Machine Learning

Dr Tamara Clelford

To download a copy of the slides go to
https://tamaraclelford.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

Not l

WIKIPEDIA
The Free Encyclopedia

Article | Talk | Read | Edit | View

Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

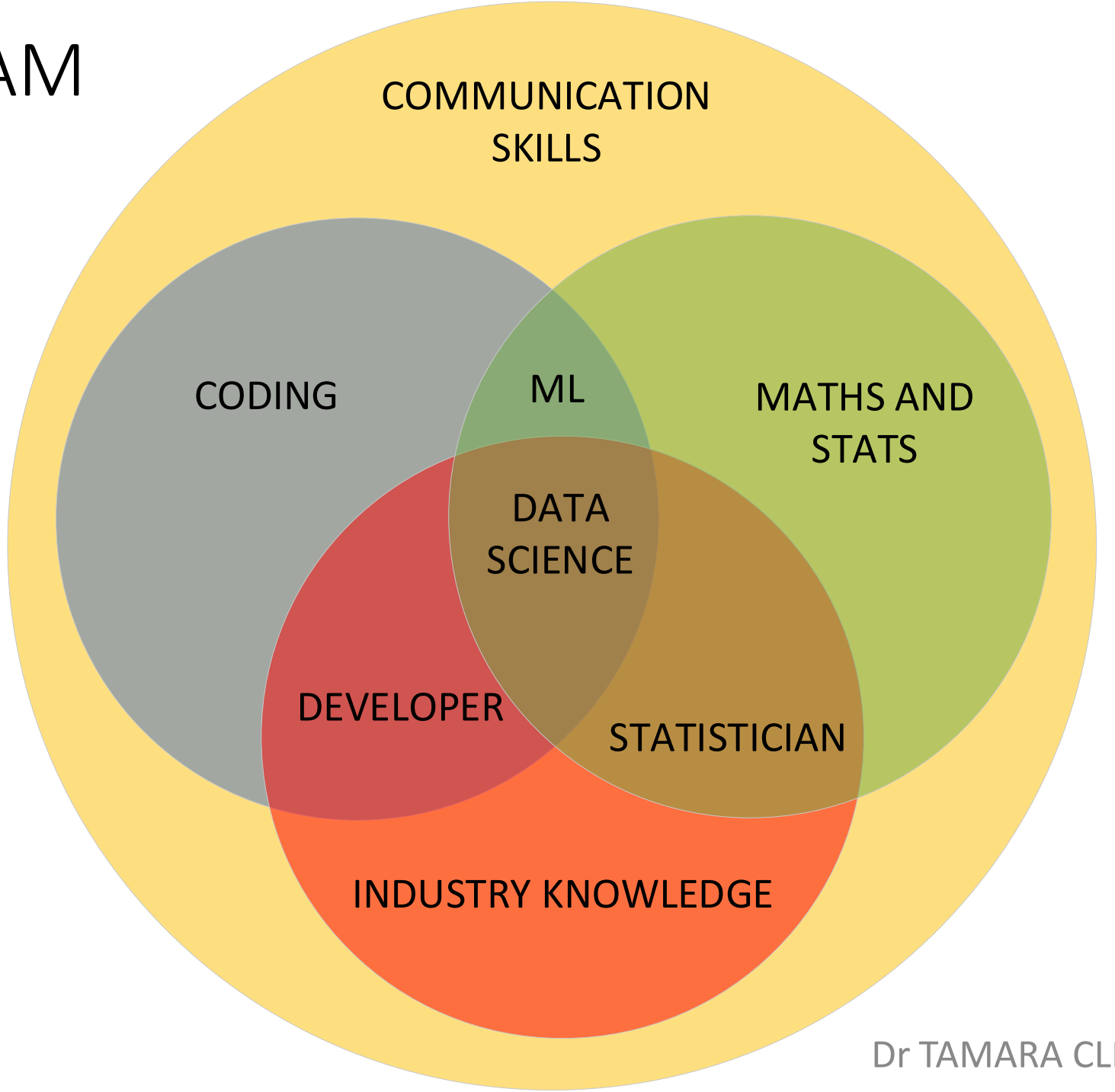
Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining.

