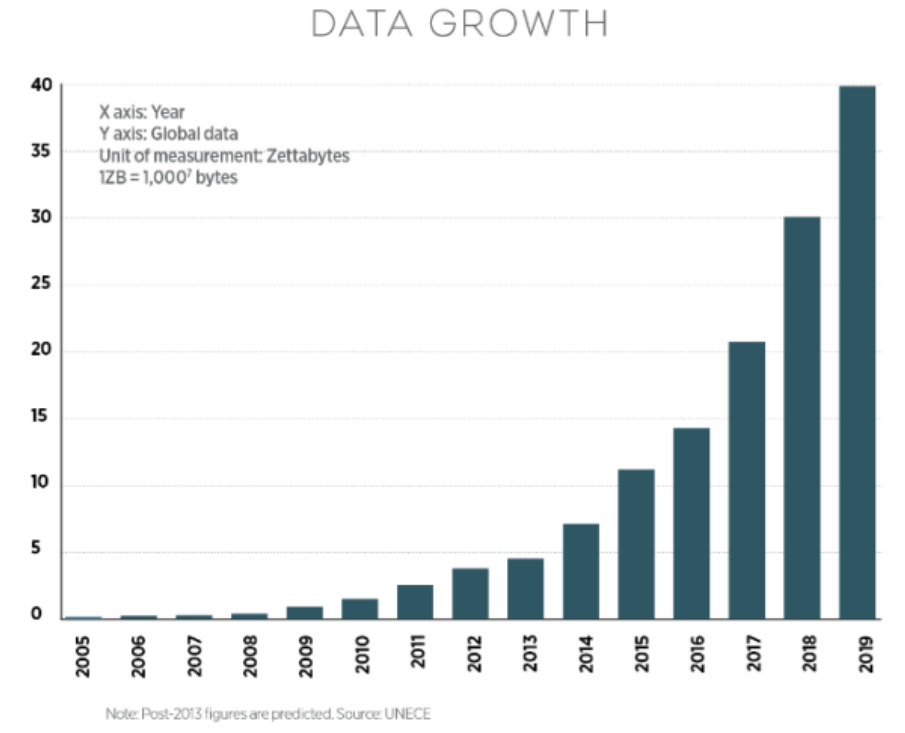
Coorganizer
@ R-Ladies Lisboa

e.mirotshnik@gmail.com
elizabeth@rladies.org



Global Data Growth

“The United Nations Economic Commission for Europe predicts that data growth will be 350% higher in 2019 than it is in 2015.”



Data \neq Information
Information = Value

What is Data Science?



Josh Wills

@josh_wills



Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.



