Agenda

- 18:30 - Doors open , Drinks, and Networking
- 18:45 - Welcome and Opening remarks
- 18:55 - Hands-on Exercise!
- 19:30 - Break
- 19:40 - Data Science @ Dalia
- 20:10 - Drinks, Finger Food and More Networking

Irati

Academia & Industry

Data Scientist at Dalia,

with background in Computational Neuroscience

Kostas

Academia & Industry

Support data science - from an engineering point of view

Why?

Why (we hope) you are here

- Get an idea what data science is (and what it is not)
- Follow a hands-on walkthrough
- Learn more on how we use it to solve our surveys / ad-tech problem
- Drinks / Finger-food / Networking

Why are we here

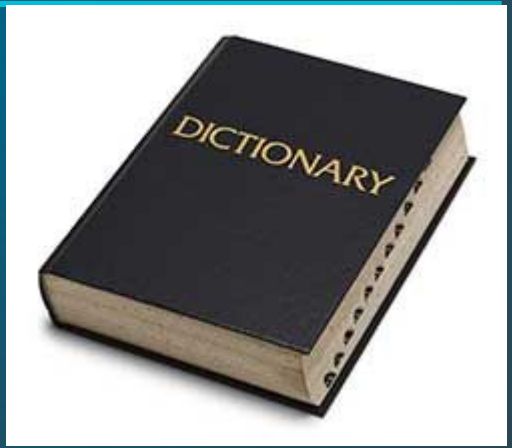
- We like to share our technical approaches
- Because we want to get some of your ideas and thoughts about it
- To let you know about Dalia

Data Process Walkthrough

Kostas Christidis, PhD

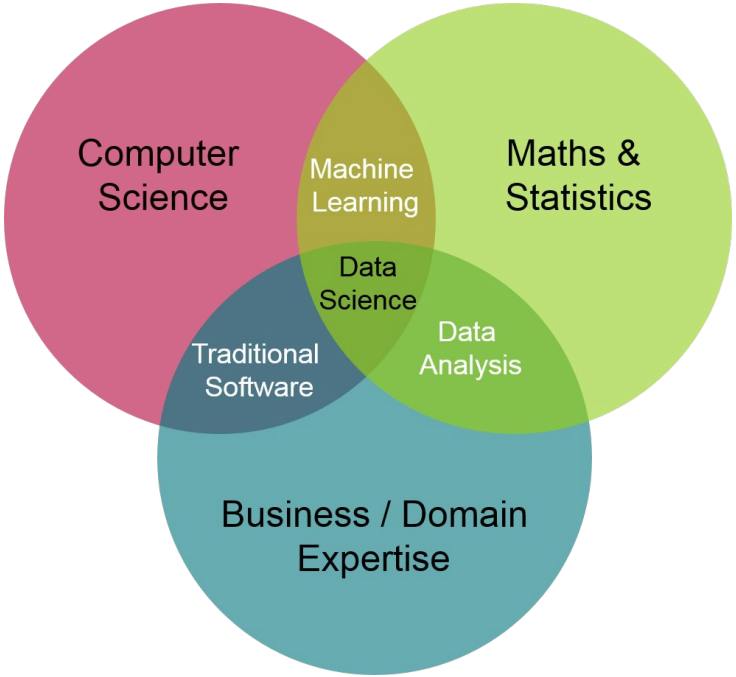
June 18, 2018

Definitions



Data Science

- Using data to understand the world.
- The art of uncovering the insights and trends behind data.
- Translate data into a story.



Eventually:

Study of data

Agenda

In the coming 30 minutes we will do the following:

- Tools of the Trade
- Data Loading / Storage
- Data Wrangling: clean, transform
- Plot and visualize

Please download the following:

https://github.com/DaliaResearch/DataScienceMeetup

Tools of the Trade

What we use

Amazon RDS

Please connect to WIFI

```
SSID: Dalia Guest

yes I am a guest
```

Installing Python

Install Anaconda

https://www.anaconda.com/download

Run the following in your terminal:

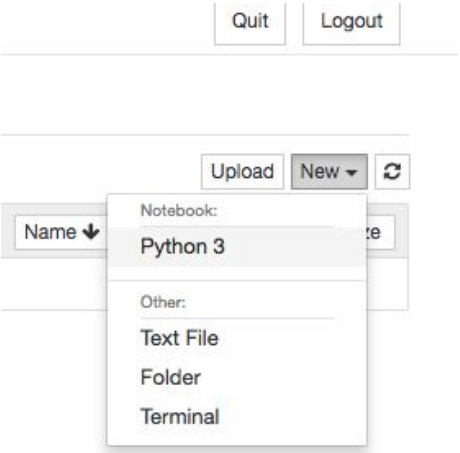
```
conda create -n dalia-meetup python=3.6

source activate dalia-meetup (conda activate
dalia-meetup)

conda install anaconda

jupyter notebook
```

Isolated python environments

Say it if something doesn't work!