Main page
Contents
Featured content
Current events
Random article

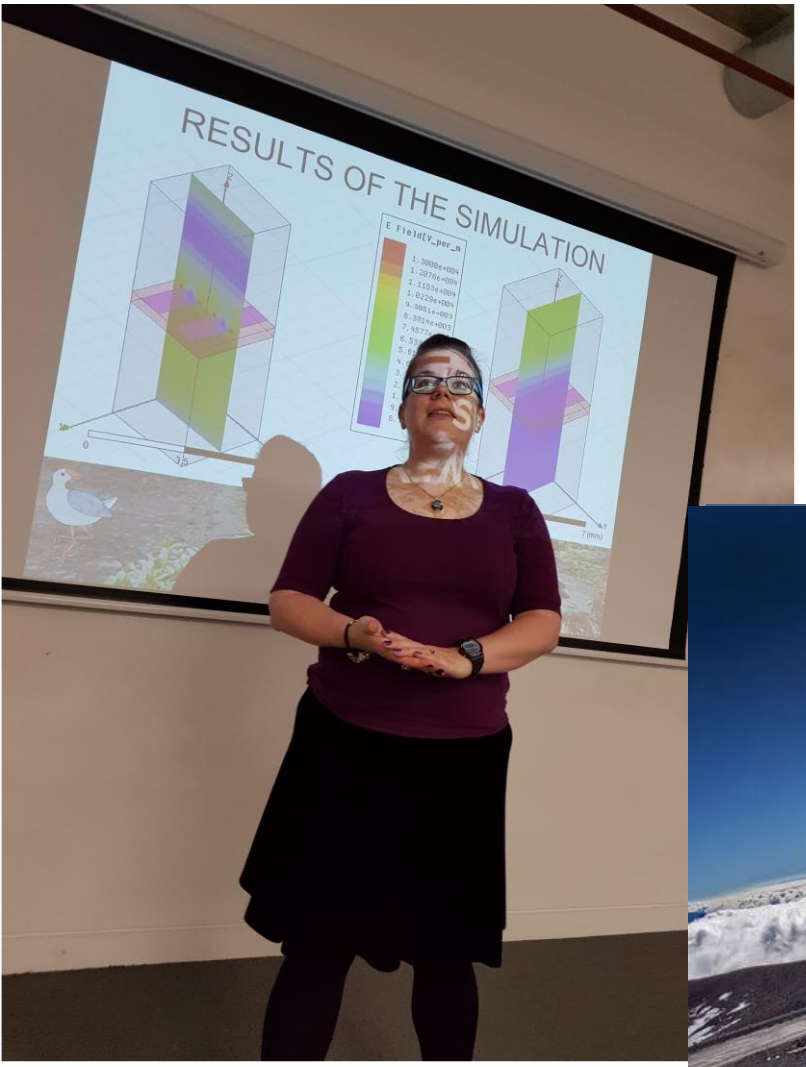
WHAT IS THIS THING CALLED DATA SCIENCE?

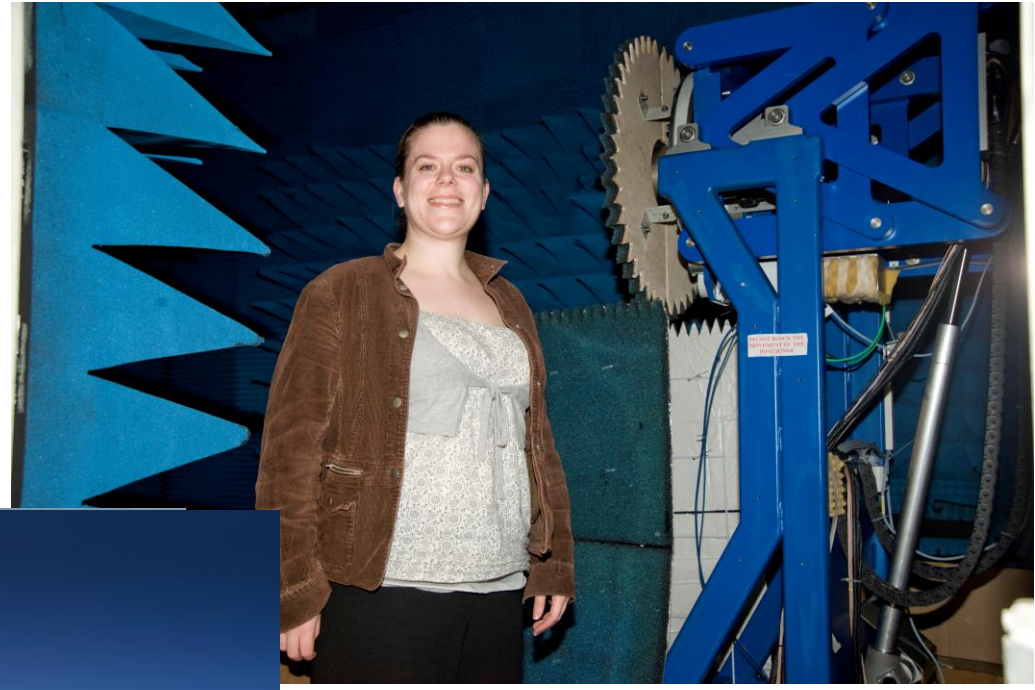
- Any data can be looked at, multiple data sets used
  - Structured, unstructured, numbers, words, clicks, actions, images
- Multi-disciplinary
  - Part physics, statistics, maths, computer science, coding, big data, business
- Clean data before use
  - Removes incorrect and partial data
- Uses scientific 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