Reply



Retweet



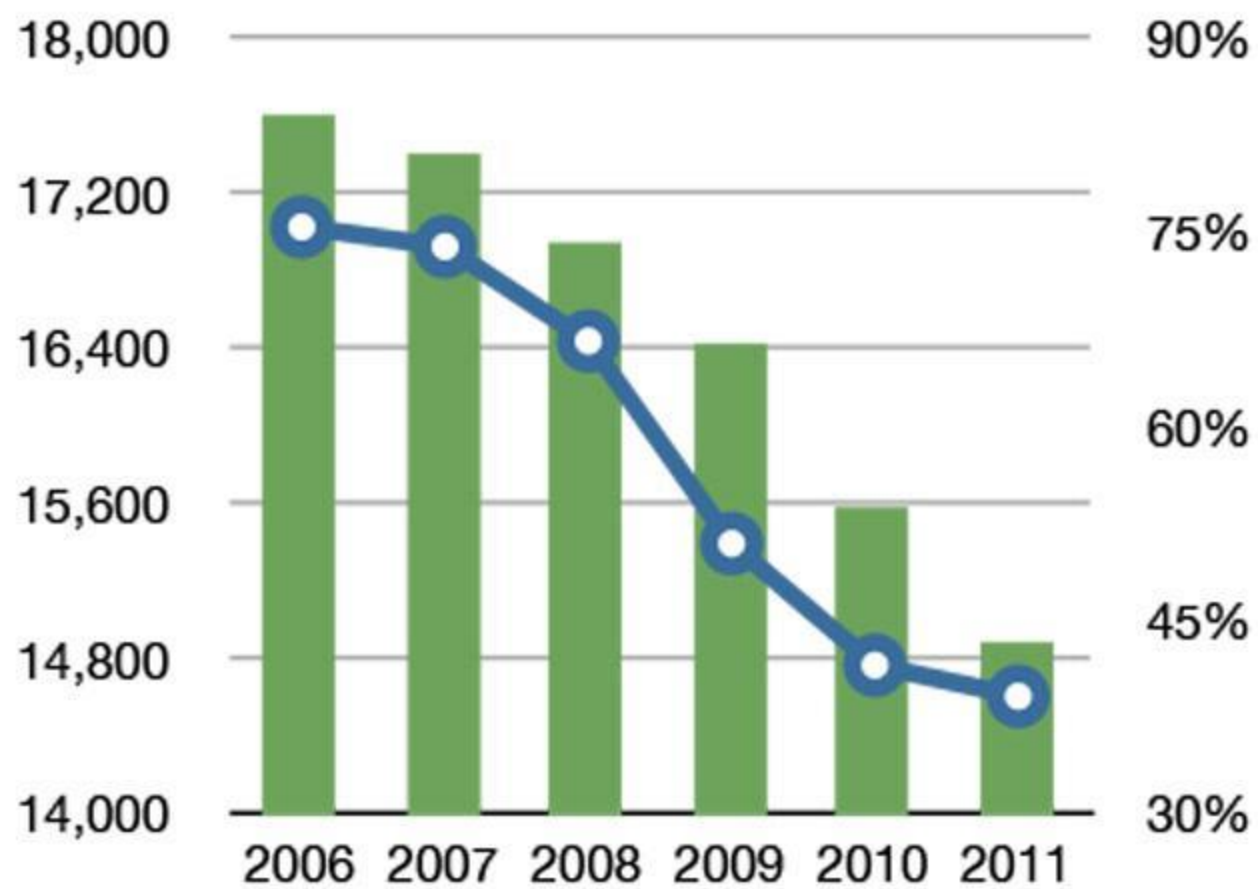
Favorite



More

9:55 AM - 3 May 12

Internet Explorer vs Murder Rate



Murders in US

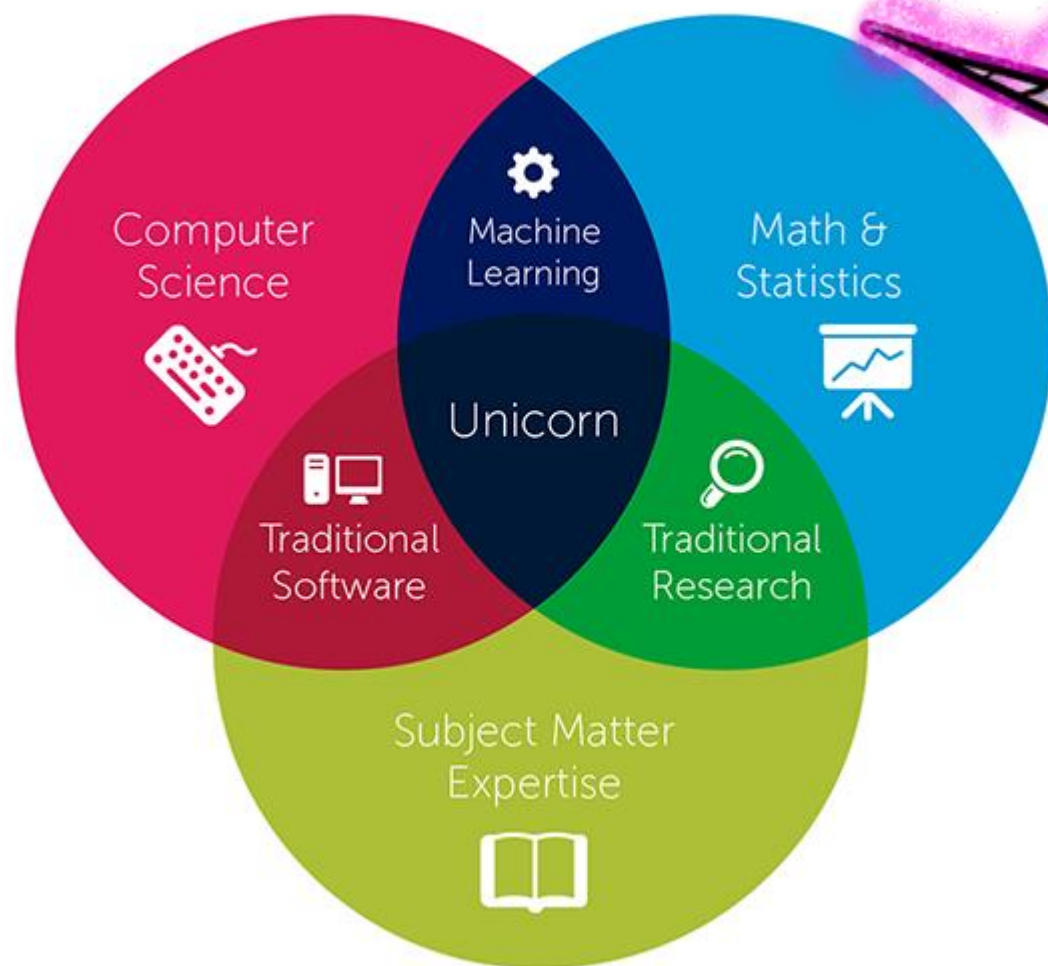


Internet Explorer Market Share

Salary differences between developers who use tabs and spaces

From 12,426 respondents in the 2017 Developer Survey results





How did I become a Data Scientist?



Data Scientists need to be comfortable
with mathematics & statistics

Mathematics

Statistical Analysis

**Linear Algebra
(Matrixes, etc.)**

**Distributions
(Binomial, Poisson, etc.)**

**Calculus
(Derivatives, etc.)**

**Summary Statistics
(Mean, Variance, etc.)**

**Probability and
Combinatorics**

Hypothesis Testing

Graph Teory

Bayesian Analysis

Mathematics

**Linear Algebra
(Matrixes, etc.)**

**Calculus
(Derivatives, etc.)**

**Probability and
Combinatorics**

Graph Teory



Statistical Analysis

Distributions
(Binomial, Poisson, etc.)

Summary Statistics
(Mean, Variance, etc.)