Jupyter Notebook: Let’s create one!

Jupyter Notebook



Jupyter Notebook - Prepare

Import required libraries

```
import numpy as np  # main calculations - linear algebra
import pandas as pd # data processing
import matplotlib.pyplot as plt # visualisation
import seaborn as sns  # visualisation


# we want to see our plots inline
%matplotlib inline
```

Data: Read

Retrieve Dataset

Please download the following:

https://www.kaggle.com/claudiodavi/superhero-set/downloads/heroes_information.csv/1

https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017/downloads/results.csv/30

Unzip the file that ends with .zip

Read Data

```
football_results = pd.read_csv('results.csv')
football_results.head(10)
```

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
0	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	False
1	1873-03-08	England	Scotland	4	2	Friendly	London	England	False
2	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	False

```
heroes_information = pd.read_csv('heroes_information.csv')
heroes_information.head(10)
```

	Unnamed: 0	name	Gender	Eye color	Race	Hair color	Height	Publisher	Skin color	Align
0	0	A-Bomb	Male	yellow	Human	No Hair	203.0	Marvel Comics	-	go
1	1	Abe Sapien	Male	blue	Icthyo Sapien	No Hair	191.0	Dark Horse Comics	blue	go

We can read also from Excel, Json, Parquet, MySQL
While also we can configure options such as separator, delimiter, names, header

Pro Tip

Create an extra cell for each action you want to do.

Data: Clean, Transform

Clean / Transform Data

In many cases we do not have values.

```
heroes_information = pd.read_csv('heroes_information.csv')

print(heroes_information.shape) # Let's look at the size

heroes_information[heroes_information.isnull().any(axis=1)]
# Heroes with None data
```

	Unnamed: 0	name	Gender	Eye color	Race	Hair color	Height	Publisher	Skin color	Alignment	Weight
46	46	Astro Boy	Male	brown	-	Black	-99.0	NaN	-	good	-99.0
86	86	Bionic Woman	Female	blue	Cyborg	Black	-99.0	NaN	-	good	-99.0
138	138	Brundlefly	Male	-	Mutant	-	193.0	NaN	-	-	-99.0
175	175	Chuck Norris	Male	-	-	-	178.0	NaN	-	good	-99.0
204	204	Darkside	-	-	-	-	-99.0	NaN	-	bad	-99.0

```
clean_data = heroes_information.dropna()
print(clean_data.shape) # Size of the clean data
```

(719, 11)

Clean / Transform Data

In a world full of lies - we need to filter heroes that have weight and height

```
# We assume that real superheroes have weight and height


real_heroes = heroes_information[(heroes_information['Weight']> 0)  &
(heroes_information['Height']> 0)]

real_heroes.shape
```

(490, 11)

Find the outlier

We first calculate the score difference
And then we find the highest!

```
football_results['difference'] = abs(football_results['home_score']-
football_results['away_score'])



football_results.loc[football_results['difference'].idxmax()]
```

date	2001-04-11
home_team	Australia
away_team	American Samoa
home_score	31
away_score	0
tournament	FIFA World Cup qualification
city	Coffs Harbour
country	Australia
neutral	False
difference	31

Name: 23569, dtype: object

Change will happen

For example a country can change its name

```
df = football_results

df[(df['home_team']=="Macedonia")| (df['away_team']=="Macedonia") ]

df['home_team'] = df['home_team'].replace({'Macedonia': 'Northern Macedonia'})
df['away_team'] = df['away_team'].replace({'Macedonia': 'Northern Macedonia'})

df[df['home_team'].str.contains("Macedonia") |
df['away_team'].str.contains("Macedonia")  ]
```

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral	difference
17881	1993-10-13	Slovenia	Northern Macedonia	1	4	Friendly	Kranj	Slovenia	False	3
18073	1994-03-23	Northern Macedonia	Slovenia	2	0	Friendly	Skopje	Macedonia	False	2
18165	1994-05-14	Northern Macedonia	Albania	5	1	Friendly	Tetovo	Macedonia	False	4
18195	1994-06-01	Northern Macedonia	Estonia	2	0	Friendly	Skopje	Macedonia	False	2

What we did:

- Remove empty values
- Remove wrong values
- Found outliers (suspicious?)
- Replace values to match the business change

Say it if something doesn’t work!