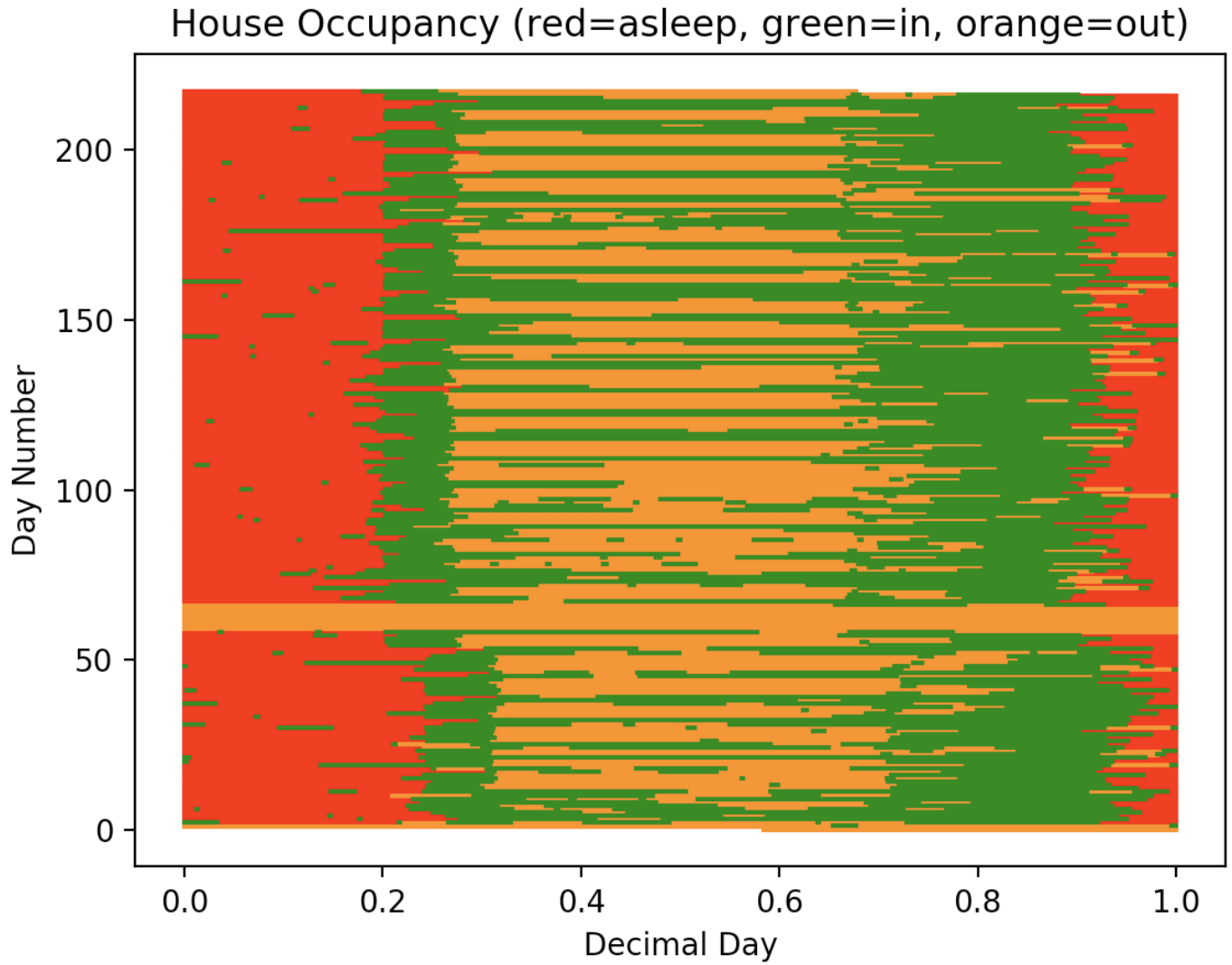
ABOUT ME





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

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

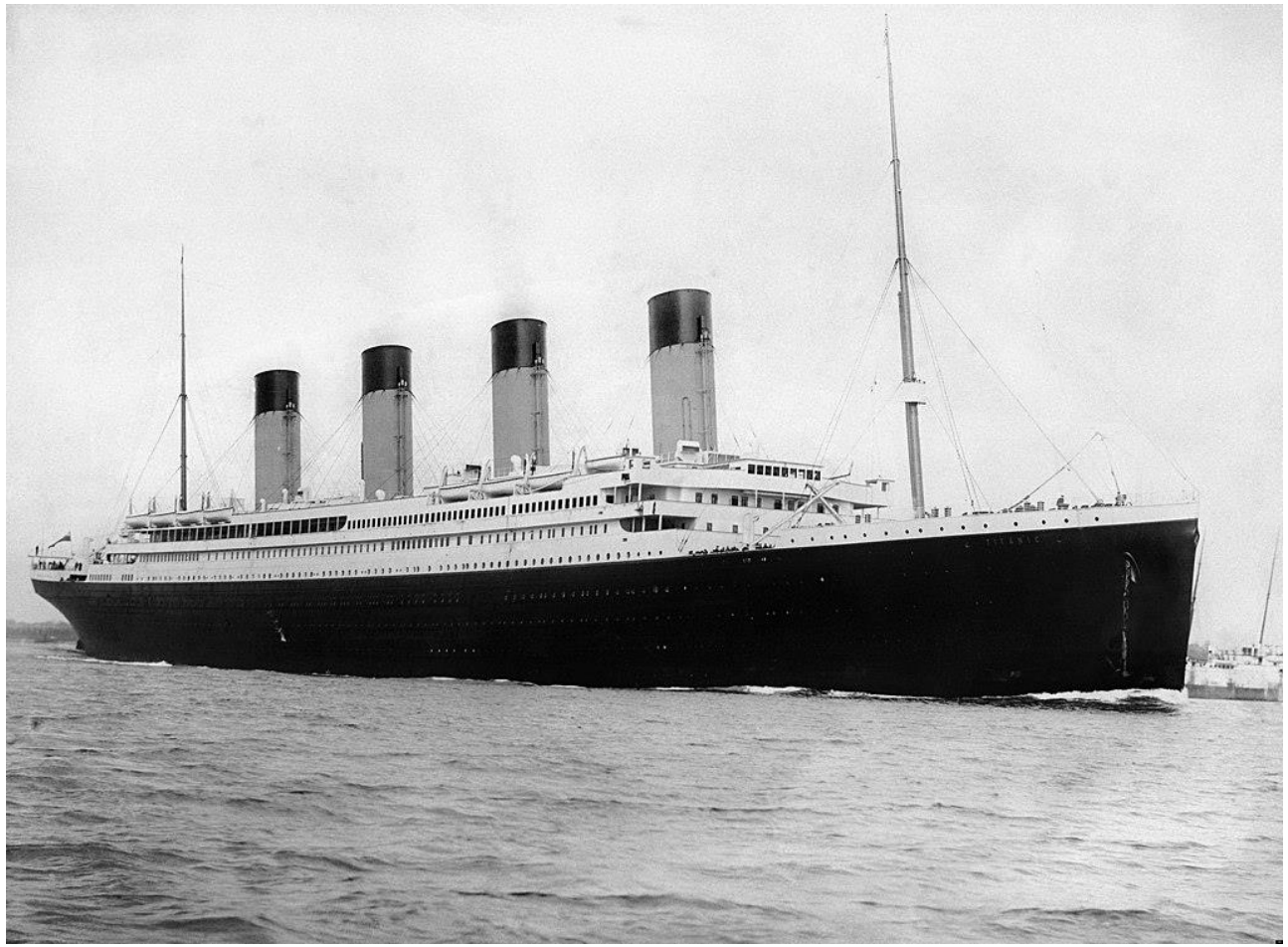
DATA SCIENCE FLOW

