DS-210: Programming for Data Science

Lecture 2: Select data science tools. Decision trees.

Survey outcomes

- Vedaant's office hours: Monday 3:45–5:45pm (**UPDATE: NOTE THE SMALL CHANGE**)

Survey outcomes

- Vedaant's office hours: Monday 3:45–5:45pm (**UPDATE: NOTE THE SMALL CHANGE**)

- Prerequisites:
    - We'll try to give you as many examples as possible
    - But you should study the self–study materials

Survey outcomes

- Vedaant's office hours: Monday 3:45–5:45pm (**UPDATE: NOTE THE SMALL CHANGE**)

- Prerequisites:
    - We'll try to give you as many examples as possible
    - But you should study the self–study materials

Task for next time:

- Get Jupyter Notebook or JupyterLab working on your computer
- Get comfortable using it

Python

- Assumption: Python already installed
- If not:
    - Either install directly
    - Or use Miniconda/Anaconda
- Make sure Python 3:

```
$ python --version
Python 3.9.9
```

Python

- Assumption: Python already installed
- If not:
    - Either install directly
    - Or use Miniconda/Anaconda
- Make sure Python 3:

```
$ python --version
Python 3.9.9
```

Jupyter Notebook

Interactive version based on iPython

If not installed, can be installed via `pip` or `conda`

```
$ pip install notebook
```

Run it with

```
$ jupyter notebook
```

Jupyter Lab

Newer, will obsolete Jupyter Notebook

```
$ pip install jupyterlab
$ jupyter-lab
```

CSV Files and Data input via `pandas`

Data often available in the CSV format:

```
Name,Number,PPG,YearBorn,TotalPoints
Kareem,33,24.6,1947,38387
Karl,32,25.0,1963,36928
LeBron,23,27.0,1984,36381
Kobe,24,25.0,1978,33643
Michael,23,30.1,1963,32292
```

- Header line optional
- Separators vary: "`,`" and "`;`" are popular
- Strings with spaces or separators may be in quotes

```
"Malone, Karl",32,25.0,1963,36928
```

CSV Files and Data input via `pandas`

Data often available in the CSV format:

```
Name,Number,PPG,YearBorn,TotalPoints
Kareem,33,24.6,1947,38387
Karl,32,25.0,1963,36928
LeBron,23,27.0,1984,36381
Kobe,24,25.0,1978,33643
Michael,23,30.1,1963,32292
```

- Header line optional
- Separators vary: "`,`" and "`;`" are popular
- Strings with spaces or separators may be in quotes

```
"Malone, Karl",32,25.0,1963,36928
```

Reading `.csv` or `.xlx` files is a popular task. Pandas are here to help you.

In [1]:
```
import pandas as pd
data = pd.read_csv('players.csv')
data
```

Out[1]:

	Name	Number	PPG	YearBorn	TotalPoints
0	Kareem	33	24.6	1947	38387
1	Karl	32	25.0	1963	36928
2	LeBron	23	27.0	1984	36381
3	Kobe	24	25.0	1978	33643
4	Michael	23	30.1	1963	32292

CSV Files and Data input via `pandas`

Data often available in the CSV format:

```
Name,Number,PPG,YearBorn,TotalPoints
Kareem,33,24.6,1947,38387
Karl,32,25.0,1963,36928
LeBron,23,27.0,1984,36381
Kobe,24,25.0,1978,33643
Michael,23,30.1,1963,32292
```

- Header line optional
- Separators vary: "`,`" and "`;`" are popular
- Strings with spaces or separators may be in quotes

```
"Malone, Karl",32,25.0,1963,36928
```

Reading `.csv` or `.xlx` files is a popular task. Pandas are here to help you.

```
In [1]: import pandas as pd
        data = pd.read_csv('players.csv')
        data
```

Out[1]:

	Name	Number	PPG	YearBorn	TotalPoints
0	Kareem	33	24.6	1947	38387
1	Karl	32	25.0	1963	36928
2	LeBron	23	27.0	1984	36381
3	Kobe	24	25.0	1978	33643
4	Michael	23	30.1	1963	32292

Access to specific cells and columns possible:

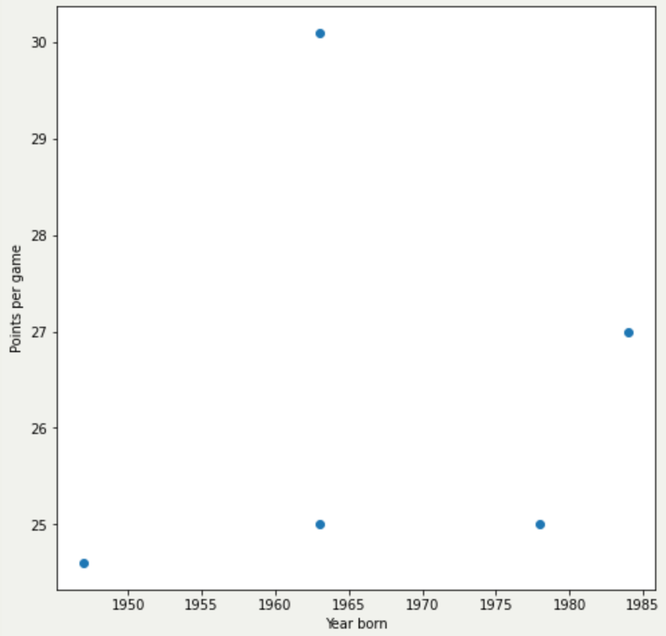
```
In [2]: data['YearBorn'][1]

Out[2]: 1963
```

Simple visualisation via `matplotlib`

In [3]:
```
print("Points per game vs. year born:")
import matplotlib.pyplot as plt
fig,ax = plt.subplots(figsize=(8,8))
ax.set_xlabel('Year born')
ax.set_ylabel('Points per game')
ax.scatter('YearBorn','PPG',data=data);
```

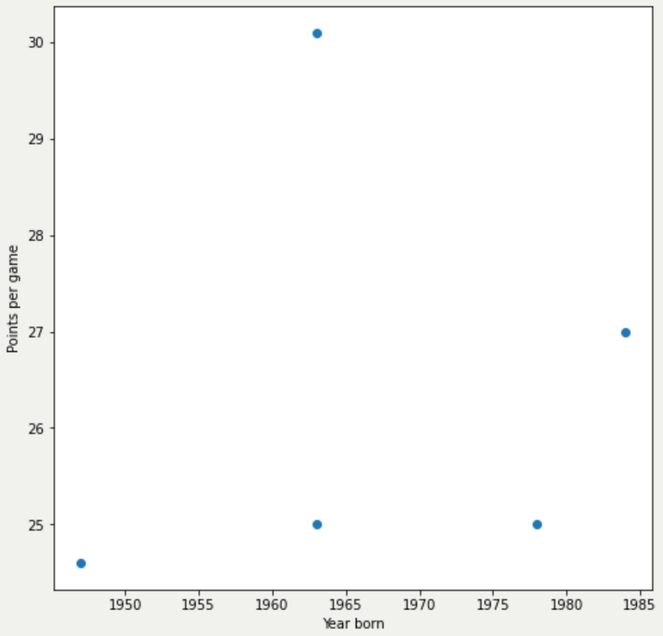
```
Points per game vs. year born:
```



Simple visualisation via `matplotlib`

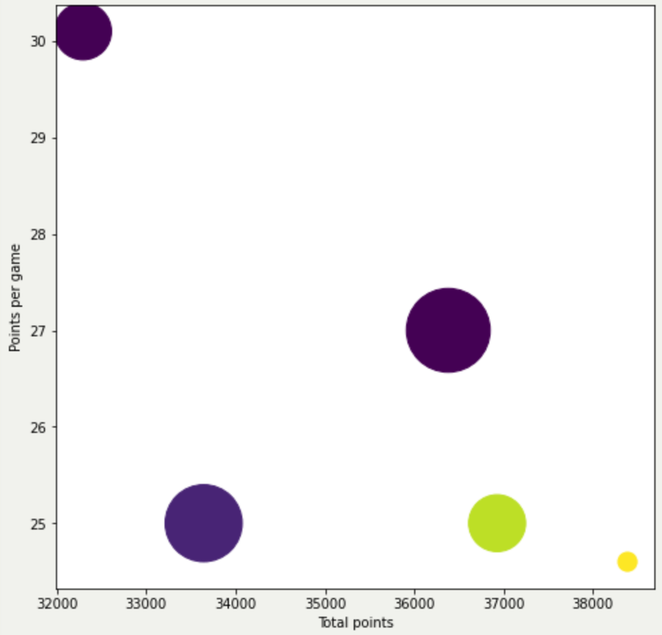
In [3]:
```
print("Points per game vs. year born:")
import matplotlib.pyplot as plt
fig,ax = plt.subplots(figsize=(8,8))
ax.set_xlabel('Year born')
ax.set_ylabel('Points per game')
ax.scatter('YearBorn','PPG',data=data);
```

```
Points per game vs. year born:
```



