

Machine Learning

Dr Tamara Celford

To download a copy of the slides go to
<https://tamaracleford.co.uk/iop.html>



INTRODUCTION TO DATA SCIENCE

- What is data science
- Examples from my work
- Data science workflow walkthrough
- How can I learn more?



IS DATA SCIENCE: “THE SEXIEST JOB OF THE 21ST CENTURY”?

- In 2012 the Harvard Business Review described data science as being
‘The sexiest job of the 21st Century’
- ‘Data will be the raw material of the 21st Century’
Angela Merkel, German Chancellor, Davos 2018



WHAT IS THIS THING CALLED DATA SCIENCE?

← → ↻ https://en.wikipedia.org/wiki/Data_science



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)

Not logged in

Article

[Talk](#)

Read

[Edit](#)

[View](#)

Data science

From Wikipedia, the free encyclopedia

Not to be confused with [information science](#).

Data science is a [multi-disciplinary](#) field that uses scientific methods, processes, algorithms and systems to extract [knowledge](#) and insights from [data](#) in various forms, both structured and unstructured,^{[1][2]} similar to [data mining](#).

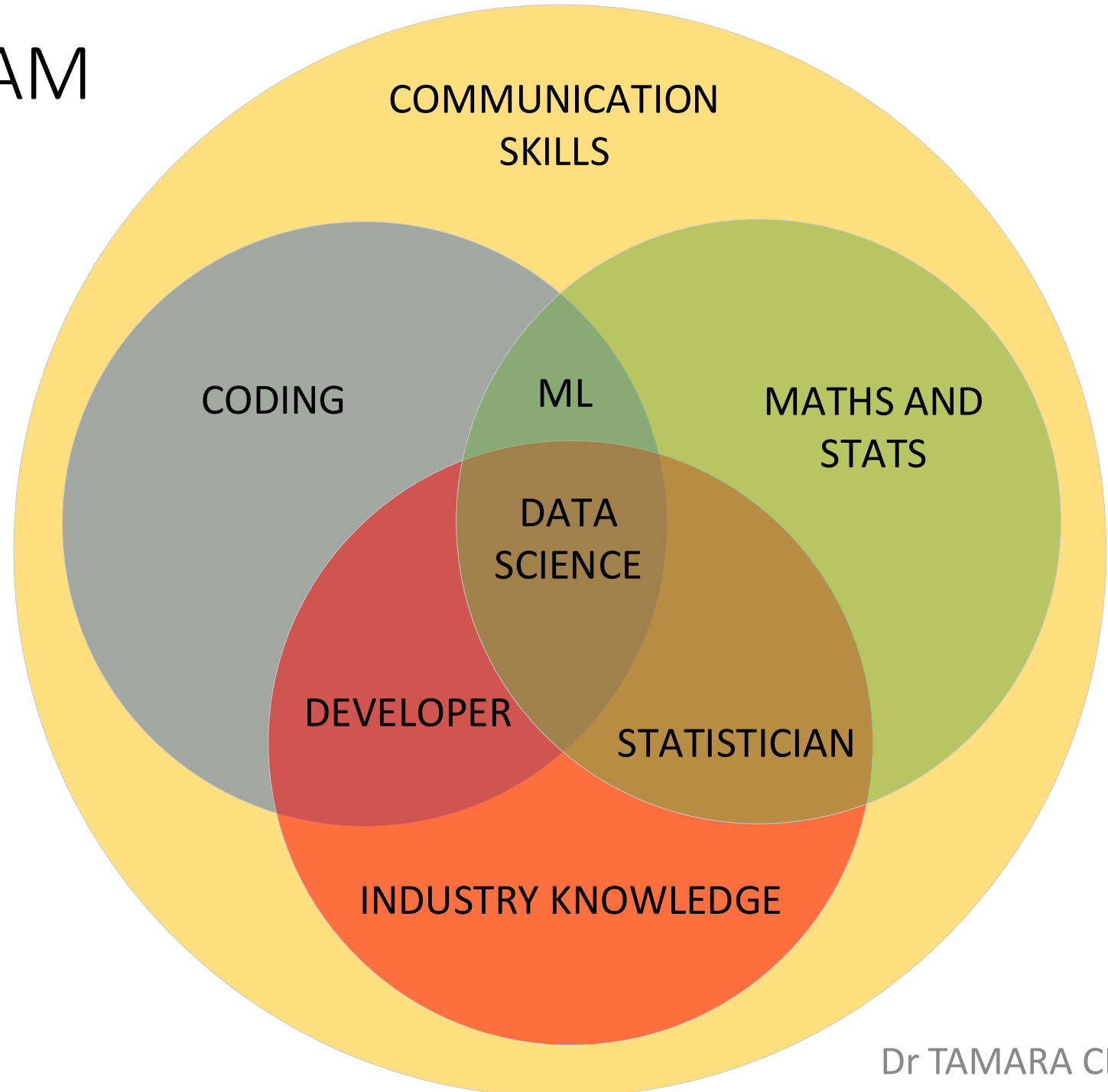


WHAT IS THIS THING CALLED DATA SCIENCE?