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

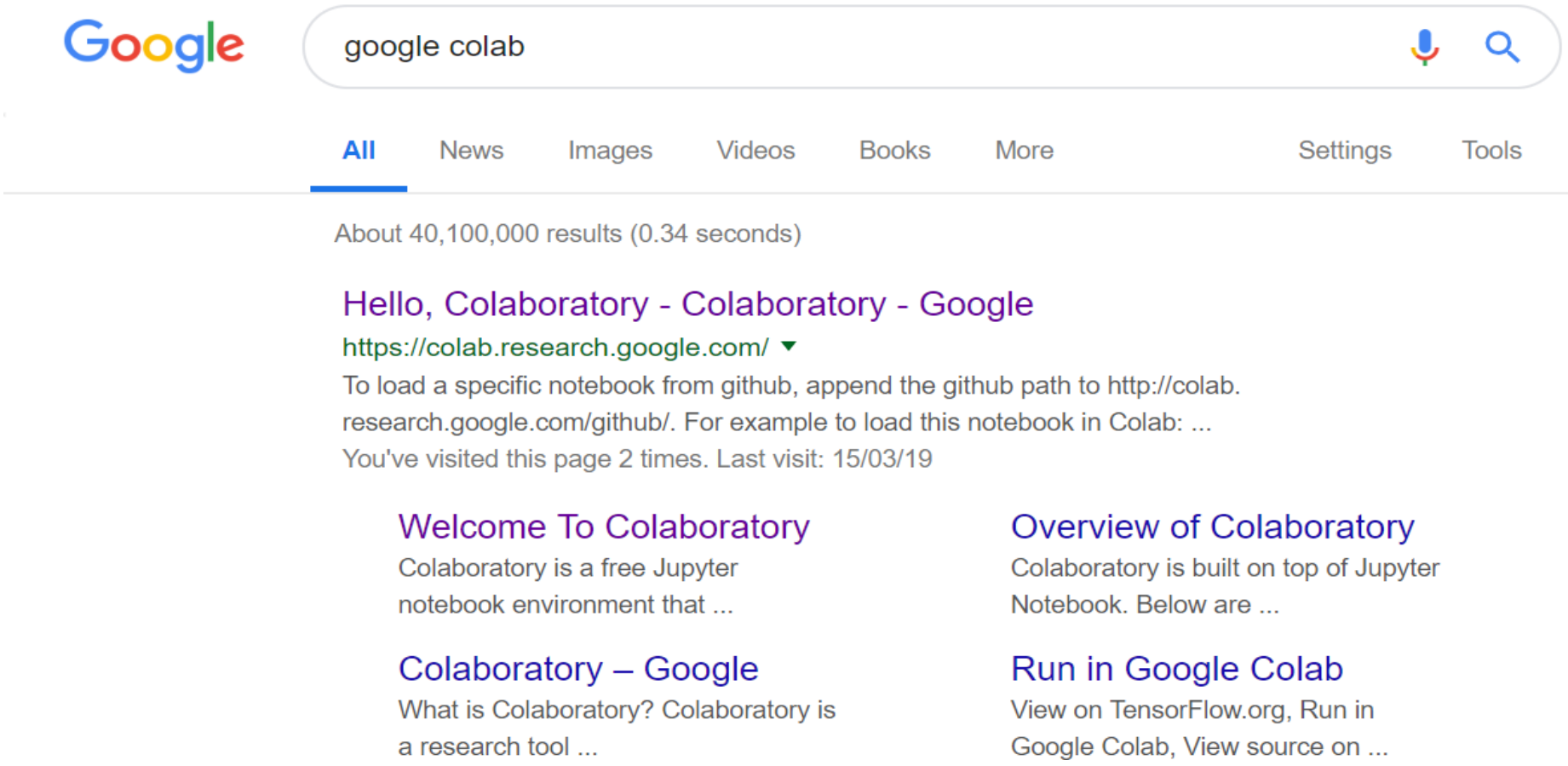
1. Identify requirement
2. Setup data science tools
3. Get data
4. Look at data
5. Need more data?
6. Clean data set