Hypothesis Testing

Bayesian Analysis



UDACITY

Intro to Statistics

<https://www.udacity.com/course/intro-to-statistics--st101>

Statistics

<https://www.udacity.com/course/statistics--st095>



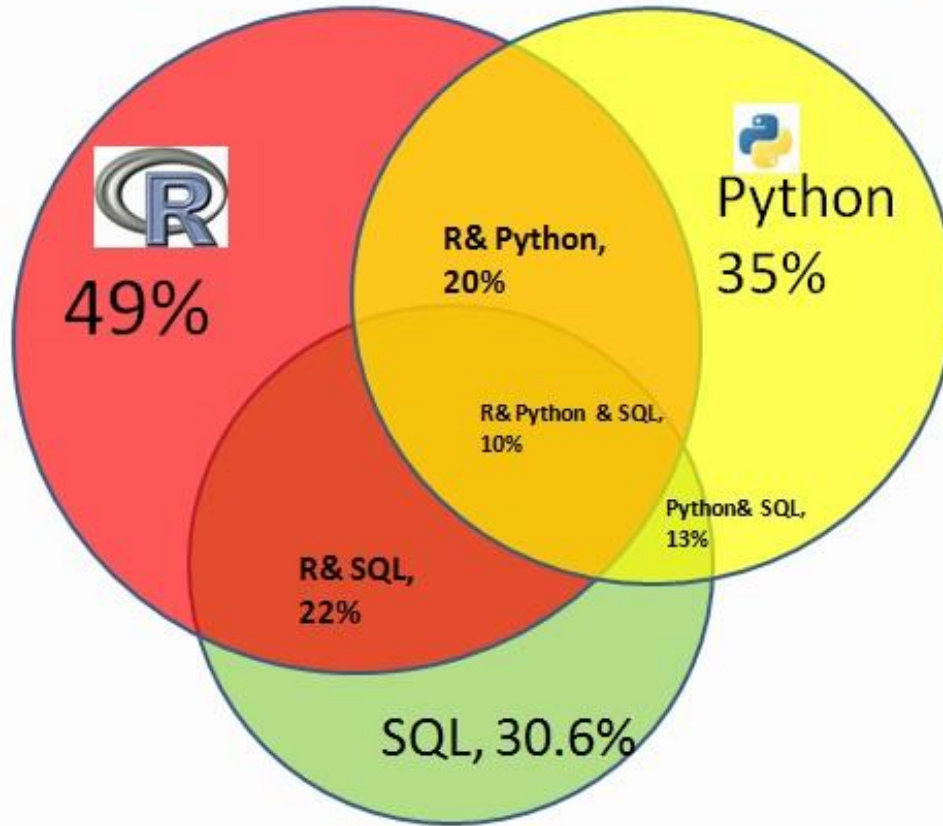
Statistics and Probability

<https://www.khanacademy.org/math/statistics-probability>

```
code.learn(you);
```

Data Scientists need to know how to code

KDnuggets 2014 Poll: Languages used for Analytics/Data Mining



Learn to code: Programming Languages

Where to learn for free?



Python

<https://www.codecademy.com/learn/python>

Javascript

<https://www.codecademy.com/learn/javascript>

Learn SQL

<https://www.codecademy.com/learn/learn-sql>

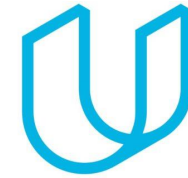


R Programming

<https://www.coursera.org/learn/r-programming>