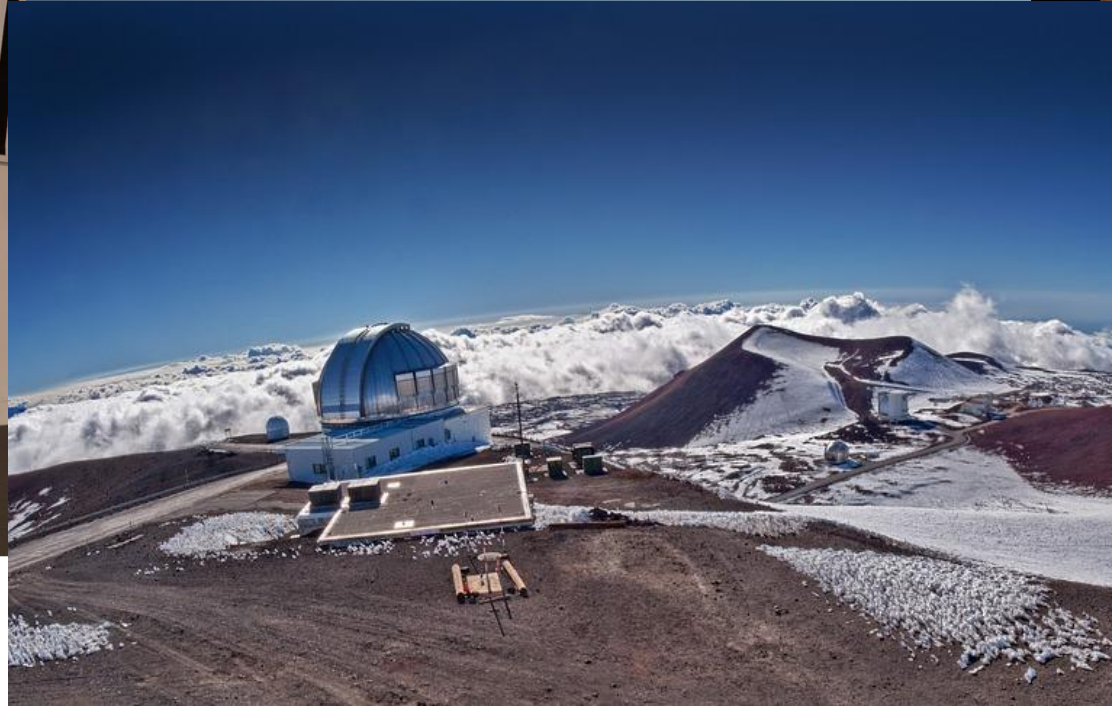
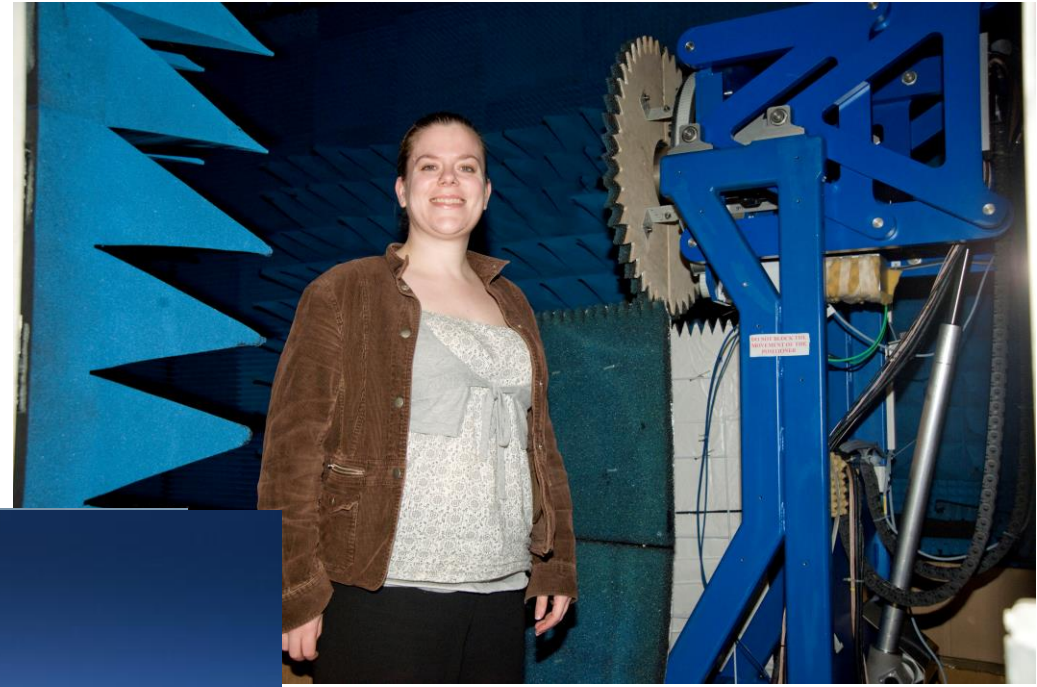
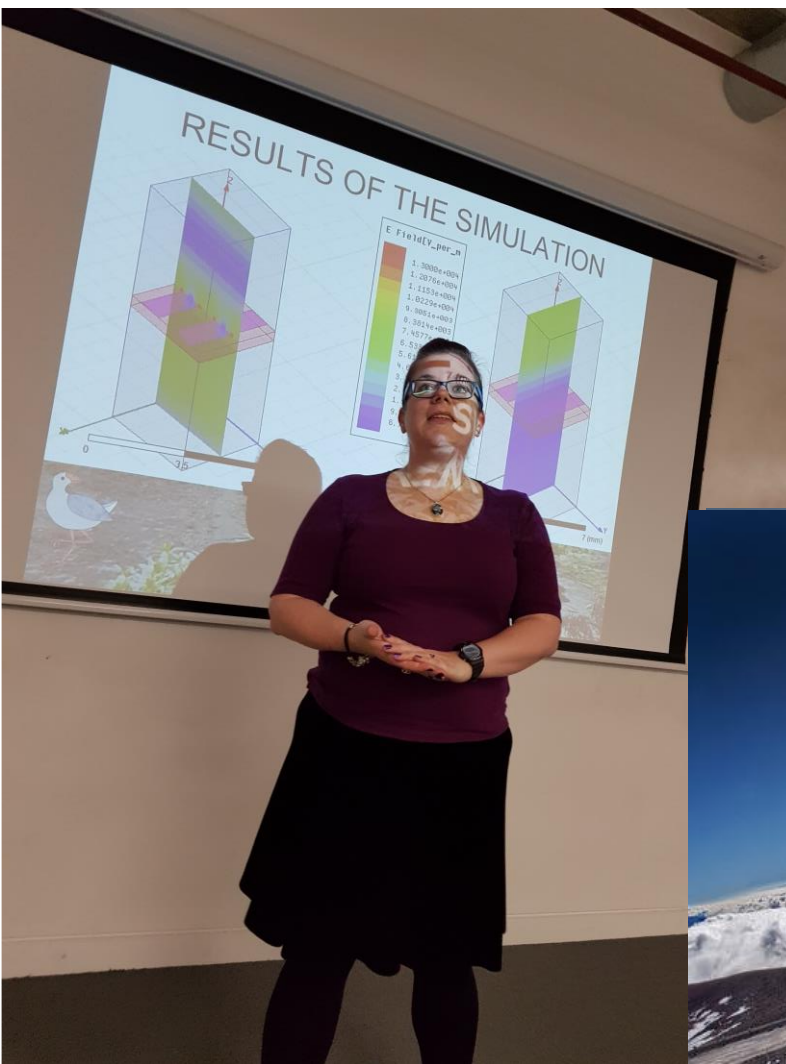
- Any data can be looked at, multiple data sets used
 - Structured, unstructured, numbers, words, clicks, actions, images
- Multi-disciplinary
 - Part physics, statistics, maths, computer science, coding, big data, business
- Clean data before use
 - Removes incorrect and partial data
- Uses scientific methods and algorithms
 - Coding, decision trees, machine learning, neural networks
- Extracts knowledge and insights from data
 - The output is verified as being plausible