7. Setup model
8. Evaluate model
9. Calculate results
10. Present results

2) SET UP DATA SCIENCE TOOLS

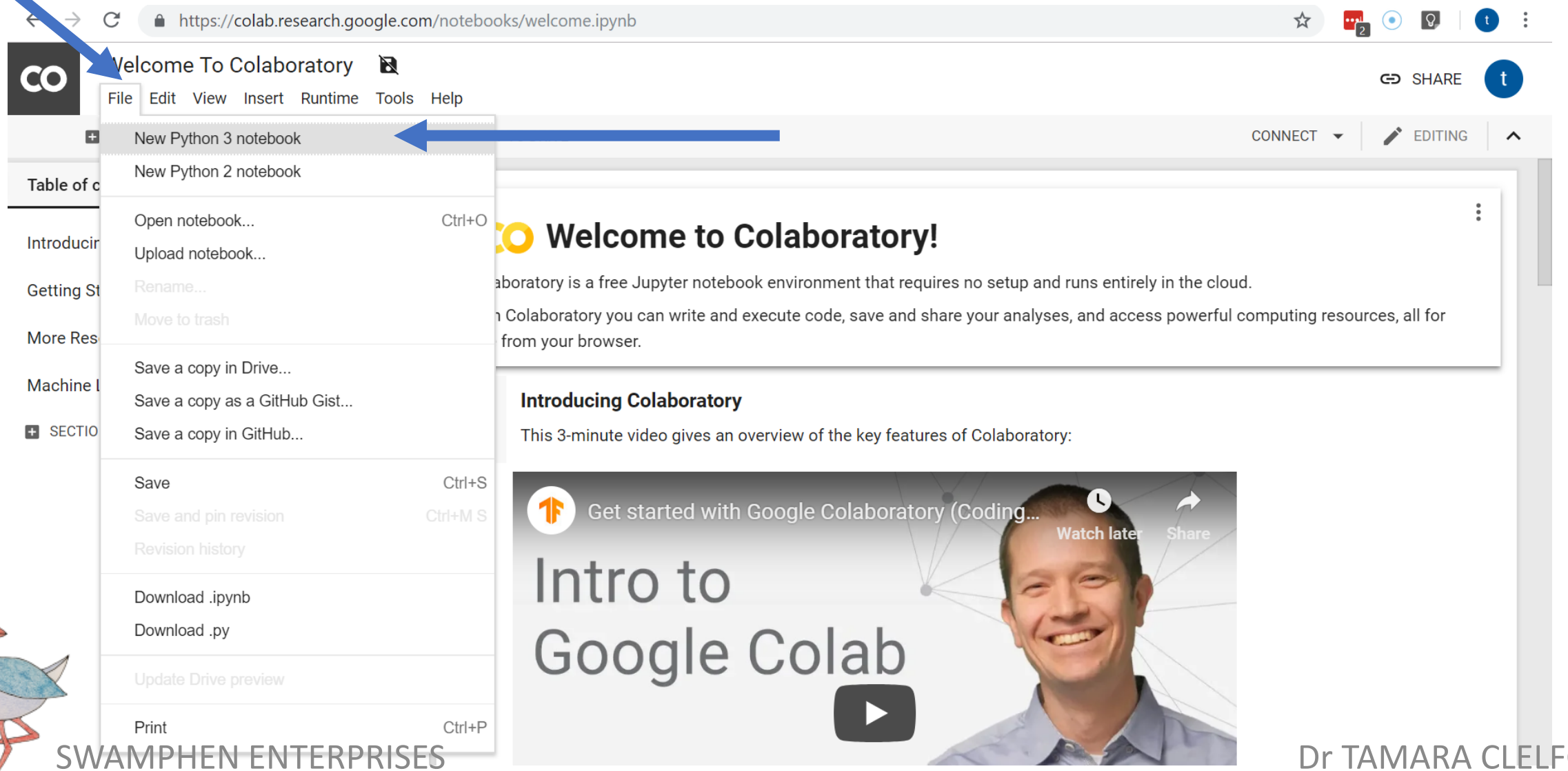
- Python programming language
  - Toolboxes such as NumPy, pandas and scikit-learn

