DATA7001
Data Science Process Recap

Nan Ye

Data Science Process

- Problem Solving with Data
- Getting the data I need
- Is my data fit for use
- Making the data confess
- Storytelling with data

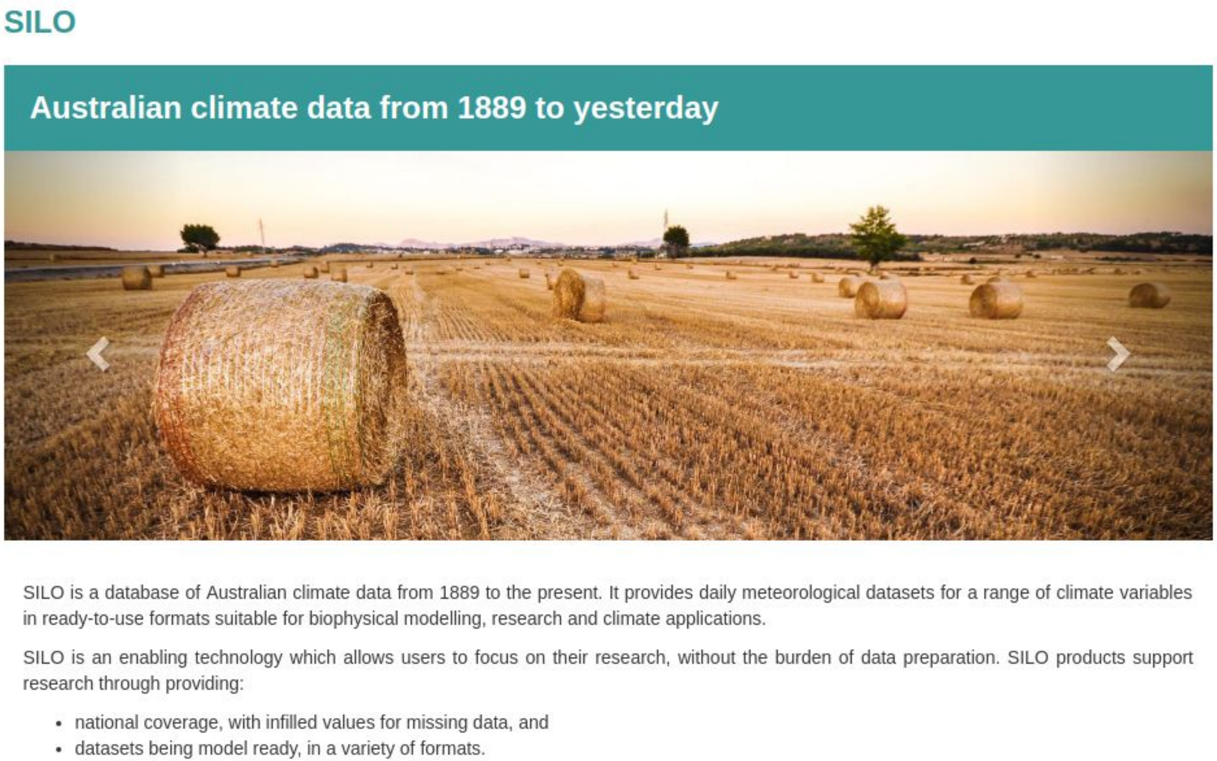
In theory, follow the steps => Done.
In practice, step 1 -> step 2 -> step 1... it often takes a few iterations.

This lecture: brief walk through of the process using a case study.

Task

- Often, as a data scientist, we need to solve other people’s problems with data
- Assume that you’re hired to study how climate changes
  - Your task: What happened? What will happen?
  - Importantly, you would like to know why the task is significant.
- Plan: let’s do this for all places in the world
  Reality: missing data for many places, different data formats,...
  Revised: let’s focus on Australia using the SILO dataset

SILO Dataset



SILO

Australian climate data from 1889 to yesterday

SILO is a database of Australian climate data from 1889 to the present. It provides daily meteorological datasets for a range of climate variables in ready-to-use formats suitable for biophysical modelling, research and climate applications.