THE VENN DIAGRAM OF DATA SCIENCE



ABOUT ME



SWAMPHEN ENTERPRISES

<https://www.needpix.com/photo/920796/mountain-telescope-hawaii-summit-astronomy-astrophysics-mauna-kea-keck-telescope-subaru-telescope>

Dr TAMARA CLELFORD

HOUSE OCCUPANCY CALCULATIONS



COMMERCIAL EXAMPLES OF DATA SCIENCE

- Business analytics
- Recommender engines
- Online advertisements
- Financial predictions
- Internet of Things
- Diagnostics
- Kaggle competitions



THE RISE OF DATA SCIENCE

- As more data is collected in all walks of life the need to analyse this data and find it's hidden secrets becomes increasingly important
- Blurred line between engineer, statistician, data analyst, data scientist, data engineer, business analyst
- People from different backgrounds are re-training and becoming data scientists



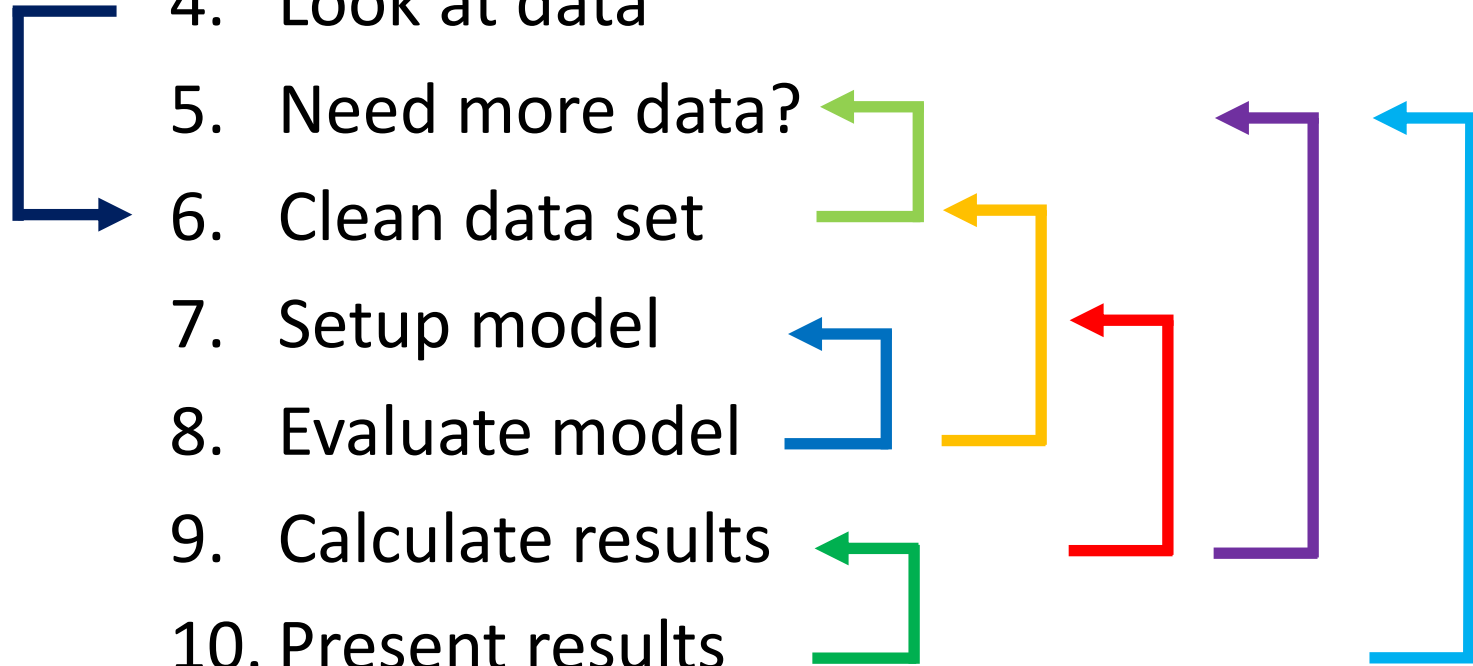
DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results