Data: Explore

What’s the size of dataset

Shape

```
heroes_information = heroes_information.dropna()

heroes_information =
heroes_information[(heroes_information['Weight']> 0)  &
(heroes_information['Height']> 0)]

print(heroes_information.shape)
```

(489, 11)

Describe the dataset - in groups

Groupby count, sum

```
heroes_information.groupby('Gender')['name'].count()
```

```
Gender
-         14
Female    141
Male      334
```

```
weights_by_race = heroes_information.groupby('Race')['Weight'].mean()
weights_by_race.sort_values()
```

```
Race
Flora Colossus        4.000000
Cosmic Entity        16.000000
Yoda's species       17.000000
Kakarantharaian      18.000000
Animal               25.000000
……
```

What we did:

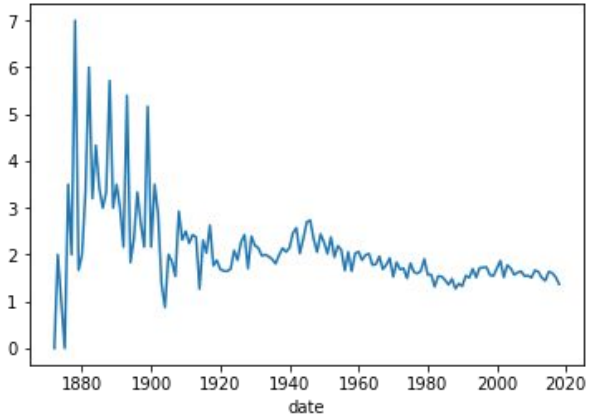
Looked at the shape of a dataset
Looked at some statistics - groups

Say it if something doesn’t work!

Data: Visualise

Line Plot

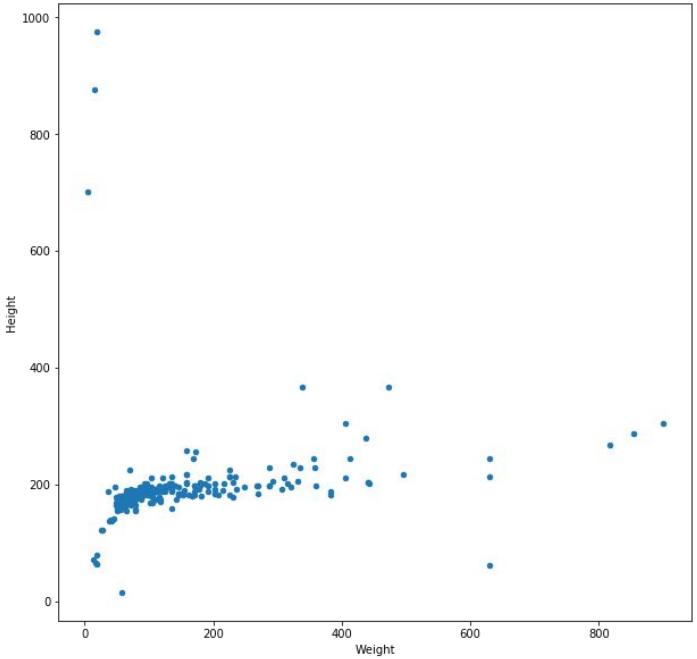
```
df = football_results.copy()

df['difference'] = abs(df['home_score']- df['away_score'])
df['date'] = pd.to_datetime(df['date']).map(lambda x: x.year)
result = df.groupby('date')['difference'].mean()

result.plot.line()
```



Scatter Plot

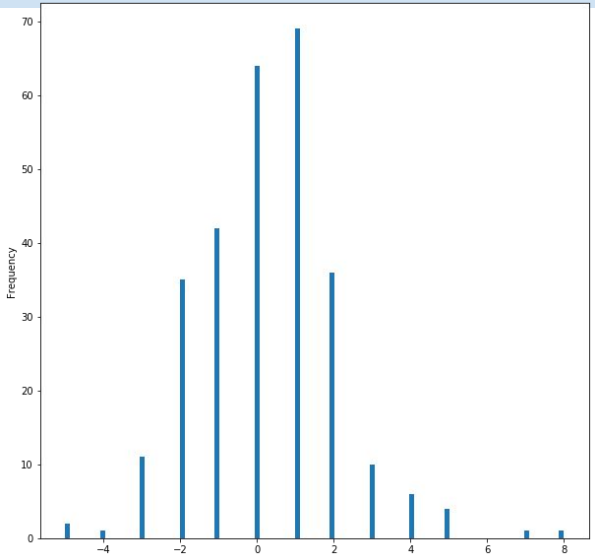
```
heroes_information[['Weight','Height']].plot.scatter(x='Weight',
y='Height')
```

What’s wrong with this plot?



Histogram Plot

```
greece_data = df[df['home_team']=='Greece'].copy()
greece_data['difference'] = greece_data.apply(lambda row: row['home_score'] -
row['away_score'],axis=1)


greece_data['difference'].plot.hist( bins=100,figsize =(10,10))
```



Notebook: Export

Export

```
greece_data.to_csv('greece_results.csv')
```

Q & A

Break Time!