SILO is an enabling technology which allows users to focus on their research, without the burden of data preparation. SILO products support research through providing:

- national coverage, with infilled values for missing data, and
- datasets being model ready, in a variety of formats.

https://www.longpaddock.qld.gov.au/silo/

Preliminary Study

- Who provides the data?
- How is the data collected?
- What are recorded?
- How to access the data?
- Quality of the data?

- Provided by Queensland Government and Bureau of Meteorology
- SILO (Scientific Information for Land Owners) is a database of Australian climate data from 1889.
- Observed data at 4000+ meteorological stations
- Access data: https://www.longpaddock.qld.gov.au/silo/about/access-data/
- A downloaded subset for Master of Data Science Program
  - https://mdatascience.uqcloud.net/SILO_PPD/
  - Metadata
    - filename, and the id, name, elevation and lat-lon of station.
  - Fields in data file
    - Date, Max temp, Min temp, Rainfall, Evaporation, Radiation and Vapour Pressure...
    - From different sources (station, nearby station, ...)

- The dataset has missing values - they are ‘patched’ with interpolated data.
  - Can we tell whether we are using measured data or interpolated data?
  - There is a data source field indicating the data source

    " 0 = Station data, as supplied by Bureau, 23 = Nearby station, data from BoM "
    " 13 = Deaccumulated using nearby station, 15 = Deaccumulated using interpolated data"
    " 35 = interpolated from daily observations using anomaly interpolation method for CLIMARC data
    " 25 = interpolated daily observations,   75 = interpolated long term average"
    " 26 = synthetic pan evaporation "

  - Can we visualize the sources of the data for a station?