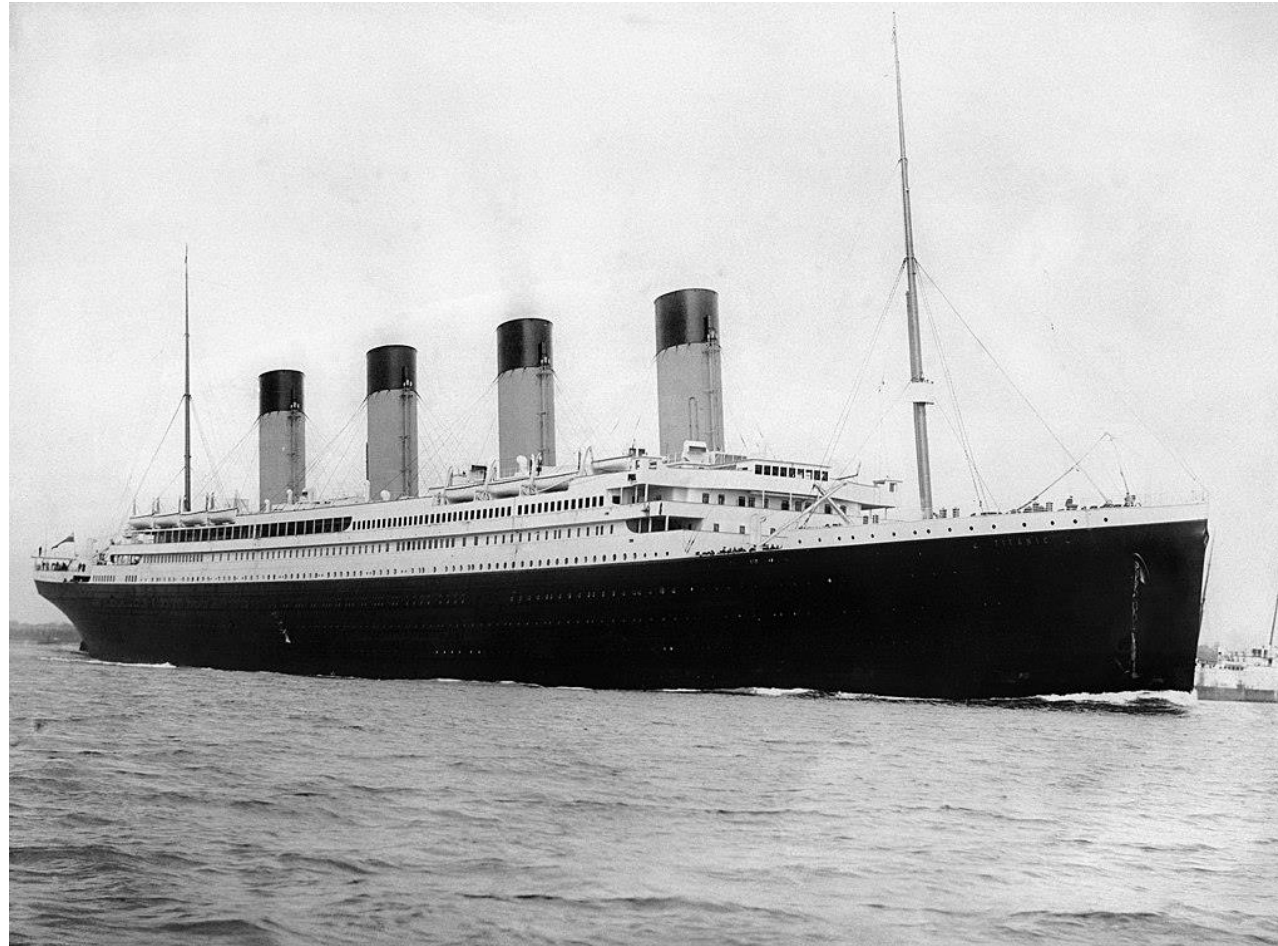


DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



SURVIVAL OF THE BEST DATA



By F.G.O. Stuart (1843-1923) - <http://www.uwants.com/viewthread.php?tid=3817223&extra=page%3D1>, Public Domain, <https://commons.wikimedia.org/w/index.php?curid=2990792>

DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



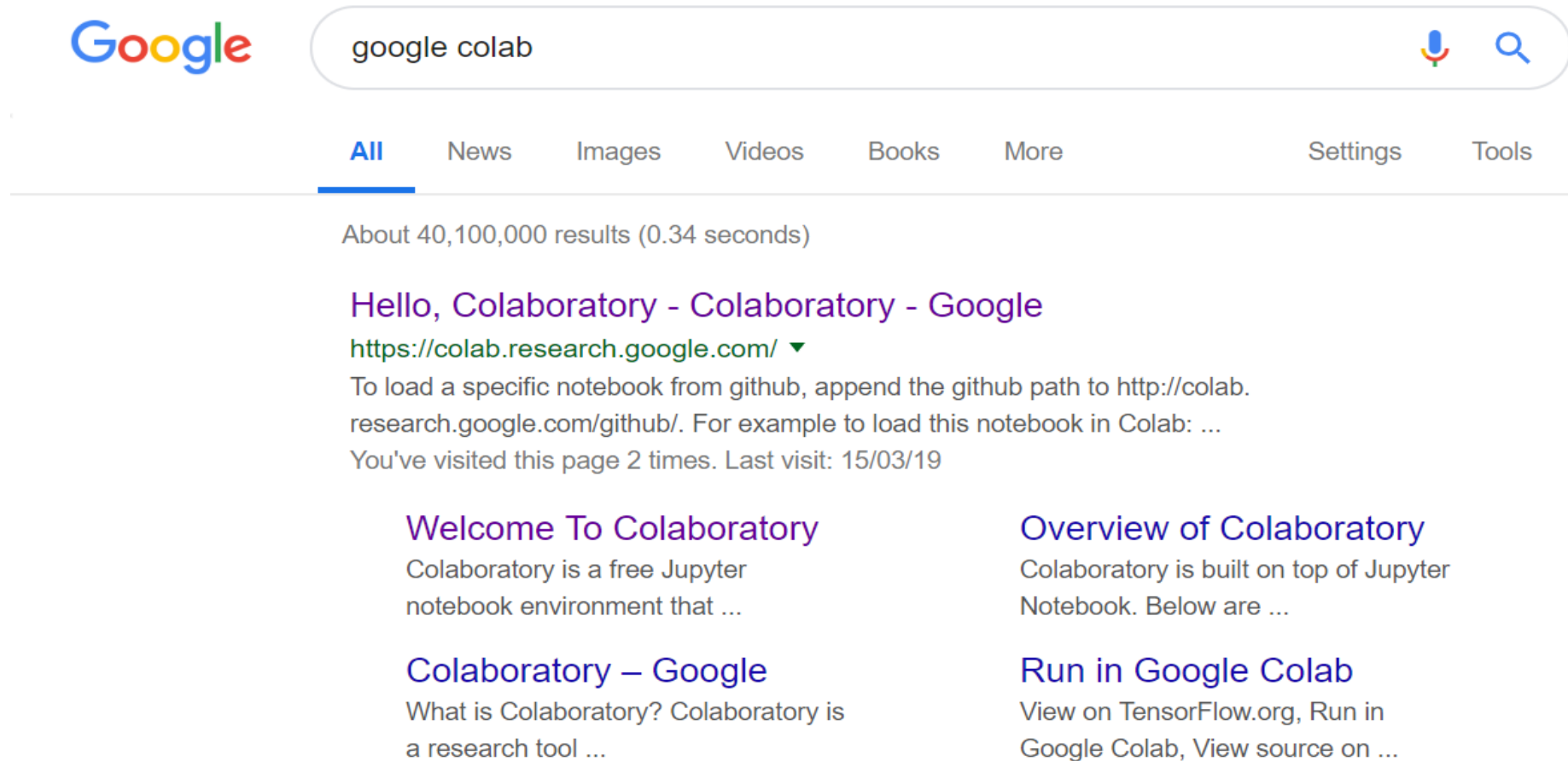
2) SET UP DATA SCIENCE TOOLS

- Python programming language
 - Toolboxes such as NumPy, pandas and scikit-learn
- Integrated Development Environment (IDE)
 - Jupyter notebook style
 - Google Colab



GO TO GOOGLE COLAB

<https://colab.research.google.com>



The screenshot shows a Google search interface. The Google logo is on the left. The search bar contains the text "google colab". To the right of the search bar are a microphone icon and a magnifying glass icon. Below the search bar are tabs for "All", "News", "Images", "Videos", "Books", "More", "Settings", and "Tools". The "All" tab is selected and underlined. Below the tabs, it says "About 40,100,000 results (0.34 seconds)". The first search result is titled "Hello, Colaboratory - Colaboratory - Google" in purple. Below the title is the URL "https://colab.research.google.com/" in green, followed by a small downward arrow. The description of the result says: "To load a specific notebook from github, append the github path to http://colab.research.google.com/github/. For example to load this notebook in Colab: ... You've visited this page 2 times. Last visit: 15/03/19". Below the search results are four cards. The first card is titled "Welcome To Colaboratory" in purple, with the text "Colaboratory is a free Jupyter notebook environment that ...". The second card is titled "Overview of Colaboratory" in purple, with the text "Colaboratory is built on top of Jupyter Notebook. Below are ...". The third card is titled "Colaboratory – Google" in purple, with the text "What is Colaboratory? Colaboratory is a research tool ...". The fourth card is titled "Run in Google Colab" in purple, with the text "View on TensorFlow.org, Run in Google Colab, View source on ...".

