Block 3

Basic Statistics and Data Wrangling

Outline Block 3

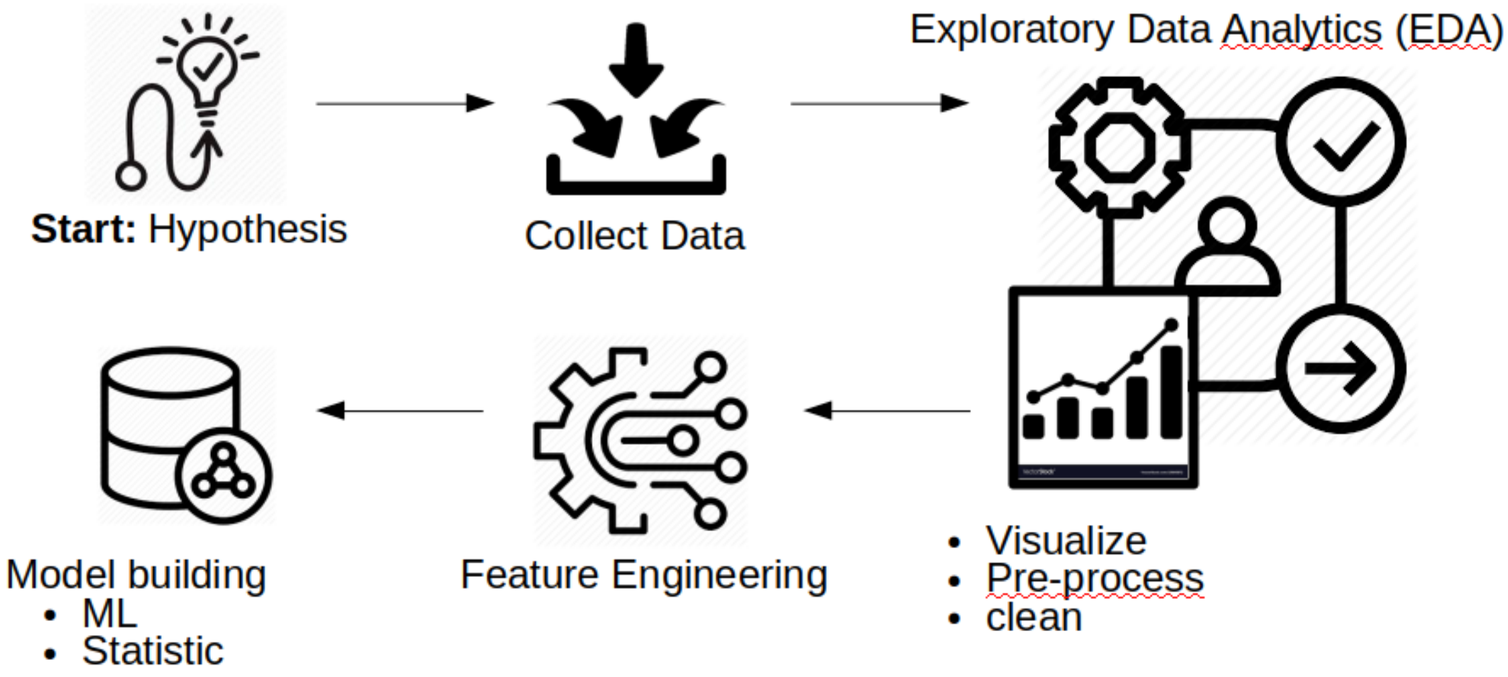
- Lecture 1: Data Wrangling
    - with sample use cases in ***Python***
- Lecture 2: Basic Statistics
    - with sample use cases in ***Python***
- Lab 1: Introduction to ***Pandas***
    - with use cases
    - and exercises
- Lecture 3: the "Group by" Operator
- Lab 2: Group by exercises