- Integrated Development Environment (IDE)
  - Jupyter notebook style
  - Google Colab

GO TO GOOGLE COLAB
https://colab.research.google.com



OPEN A NEW PYTHON 3 NOTEBOOK



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

```
# import the data to google colab
from google.colab import files
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

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

```
# print out the headder
raw_data.head()
# try putting a number in the brackets
# what does .tail() do?
```

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

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

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

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

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

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

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

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

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

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

```
# describe the data set again
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

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

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

```
# check include
raw_data[include]
```

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

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

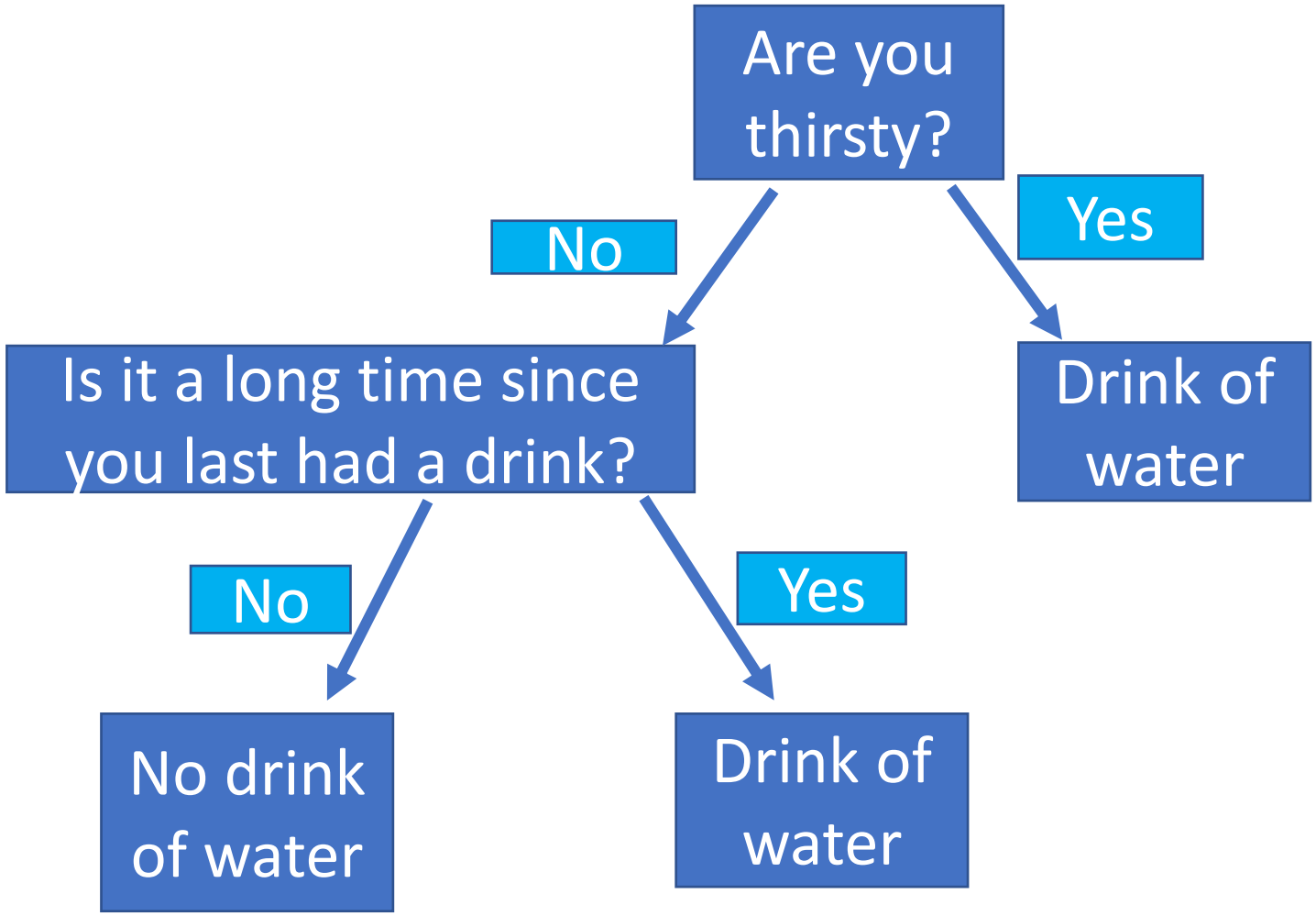
DATA SCIENCE FLOW

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

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

```
# import decision tree
from sklearn import tree
```