Google

google colab

All News Images Videos Books More Settings Tools

About 40,100,000 results (0.34 seconds)

Hello, Colaboratory - Colaboratory - Google

<https://colab.research.google.com/> ▼

To load a specific notebook from github, append the github path to <http://colab.research.google.com/github/>. For example to load this notebook in Colab: ...

You've visited this page 2 times. Last visit: 15/03/19

Welcome To Colaboratory

Colaboratory is a free Jupyter notebook environment that ...

Overview of Colaboratory

Colaboratory is built on top of Jupyter Notebook. Below are ...

Colaboratory – Google

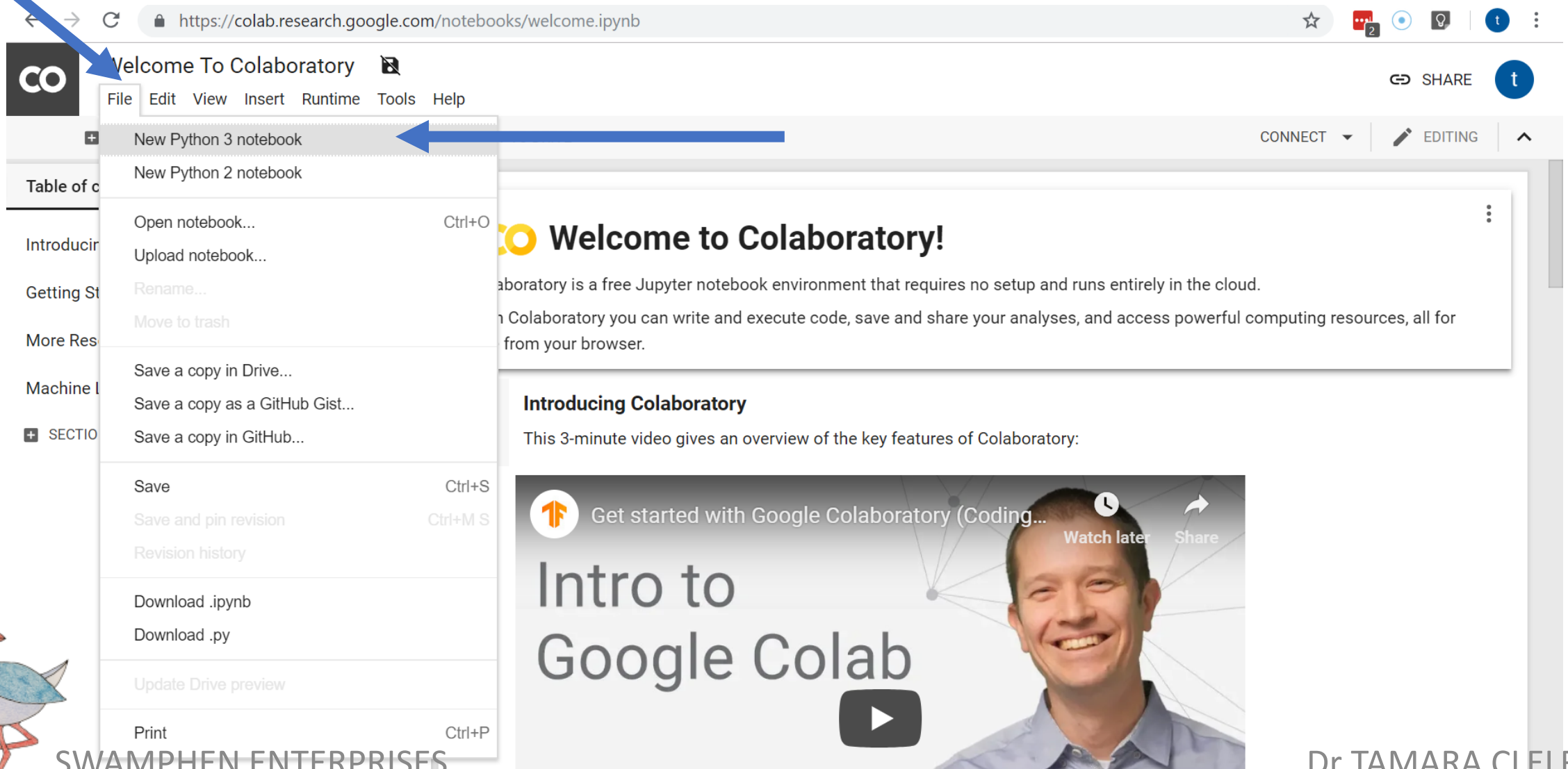
What is Colaboratory? Colaboratory is a research tool ...

Run in Google Colab

View on TensorFlow.org, Run in Google Colab, View source on ...



OPEN A NEW PYTHON 3 NOTEBOOK



The screenshot shows the Google Colaboratory web interface. The browser address bar displays <https://colab.research.google.com/notebooks/welcome.ipynb>. The 'File' menu is open, showing options like 'New Python 3 notebook', 'New Python 2 notebook', 'Open notebook...', 'Upload notebook...', 'Rename...', 'Move to trash', 'Save a copy in Drive...', 'Save a copy as a GitHub Gist...', 'Save a copy in GitHub...', 'Save', 'Save and pin revision', 'Revision history', 'Download .ipynb', 'Download .py', 'Update Drive preview', and 'Print'. The 'New Python 3 notebook' option is highlighted. The main content area displays a 'Welcome to Colaboratory!' message and a video titled 'Intro to Google Colab'.

Table of contents:

- Table of contents
- Introducing Colaboratory
- Getting Started
- More Resources
- Machine Learning
- SECTION

CONNECT | EDITING

Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. In Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all from your browser.

Introducing Colaboratory

This 3-minute video gives an overview of the key features of Colaboratory:

Get started with Google Colaboratory (Coding...)