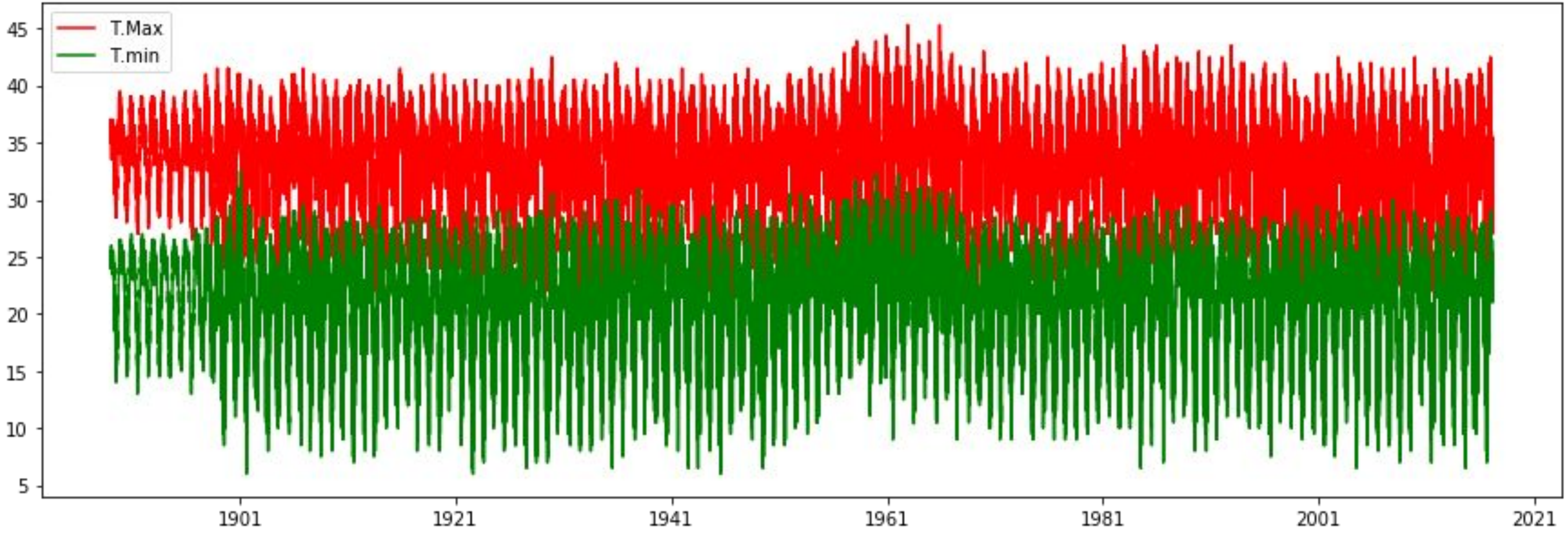
Data Collection

- We have already downloaded a subset of the dataset for you.
- How will you collect the data, if you need to do it on your own?
  - Do you need to write a crawler?
  - Do you need to ask for permission?
  - Any conditions for using the data?
- How can you store the data?

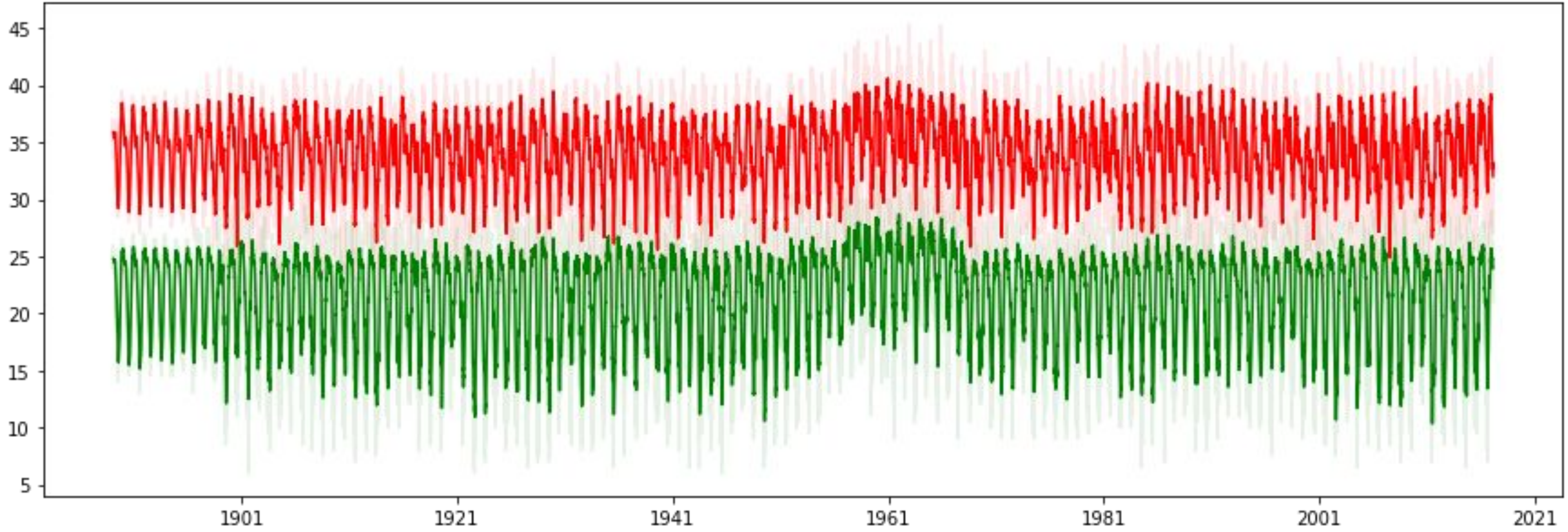
Exploratory Analysis

- We generate various plots to help us better understand the data
  - E.g. plot of one variable against another, histogram, boxplot,...
- Some questions
  - How is a variable distributed?
  - Are two variables correlated?
  - Are there any obvious issues with the data?
- Useful if we want to build a model for the data