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

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

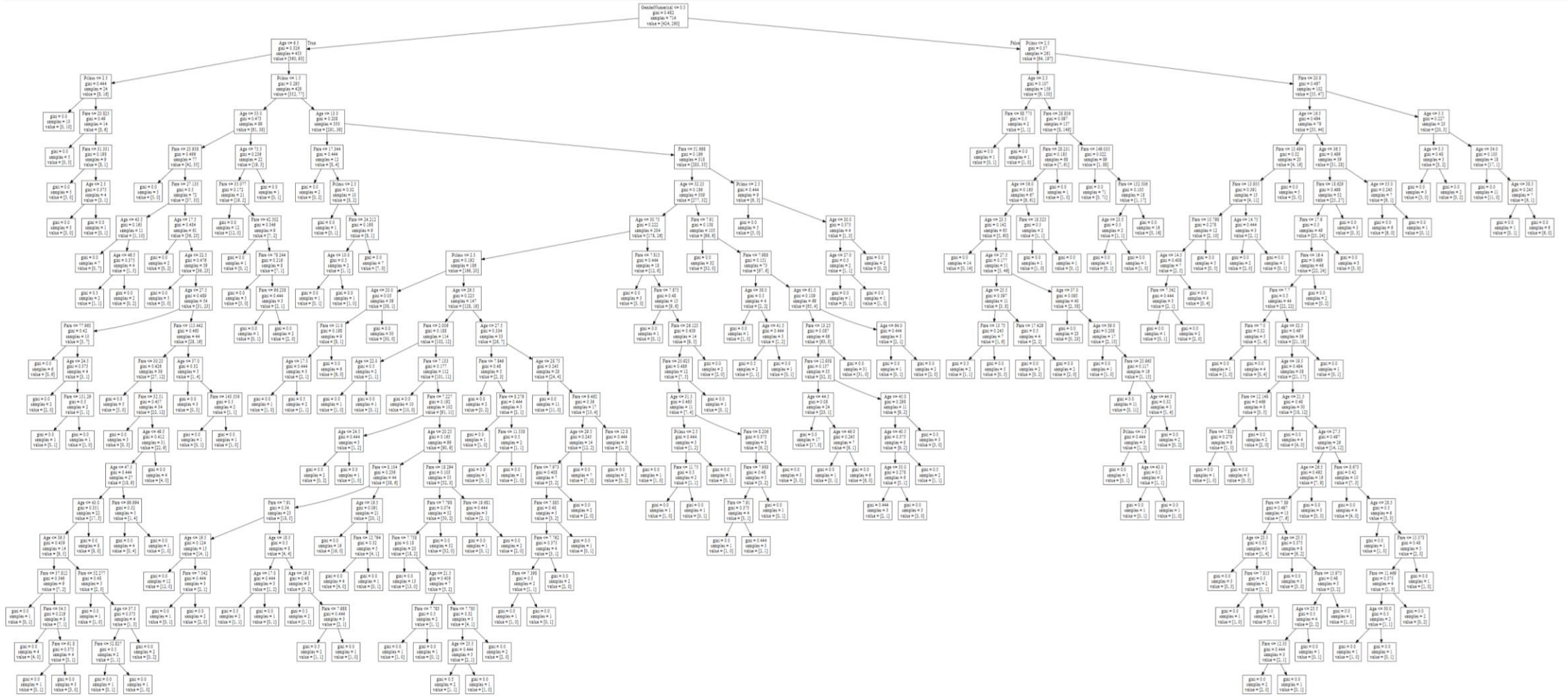
DATA SCIENCE FLOW

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))
```

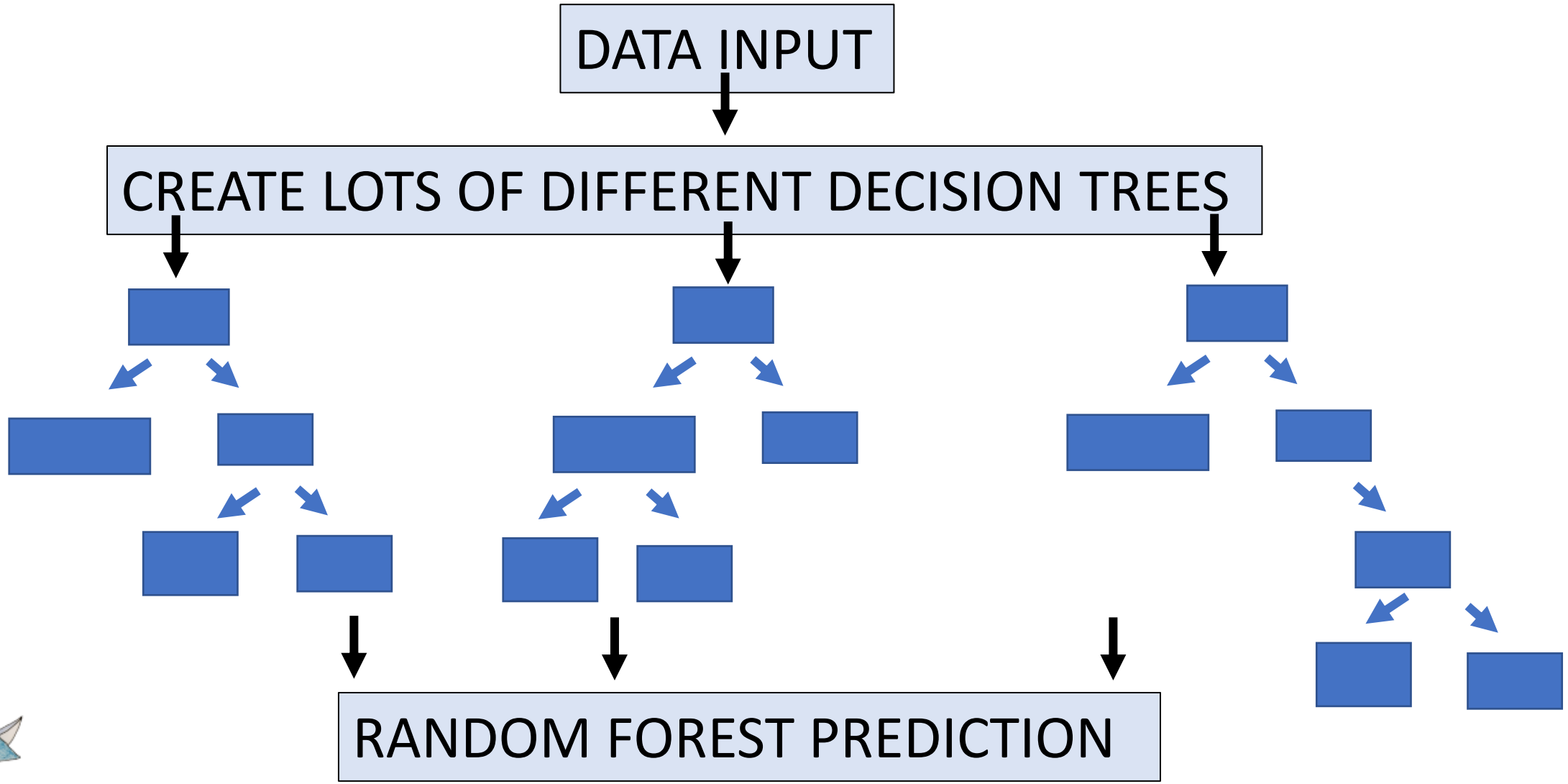
DECISION TREE



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

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

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

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

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

```
# see how good model is at predicting survival
from sklearn.model_selection import cross_val_score as CVS
scores_forest = CVS(titanic_forest, data, answer)
print(scores_forest)
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

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

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

```
# calculate survival score
survival_score = titanic_forest.predict_proba(person)
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

```
# did the person survive the sinking of the titanic?
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

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

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

```
# to look at male and female survival probability
survival_female = []
survival_male = []
selected_class = 2
average_fare = np.mean(data_set['Fare'])
gender_list = [0,1]
for i in age_list:
    for j in gender_list:
        if j == 1:
            array = survival_female
        else:
            array = survival_male
        person = np.array([[selected_class, i, j, average_fare]])
        survival_chances = titanic_forest.predict_proba(person)
        array.append(survival_chances[0,1]*100)
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

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

```
# plot male and female survival chances
plt.plot(age_list, survival_female, label = 'female')
plt.plot(age_list, survival_male, label = 'male')
plt.title('survival in ' + str(selected_class) + ' class')
plt.xlabel('age (y)')
plt.ylabel('survival chances (%)')
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 Clelford

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