Watch later | Share

Intro to Google Colab



2) SETUP DATA SCIENCE TOOLS

```
1  # open up google colab
2  # open a python 3 notebook and rename to 'Titanic_analysis'
3  # import packages going to use
4  import numpy as np
5  import pandas as pd
6  import matplotlib.pyplot as plt
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



3) GET DATA

```
1 # import the data to google colab  
2 from google.colab import files  
3 uploaded = files.upload()
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



4) LOOK AT YOUR DATA SET

- Important step to get an understanding of the data
 - If you don't understand your data set it will be hard to get insights from it
- Check that is has been read in correctly
- Look at raw data and plot things to help understanding



TITANIC DATA INVESTIGATION

```
1 # import data set
2 raw_data = pd.read_csv('titanic.csv')
```

```
1 # print out the header
2 raw_data.head()
3 # try putting a number in the brackets
4 # what does .tail() do?
```



```
1 # how many rows in the data set
2 len(raw_data)
```

```
1 # get basic shape information
2 raw_data.shape
```

```
1 # print out a selection of the data
2 raw_data[20:25][['Name']]
```

```
1 # sort wrt one column
2 raw_data.sort_values('Fare')
```

```
1 # calculate average fare
2 np.mean(raw_data['Fare'])
```

```
1 # describe data set - nothing with letters in the result
2 raw_data.describe()
```

```
1 # get the range of ages on board
2 age_female = raw_data.loc[raw_data['Sex'] == 'female', 'Age']
3 age_male = raw_data.loc[raw_data['Sex'] == 'male', 'Age']
4 print(age_female)
5 print(age_male)
```

```
1 # Look at the range of ages on board, often scatter graph, but not in this case
2 fig, graph = plt.subplots()
3 graph.hist([age_female, age_male])
```

```
1 # graph options
2 fig, graph = plt.subplots()
3 graph.hist([age_female, age_male], color = ['b', 'g'], label = ['female', 'male'])
4 # pick any colours to use from b, g, k, r, c, m, y, w
5 # print out the axis labels
6 graph.set_ylabel("number of people")
7 graph.set_xlabel("age (y)")
8 # add the legend
9 plt.legend(loc = 'best')
10 # move the legend around, options are:
11 # upper left, upper right, lower left, lower right, best, upper center, lower center,
12 # center left, center right
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



5) NEED MORE DATA?

```
1 # we want to be able to include gender information
2 # change female = 1, male = 0
3 raw_data['GenderNumerical'] = raw_data.apply(lambda row:1 if row.Sex == 'female' else 0, axis = 1)
4 raw_data.head()
```

```
1 # describe the data set again
2 raw_data.describe()
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



6) CLEAN DATA SET

- This is where most of your time is spent as a data scientist
- Need to ensure you do not introduce bias or errors to the data set during cleaning
- Examples of cleaning practices:
 - Remove lines with missing data
 - Use an average to fill missing data
 - Use a different data set to fill missing data
 - Interpolate between points



6) CLEAN DATA SET

```
1 raw_data.head(7)
```

```
1 # pick data want to include  
2 include = ['Survived', 'Pclass', 'Age', 'GenderNumerical', 'Fare']
```

```
1 # check include  
2 raw_data[include]
```

```
1 # remove nans  
2 data_set = raw_data[include].dropna()
```

```
1 # check removal of nans  
2 data_set.head(7)  
3 data_set[-5:]
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results

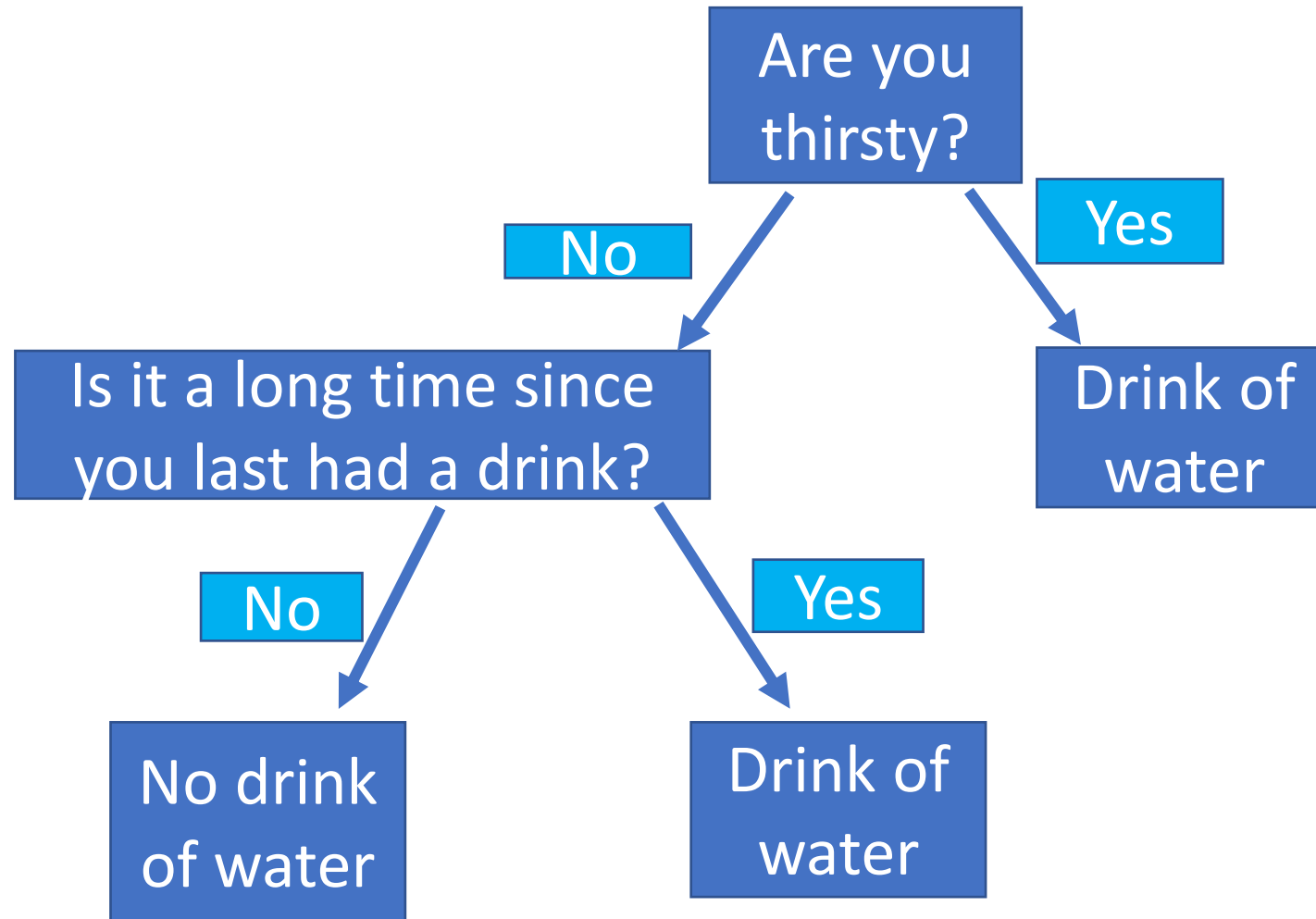


MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

- AI is getting computers to behave like humans
- ML is an application of AI
 - gives the ML code access to data and lets it learn for itself
- Two categories of ML:
 - Supervised
 - decision trees, random forests, linear regression, Gaussian regression, Bayesian statistics
 - Unsupervised
 - clustering, k-means, anomaly detection, neural networks, method of moments



DECISION TREE



7) SET UP MODEL