Data Wrangling

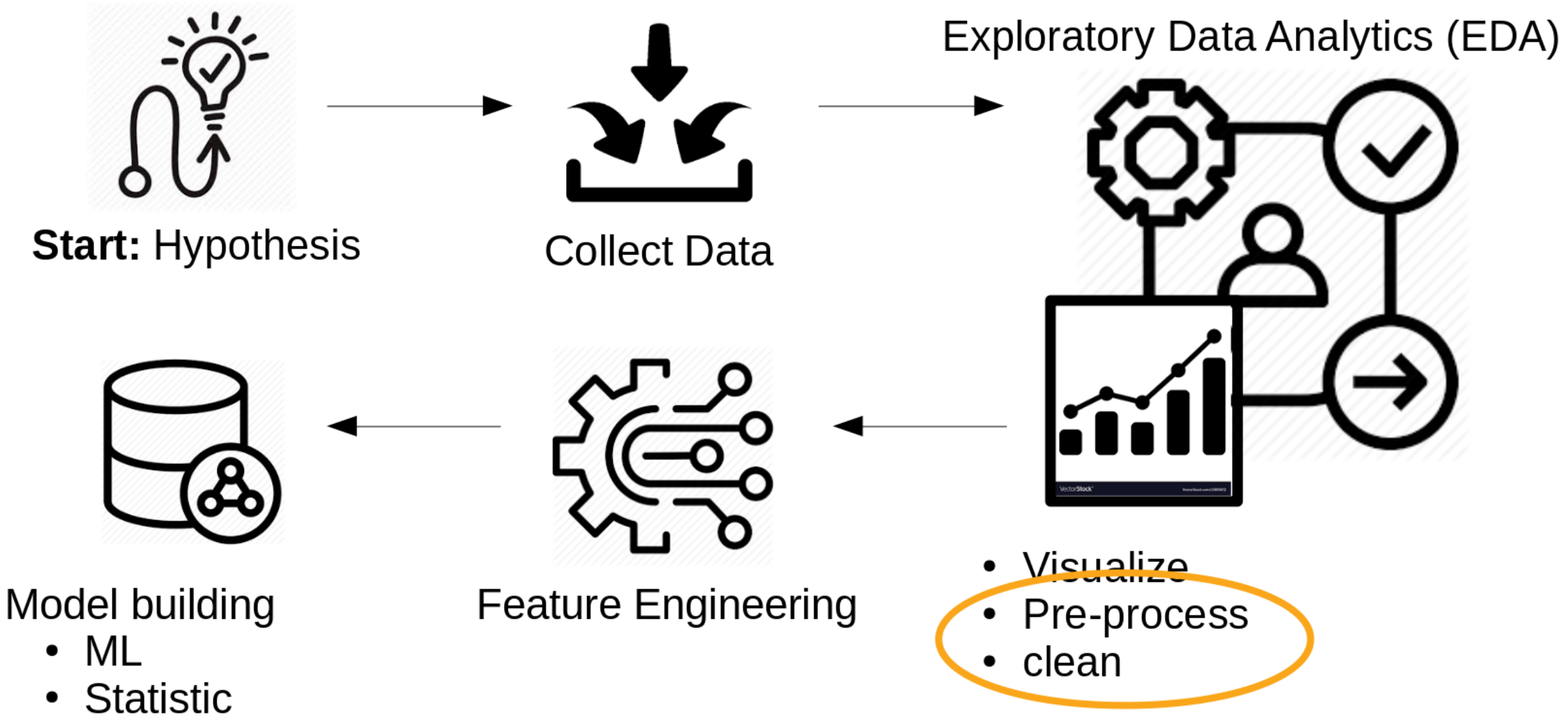
Outline

- Data Science Processing Pipeline
- What is ***Data Wrangling***?
    - Stages of ***Data Wrangling***
- Short Introduction to ***Pandas***
- ***Wrangling*** by Use Cases (Lab session)

Data Science Processing Pipeline



Data Science Processing Pipeline

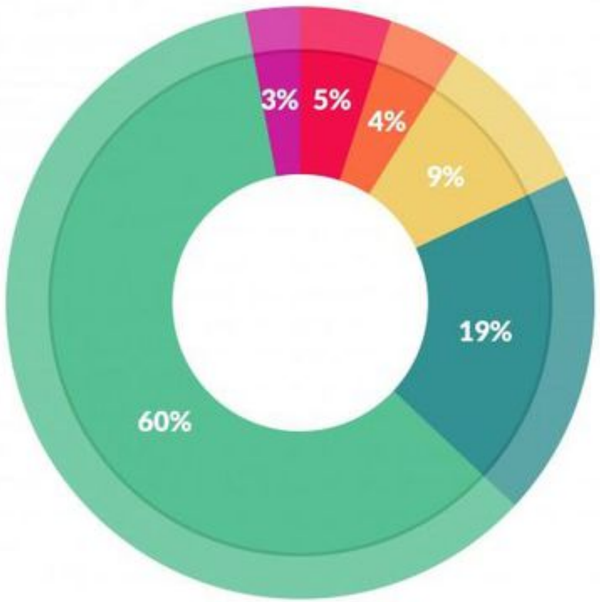


What is *Data Wrangling* ?