Programming for Everybody (Getting Started with Python)

<https://www.coursera.org/learn/python>



UDACITY

Programming Foundations with Python

<https://www.udacity.com/course/programming-foundations-with-python--ud036>

Intro to Relational Databases

<https://www.udacity.com/course/intro-to-relational-databases--ud197>



Introduction to Computer Science and Programming

<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-00-introduction-to-computer-science-and-programming-fall-2008/index.htm>

Data Scientist need to
know how to do
Exploratory Analysis



What is EDA?

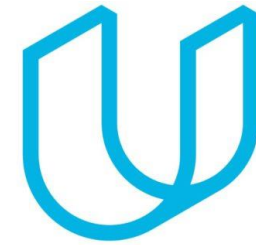
“In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.”

Where to learn for free?



Exploratory Data Analysis

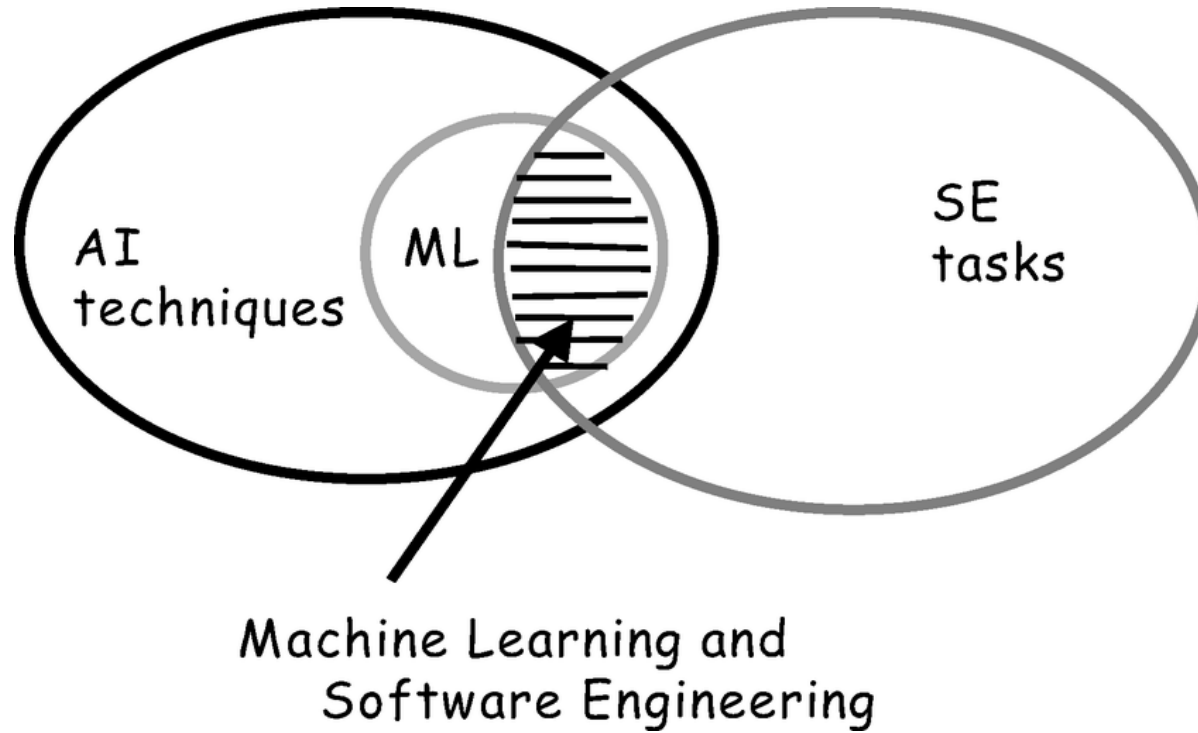
<https://www.coursera.org/learn/exploratory-data-analysis>