```
1 # select data for random forest
2 data_chosen = include[1:]
3 data = data_set[data_chosen]
4 answer = data_set['Survived']
```

```
1 # import decision tree
2 from sklearn import tree
```

```
1 # set up the model
2 tree_model = tree.DecisionTreeClassifier(random_state = 42)
```

```
1 # run the model on our data
2 titanic_tree = tree_model.fit(data, answer)
3 print(titanic_tree)
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results

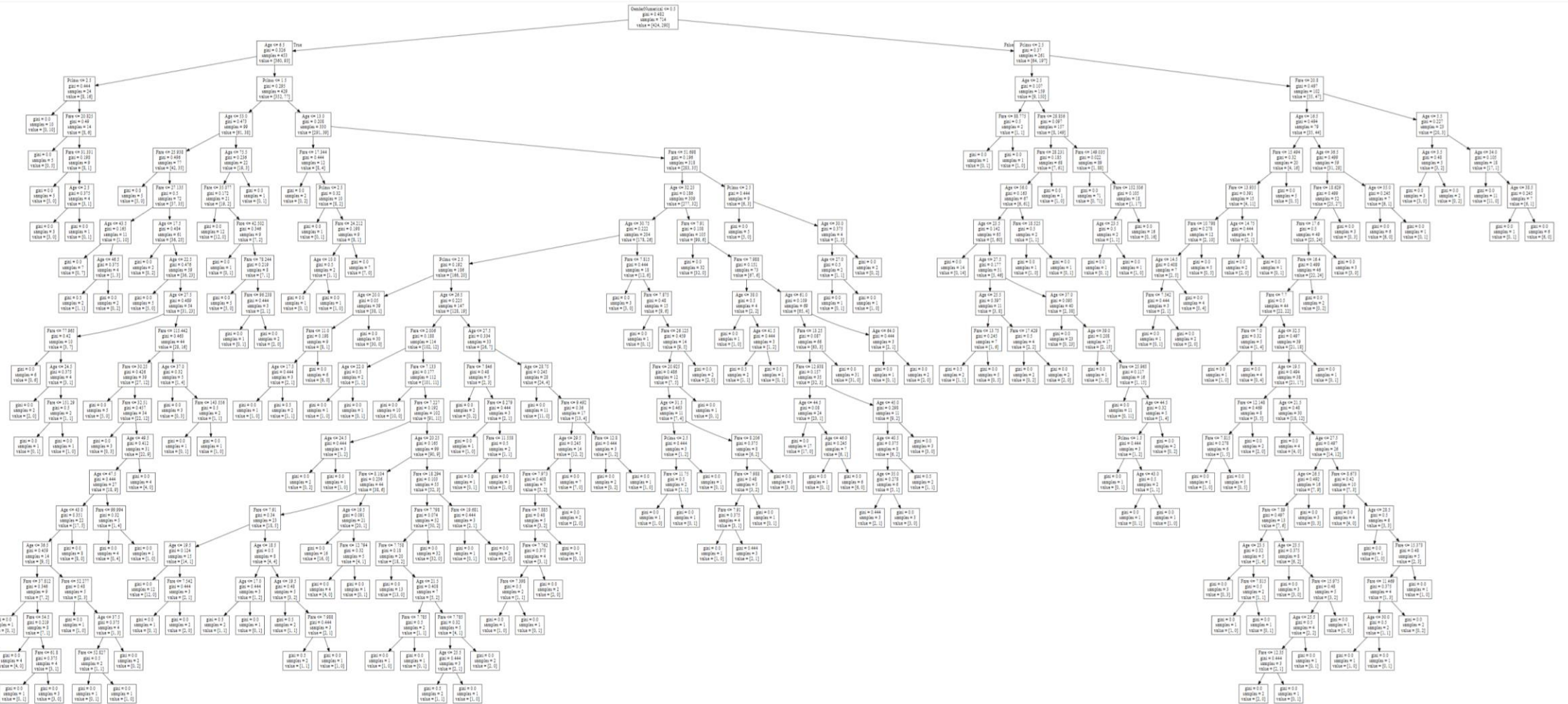


8) EVALUATE MODEL

```
1 # see how good model is at predicting survival
2 from sklearn.model_selection import cross_val_score as CVS
3 scores_tree = CVS(titanic_tree, data, answer)
4 print(scores_tree)
5 print(np.mean(scores_tree))
```