What is *Data Wrangling* ?

Definition:

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. [wikipedia]



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

[source: study by forbes.com: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#]

Phases of *Data Wrangling*

- (Scrape)
- Clean
- Transform
- Merge
- Reshape -> Rectify

Phases of *Data Wrangling*

- (Scrape): ***get data from sensors, internet, databases, ...***
- Clean
- Transform
- Merge
- Reshape -> Rectify

Phases of *Data Wrangling*

- (Scrape)
- Clean : ***remove "bad data"***
- Transform
- Merge
- Reshape -> Rectify

Phases of *Data Wrangling*

- (Scrape)
- Clean
- Transform : ***change/correct data formats, recompute, ...***
- Merge
- Reshape -> Rectify

Phases of *Data Wrangling*

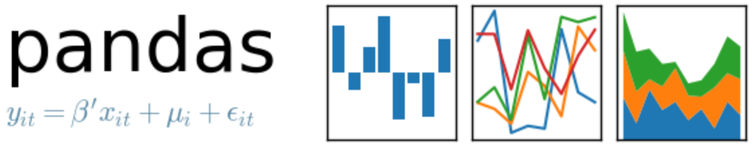
- (Scrape)
- Clean
- Transform
- Merge: ***combine and connect data sources***
- Reshape -> Rectify

Phases of *Data Wrangling*

- (Scrape)
- Clean
- Transform
- Merge
- Reshape -> Rectify: ***output: vectors, arrays, tables***

Wrangling in Python with Pandas

Started as *"**spread sheets for python**"* - now has become one of the most important ***Data Wrangling*** and **EDA** tools in ***Python***



pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python has long been great for data munging and preparation, but less so for data analysis and modeling. pandas helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like **R**.[pandas website]

Pandas Documentation

- Pandas website: https://pandas.pydata.org/
- Pandas user guide: http://pandas.pydata.org/pandas-docs/stable/user_guide/index.html
- Pandas API documentation: http://pandas.pydata.org/pandas-docs/stable/reference/index.html
- VERY USEFULL: Pandas Cheat Sheet: https://github.com/pandas-dev/pandas/blob/master/doc/cheatsheet/Pandas_Cheat_Sheet.pdf

Pandas in a Nutshell

In [2]:
```
#import the pandas module
import pandas as pd #naming convention for pandas is pd
```

The central element of *Pandas* is the *DataFrame*

- spreadsheet like data structure
- rectifies data into tables
- database like functionality
- arrray compatible

In [3]:
```
d=pd.read_csv(path+'/DATA/weather.csv') #read some data from file
d.head()#show first rows of the DataFrame
```

Out[3]:

	Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Loud Cover	Pressure (millibars)	Daily Summary
0	2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251.0	15.8263	0.0	1015.13	Partly cloudy throughout the day.
1	2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259.0	15.8263	0.0	1015.63	Partly cloudy throughout the day.
2	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204.0	14.9569	0.0	1015.94	Partly cloudy throughout the day.
3	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269.0	15.8263	0.0	1016.41	Partly cloudy throughout the day.
4	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259.0	15.8263	0.0	1016.51	Partly cloudy throughout the day.

Pandas Features

- Data in- and export
- DataFrame (DF) data structure with functionality of
    - spreadsheet
    - relational data base
- DF Statistcs
- DF Visualization
- Rich library of ***wrangling*** methods

Pandas Features

- Data in- and export
- DataFrame (DF) data structure with functionality of
    - spreadsheet
    - relational data base
- DF Statistcs
- DF Visualization
- Rich library of ***wrangling*** methods

Detailted introduction in the Lab session!

- With wrangling use cases ...