UDACITY

Data Analysis with R

<https://www.udacity.com/course/data-analysis-with-r--ud651>



Data Scientist need to know
Machine Learning and Software
Engineering

Machine Learning

**Supervised Learning
(SVM, NB, RF, etc.)**

**Unsupervised
(K-Means, etc.)**

**NLP / Information
Retrieval**

**Validation / Model
Comparison**



UDACITY

Introduction to Machine Learning

<https://www.udacity.com/course/intro-to-machine-learning--ud120>



DataCamp

Kaggle R Tutorial on Machine Learning

<https://www.datacamp.com/community/open-courses/kaggle-tutorial-on-machine-learning-the-sinking-of-the-titanic#gs.Q8Mgsaw>

Kaggle Python Tutorial on Machine Learning

https://www.datacamp.com/community/open-courses/kaggle-python-tutorial-on-machine-learning#gs.qfvND_o

Software Engineering

Algorithms & Data
Structure

Data Visualization

Data Munging

Distributed
Computing



UDACITY

**Intro to Hadoop and
MapReduce**

<https://www.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617>

**Data Visualization and
D3.js**

<https://www.udacity.com/course/data-visualization-and-d3js--ud507>



DataCamp

coursera

Getting and Cleaning Data

<https://www.coursera.org/learn/data-cleaning>



**Introduction to Computer
Science and Programming**

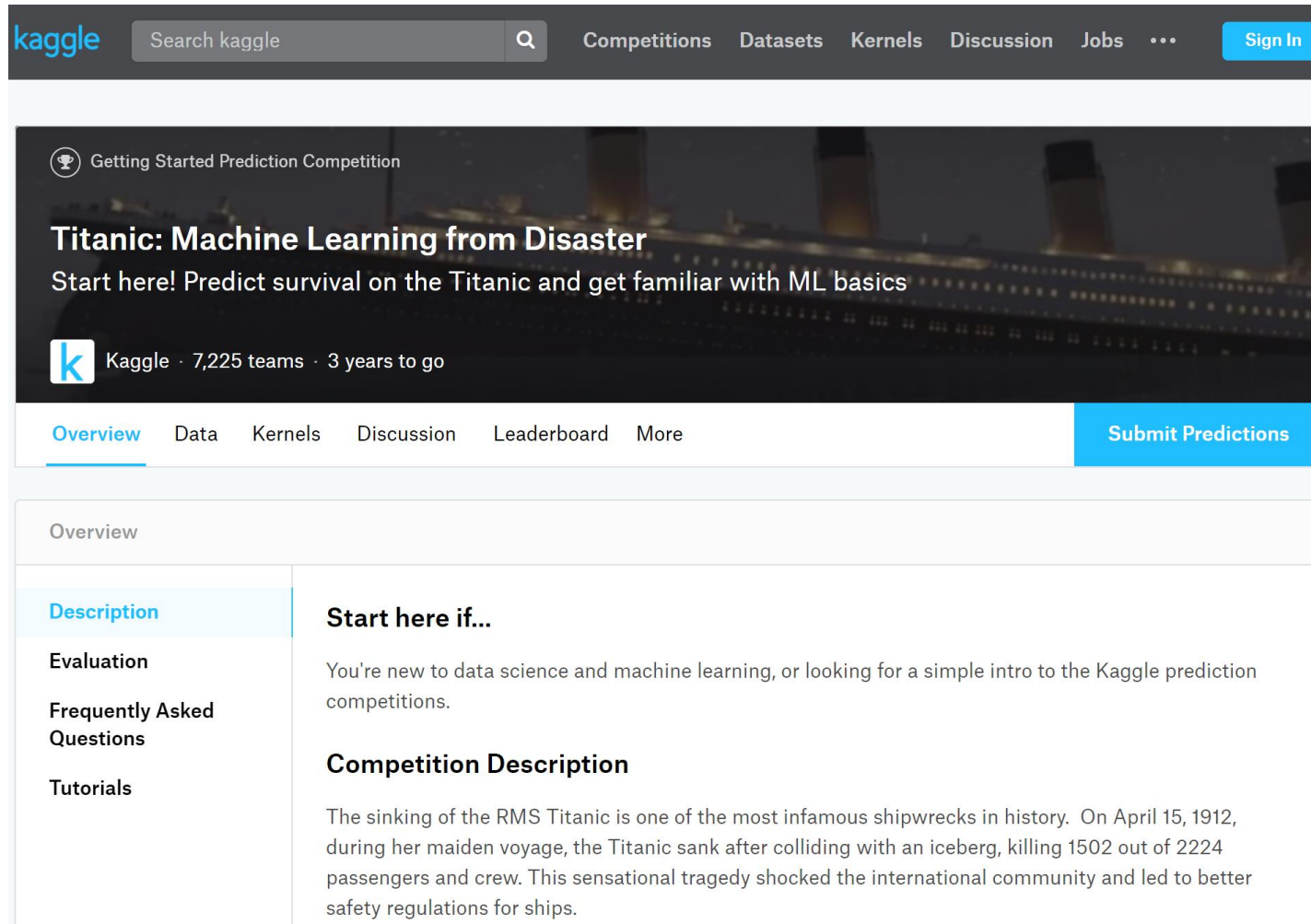
<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-00-introduction-to-computer-science-and-programming-fall-2008/index.htm>

**Introduction to Data Visualization with
Python**

<https://www.datacamp.com/courses/introduction-to-data-visualization-with-python>

The best way to learn Data Science
is to do Data Science.

Kaggle



The screenshot shows the Kaggle website interface. At the top is a dark navigation bar with the Kaggle logo, a search bar, and links for Competitions, Datasets, Kernels, Discussion, Jobs, and a Sign In button. Below this is a large banner for the 'Titanic: Machine Learning from Disaster' competition, featuring a night-time image of the Titanic. The banner includes the title, a subtitle 'Start here! Predict survival on the Titanic and get familiar with ML basics', and the Kaggle logo with text indicating 7,225 teams and 3 years to go. Below the banner is a horizontal menu with 'Overview' (selected), 'Data', 'Kernels', 'Discussion', 'Leaderboard', and 'More', followed by a 'Submit Predictions' button. The main content area shows the 'Overview' section with a sidebar containing 'Description' (selected), 'Evaluation', 'Frequently Asked Questions', and 'Tutorials'. The 'Description' section is titled 'Start here if...' and contains two sub-sections: 'You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.' and 'Competition Description', which provides a detailed account of the Titanic's sinking and its historical significance.

kaggle Search kaggle Q Competitions Datasets Kernels Discussion Jobs ... Sign In

Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

k Kaggle · 7,225 teams · 3 years to go

Overview Data Kernels Discussion Leaderboard More Submit Predictions

Overview

Description

Evaluation

Frequently Asked Questions

Tutorials

Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

Competition Description

The sinking of the RMS 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.





