The UK's European university

University of Kent

COMP8270 / PROGRAMMING FOR ARTIFICIAL INTELLIGENCE

Fernando Otero
febo@kent.ac.uk
cs.kent.ac.uk/people/staff/febo

Overview:

1. Data Sets

2. Pandas
   - Load and Creating
   - Slicing
   - Indexing

Data Sets

Data Sets

- We have seen how it is advantageous to encapsulate related data.
  - E.g. a bank account
- All the bank accounts are a vital source of insight for a bank.
- To perform analysis (data analytics) we usually want to group by field, not by object.
- For online transactions we group by object.
- For offline analysis we group by object data attributes (fields).

Group by Field, not Object

```
Account_t
     Balance = [100]
     Owner = "Sally”
     Age = 25

Account_t
     Balance = [1500]
     Owner = "Tariq”
     Age = 34

Account_t
     Balance = [10000]
     Owner = “Frank”
     Age = 55

Account_t
     Balance = [30000]
     Owner = “Katie”
     Age = 65
```

Balance	Owner	Age
[100]	"Sally”	25
[1500]	"Tariq”	34
[10000]	“Frank”	55
[30000]	“Katie”	65

- Suppose the bank wanted to understand the relationship between age and balance?

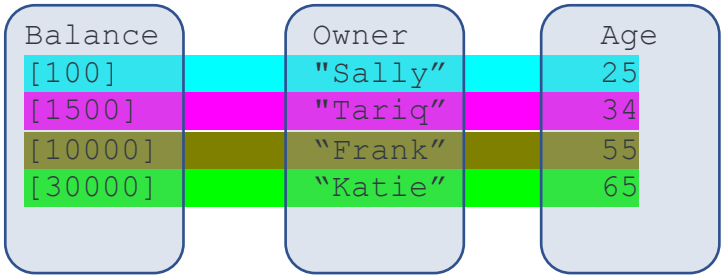
Group by Field, not Object

```
Account_t
     Balance = [100]
     Owner = "Sally”
     Age = 25

Account_t
     Balance = [1500]
     Owner = "Tariq”
     Age = 34

Account_t
     Balance = [10000]
     Owner = “Frank”
     Age = 55

Account_t
     Balance = [30000]
     Owner = “Katie”
     Age = 65
```



- Suppose the bank wanted to understand the relationship between age and balance?

Pandas Data Frames

Pandas Data Frames

- Pandas is a contraction of “panel data”
- The primary abstraction is the data frame.
- A Pandas dataframe is a Python object that encapsulates all of our data.
- A dataframe can passed around and tracked.
- We can:
  - Process
  - Search
  - Subset
  - Save

Using Pandas…