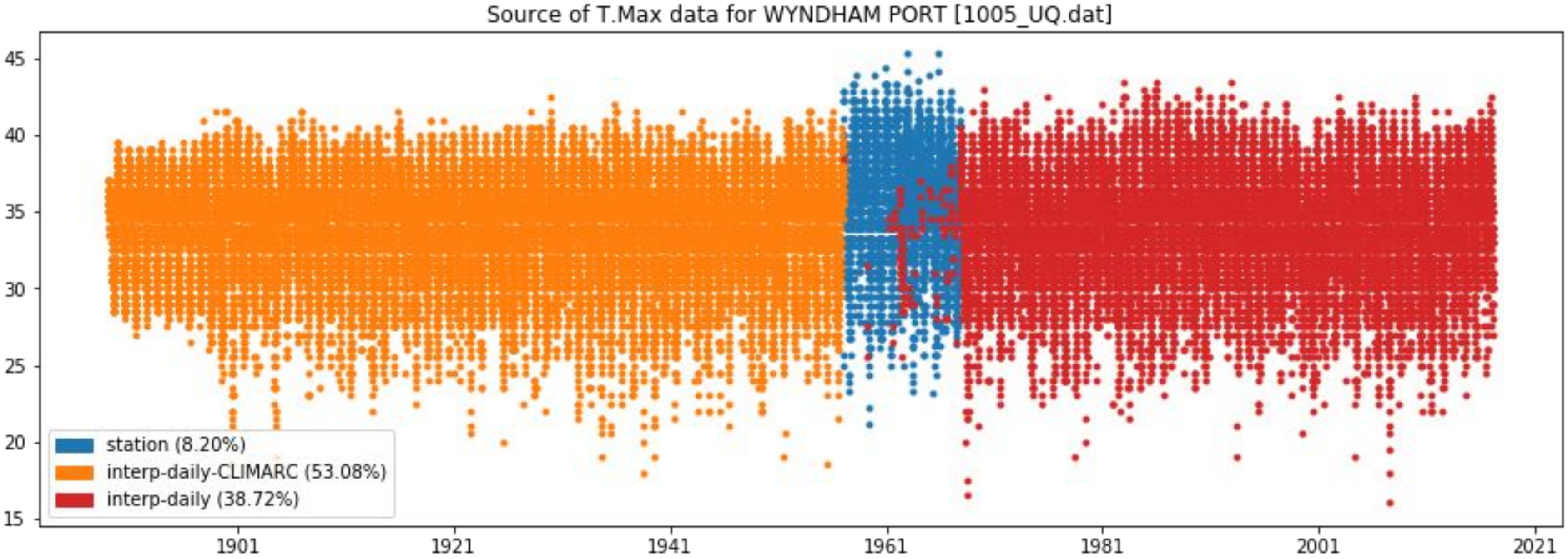
- Any pattern in the data?



- Smoothing the data sometimes help to make pattern clearer



- Is it climate change or a problem with data?



- Further EDA
    - Are there any spatial patterns?
    - Is there a hottest or coldest area?
    - Is the temperature increasing on average?
    - How is min daily temperature correlated with the max daily temperature?

Making the Data Confess

- What models are likely to be a good fit?
  - Is a model’s assumptions obviously violated?
- Can we use the model to formulate possible explanations for the observed data?
  - Simple models like linear regression helps us to identify important factors.
- How can we check whether a model is good enough for making predictions?
  - This need to be separately evaluated on a different dataset.
  - Cross-validation score can be useful.

Storytelling with Data

- Describe not just your approach, but why you used the approach
- Always connect back to the task that you want to solve

Know Your Tools

- Python
  - Web-scraping with requests
  - Extracting information with re
  - Processing tabular data with pandas
  - Visualization using matplotlib
  - Scipy, NumPy
  - Machine learning using sklearn
- Apache Hive and SQL
- R