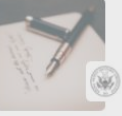


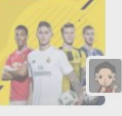



Kaggle R Tutorial on Machine Learning

<https://www.datacamp.com/community/open-courses/kaggle-tutorial-on-machine-learning-the-sinking-of-the-titanic#gs.Q8Mgsaw>

Kaggle Python Tutorial on Machine Learning

https://www.datacamp.com/community/open-courses/kaggle-python-tutorial-on-machine-learning#gs.qfvND_o

Datasets

4		Open Data 500 Companies The first comprehensive study of U.S. companies using open government data GovLab · updated 3 days ago	54 downloads 0 comments
134		Breast Cancer Wisconsin (Diagnostic) Data Set Predict whether the cancer is benign or malignant UCI Machine Learning · updated 9 months ago	8,608 downloads 29 comments
4		Amending America 11,000+ Proposed Amendments to the United States Constitution from 1787 to 2014 U.S. National Archives and Records Administration · updated 3 days ago	19 downloads 1 comment
217		Climate Change: Earth Surface Temperature Data Exploring global temperatures since 1750 Berkeley Earth · updated 3 days ago	14,642 downloads 41 comments
514		Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or not Andrea · updated 3 months ago	15,800 downloads 56 comments
48		Complete Football Dataset (Global) 15k+ players, 100+ leagues, 100+ years of data Soumitra Agarwal · updated 2 months ago	769 downloads 4 comments
110		Mushroom Classification Safe to eat or deadly poison? UCI Machine Learning · updated 7 months ago	5,130 downloads 14 comments
572		IMDB 5000 Movie Dataset 5000+ movie data scraped from IMDB website chuansun76 · updated 10 months ago	37,035 downloads 74 comments
169		Global Terrorism Database More than 150,000 terrorist attacks worldwide, 1970-2015 START Consortium · updated 7 months ago	9,742 downloads 30 comments

IMDB 5000 Movie Dataset

5000+ movie data scraped from IMDB website

chuansun76 · last updated 10 months ago

[Overview](#) [Kernels](#) [Discussion](#) [Activity](#)

[Download \(580 KB\)](#) [New Kernel](#)

Kernels	Discussion	Top Contributors
<div>Principal Component Analysis.. run a day ago 88 votes</div>	<div>Principal Component Analys.. a day ago 34 replies</div>	<div>Anisotropic 1st</div>
<div>EDA with Plotly run 10 months ago 32 votes</div>	<div>rank the directors and actor... 7 days ago 10 replies</div>	<div>AdhokshajaPradeep 2nd</div>
<div>Network Mapping Hollywood ... run 6 months ago 22 votes</div>	<div>Please update the dataset w... 14 days ago 0 replies</div>	<div>steinate 3rd</div>

Question

How can we tell the greatness of a movie before it is released in the market?

This question is asked for a long time. Some people use the number of reviews to judge the quality of a movie. Many people rely on critics to judge the quality of a movie, while others rely on their own instincts. But it is not easy to judge the quality of a movie before it is released. And human instinct sometimes is unreliable.

Method

To answer this question, I scraped 5000+ movies from IMDB website using a Python library called "scrapy".

The scraping process took 2 hours to finish. In the end, I was able to obtain all needed 28 variables for 5043 movies and 4906 posters (998MB), spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. Below are the 28 variables:

"movie_title" "color" "num_critic_for_reviews" "movie_facebook_likes" "duration" "director_name" "director_facebook_likes" "actor_3_name" "actor_3_facebook_likes" "actor_2_name" "actor_2_facebook_likes" "actor_1_name" "actor_1_facebook_likes" "gross" "genres" "num_voted_users" "cast_total_facebook_likes" "facenumber_in_poster" "plot_keywords" "movie_imdb_link" "num_user_for_reviews"

Pathways:

MS/PhD in Data
Science

Internship

Self-study

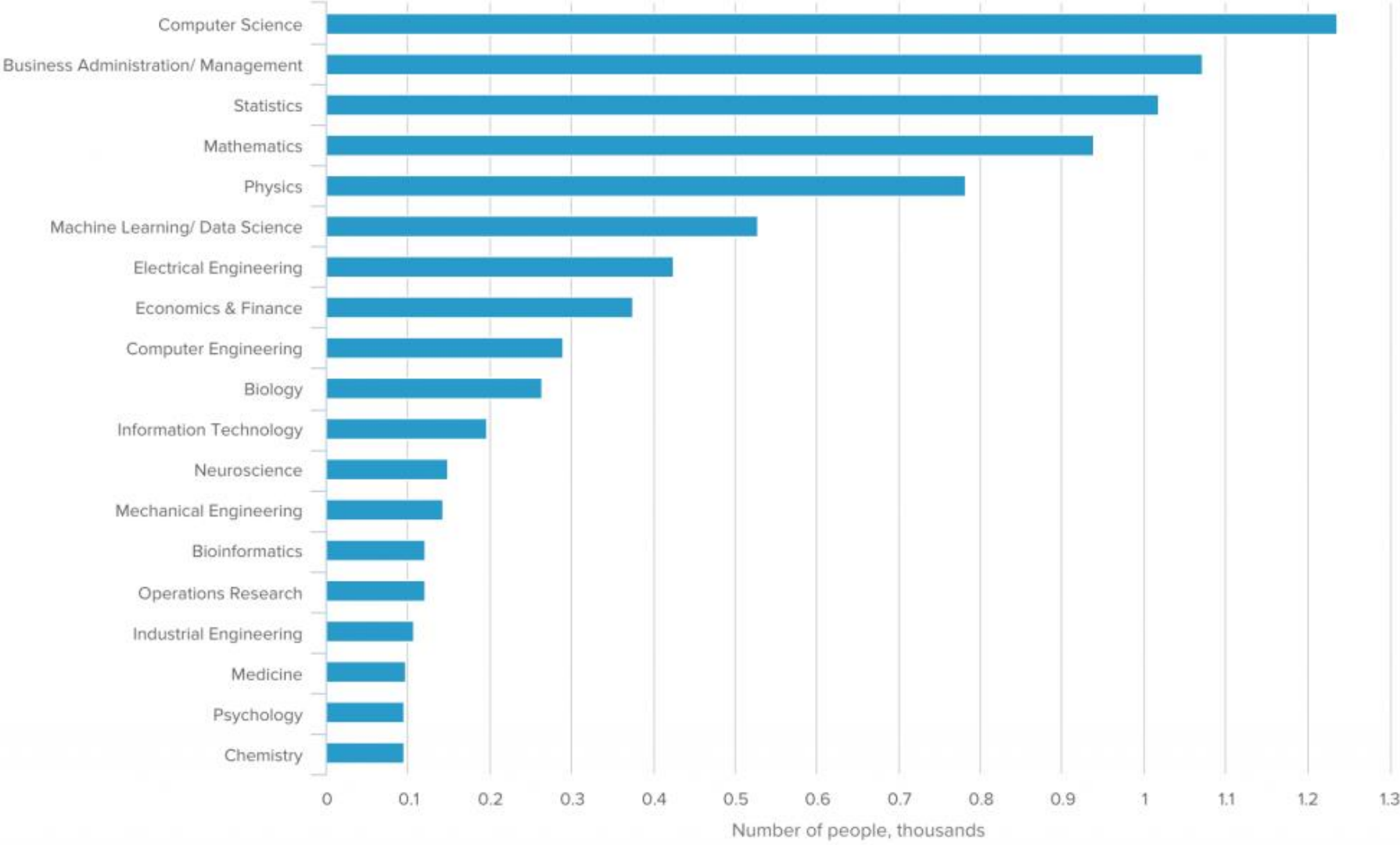
Immersive
Programs

Data Scientist

```
graph TD; A[MS/PhD in Data Science] --> D[Data Scientist]; B[Internship] --> D; C[Self-study] --> D; E[Immersive Programs] --> D;
```

The diagram illustrates four distinct pathways that lead to the role of a Data Scientist. At the top, a green box labeled 'Pathways:' introduces the section. Below it, four blue boxes are arranged horizontally, each representing a different route: 'MS/PhD in Data Science', 'Internship', 'Self-study', and 'Immersive Programs'. Green lines with arrowheads at the bottom connect each of these four boxes to a single dark blue box at the bottom labeled 'Data Scientist', indicating that all these paths converge on the same career goal.

TOP 20 BACKGROUNDS OF DATA SCIENTISTS WITH A GRADUATE DEGREE



Thank you Geek Girls Portugal!