DECISION TREE

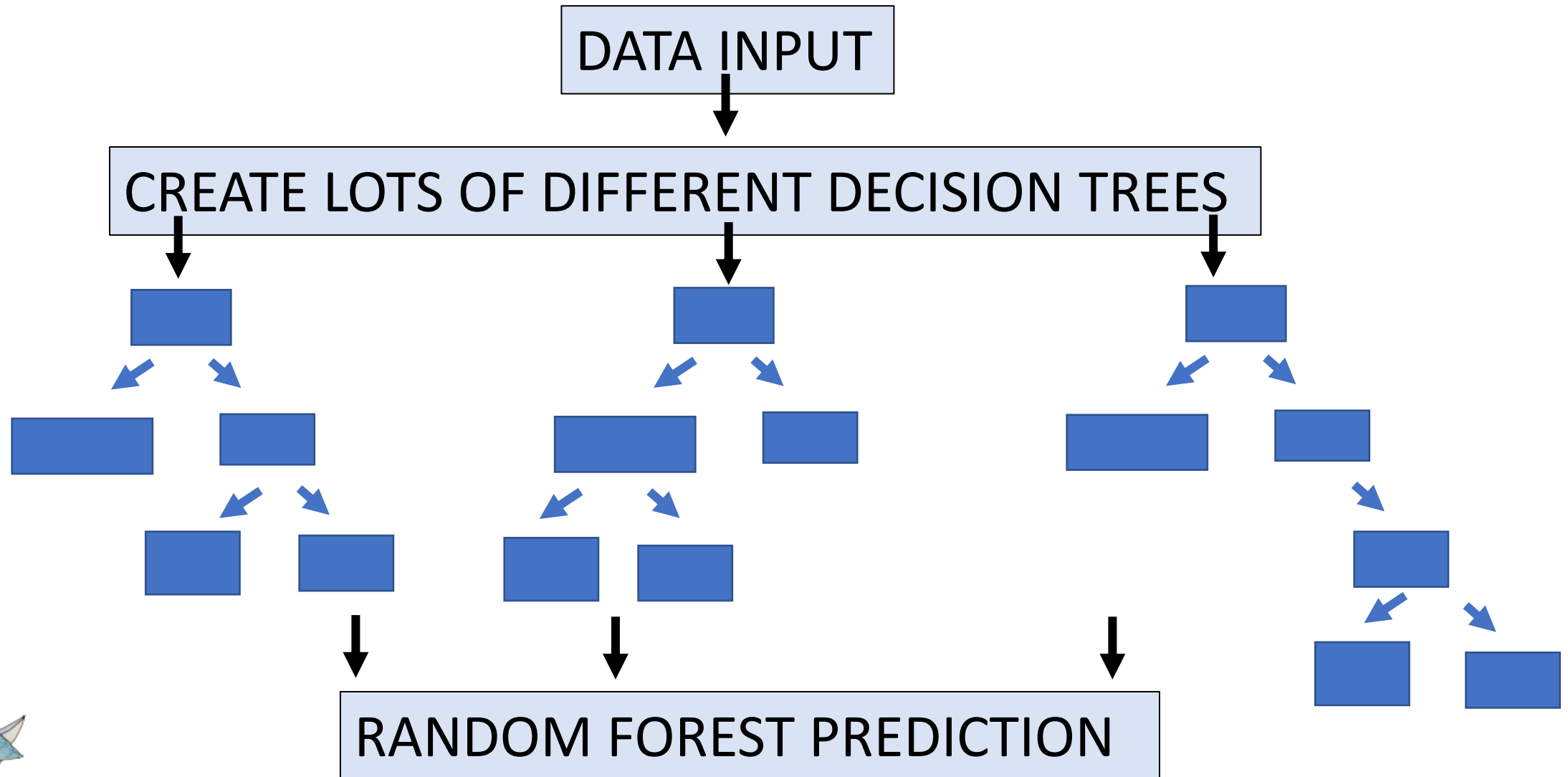


DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



RANDOM FORESTS



7) SET UP MODEL

```
1 # select data for random forest
2 include
3 include[1:]
4 data = data_set[include[1:]]
5 answer = data_set['Survived']
```

```
1 # import random forest
2 from sklearn.ensemble import RandomForestClassifier as RFC
```

```
1 # create the model
2 forest_model = RFC(n_estimators=100)
```

```
1 # run the model on the data
2 titanic_forest = forest_model.fit(data, answer)
3 print(titanic_forest)
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



CROSS VALIDATION SCORE

- Splits data randomly into k sections (fold)
- Model is fitted on $k-1$ of sections – training data
- Model is evaluated on remaining section – testing data
- Repeated k times



CROSS VALIDATION SCORE

DATA SET SPLIT			
FOLD 1	1	2	3
FOLD 2	1	2	3
FOLD 3	1	2	3

TRAINING DATA

TESTING DATA



8) EVALUATE

```
1 # see how good model is at predicting survival
2 from sklearn.model_selection import cross_val_score as CVS
3 scores_forest = CVS(titanic_forest, data, answer)
4 print(scores_forest)
5 print(np.mean(scores_forest))
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



9) CALCULATE RESULTS

```
1  # choose a person to see if they survived
2  pclass = np.mean(data_set['Pclass'])           # 2.5
3  age = np.mean(data_set['Age'])                 # 29.7 years
4  gender = np.mean(data_set['GenderNumerical']) # 0.37
5  fare = np.mean(data_set['Fare'])               # £34.69
```

```
1  # put person data into an array
2  person = np.array([[pclass, age, gender, fare]])
3  print(person)
```

```
1  # calculate survival score
2  survival_score = titanic_forest.predict_proba(person)
3  print(survival_score)
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



10) PRESENT RESULTS

```
: 1 # did the person survive the sinking of the titanic?  
 2 print({'survival chances': survival_score[0,1]*100, 'death chances': survival_score[0,0]*100})
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



9) CALCULATE RESULTS

```
1 # how does age and gender alter survival chances
2 # create a list covering all ages on the titanic
3 age_max = int(np.max(data_set['Age']))
4 print(age_max)
5 age_list = list(range(0, age_max + 1))
6 print(age_list)
```

```
1 # to look at survival probability
2 survival = []
3 selected_class = 1
4 average_fare = np.mean(data_set['Fare'])
5 gender = 1
6 for i in age_list:
7     person = np.array([[selected_class, i, gender, average_fare]])
8     survival_chances = titanic_forest.predict_proba(person)
9     survival.append(survival_chances[0,1]*100)
```

```
1 # to look at male and female survival probability
2 survival_female = []
3 survival_male = []
4 selected_class = 2
5 average_fare = np.mean(data_set['Fare'])
6 gender_list = [0,1]
7 for i in age_list:
8     for j in gender_list:
9         if j == 1:
10             array = survival_female
11         else:
12             array = survival_male
13         person = np.array([[selected_class, i, j, average_fare]])
14         survival_chances = titanic_forest.predict_proba(person)
15         array.append(survival_chances[0,1]*100)
16 len(survival_female)
```



DATA SCIENCE FLOW

1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results



10) PRESENT RESULTS

```
1 # plot just survival
2 plt.plot(age_list, survival)
3 plt.title('survival in ' + str(selected_class) + ' class')
4 plt.xlabel('age (y)')
5 plt.ylabel('survival chances (%)')
```

```
1 # plot male and female survival chances
2 plt.plot(age_list, survival_female, label = 'female')
3 plt.plot(age_list, survival_male, label = 'male')
4 plt.title('survival in ' + str(selected_class) + ' class')
5 plt.xlabel('age (y)')
6 plt.ylabel('survival chances (%)')
7 plt.legend()
```



SO, IS DATA SCIENCE: “THE SEXIEST JOB OF THE 21ST CENTURY”?

- It is certainly a job with its profile on the rise
- Analytical role with varied and interesting work
- Data Science is a well paid and sought-after skill set
 - a lot of people are re-training or re-branding
- It covers a skill set desperately needed in this new digital world
- It needs people with good analytical skills to champion its evolution as a numerical science



HOW DO I LEARN MORE?

- Learn to program in Python
- Get to grips with the data science packages
 - NumPy, pandas and scikit-learn
- Brush up on your stats
- Understand the business you want to work in
- Practice!!



QUESTIONS!

- Dr Tamara Celford 

https://tamaracleford.co.uk/online_courses.html