In [4]:
```
print("Other attributes: 3rd radius, 4th color")
data['radius'] = data['YearBorn'].subtract(1945)\
                                 .multiply(100)
fig,ax = plt.subplots(figsize=(8,8))
ax.set_xlabel('Total points')
ax.set_ylabel('Points per game')
#ax.scatter('TotalPoints','PPG','radius',data=data);
something_else = data
ax.scatter('TotalPoints','PPG','radius','Number',\
           data=something_else);
```

```
Other attributes: 3rd radius, 4th color
```



Machine learning: supervised vs. unsupervised

Supervised

- Labeled data
    - **Example 1:** images labeled with the objects: cat, dog, monkey, elephant, etc.
    - **Example 2:** medical data labeled with likelihood of cancer
- **Goal:** discover a relationship between attributes to predict unknown labels

Machine learning: supervised vs. unsupervised

Supervised

- Labeled data
  - **Example 1:** images labeled with the objects: cat, dog, monkey, elephant, etc.
  - **Example 2:** medical data labeled with likelihood of cancer
- **Goal:** discover a relationship between attributes to predict unknown labels

Unsupervised

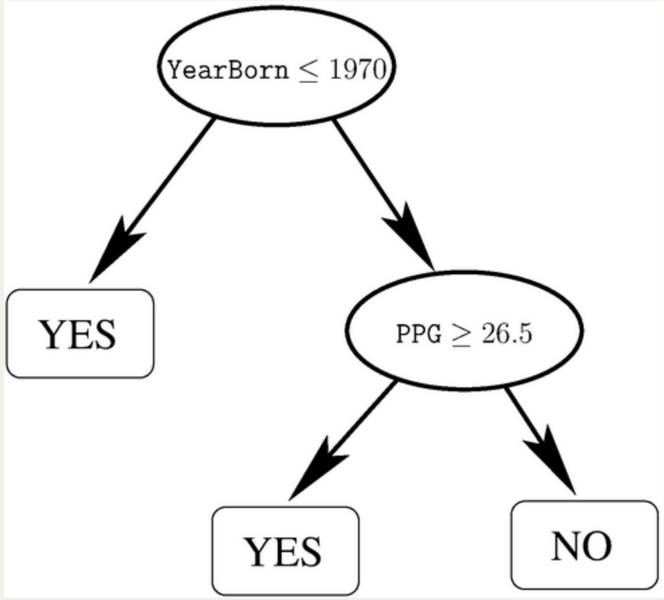
- Unlabeled data
- Want to discover a relationship between data points
- **Examples:**
  - *clustering*: partition your data into groups of similar objects
  - *dimension reduction*: for high dimensional data discover important attributes
  - generate random faces based on a sample you see

Supervised learning: Decision trees

Popular machine learning tool for predictive data analysis:

- rooted tree
- start at the root and keep going down
- every internal node labeled with a condition
    - if satisfied, go left
    - if not satisfied, go right
- leafs labeled with predicted labels

Does a player like bluegrass?

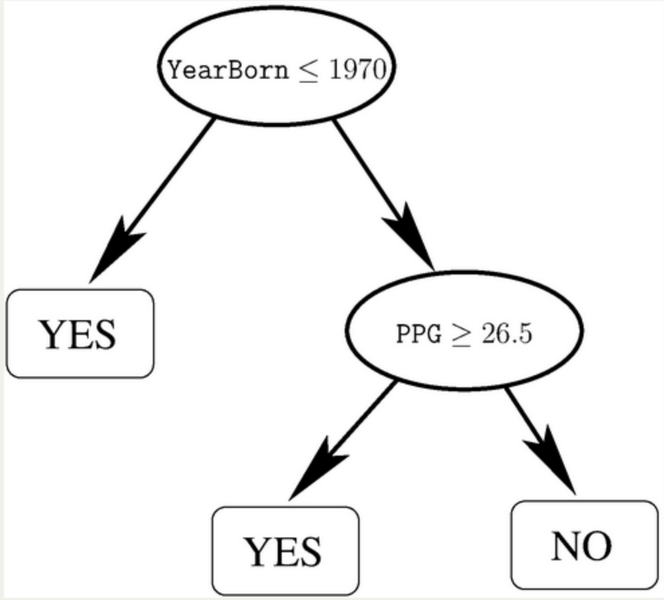


Supervised learning: Decision trees

Popular machine learning tool for predictive data analysis:

- rooted tree
- start at the root and keep going down
- every internal node labeled with a condition
  - if satisfied, go left
  - if not satisfied, go right
- leafs labeled with predicted labels

Does a player like bluegrass?



Big challenge: finding a decision tree that matches data!

Reminder: Task for next time

- Get Jupyter Notebook or JupyterLab working on your computer
- Get comfortable using it
- I will share my slides which are a Jupyter Notebook and a recording of this lecture