



**HKUST**  
VISLAB

# **COMP 4462**

## **Data Visualization Tutorial**

Leo Yu Ho, Lo  
Ming Yao

Tuesday 12 March, 2019

# Pandas and visualization

- Pandas

- Best data processing library
  - Fast and easy to use, inspired by R language
  - Can handle large dataset as long as it fits in memory (with some workarounds if not)
- Can read many common file formats, e.g. csv, json, xlsx, sql
- Easy to make plots with matplotlib / seaborn / bokeh / **altair**
- Easy to pass data around with different libraries, like numpy, scikit-learn, etc.

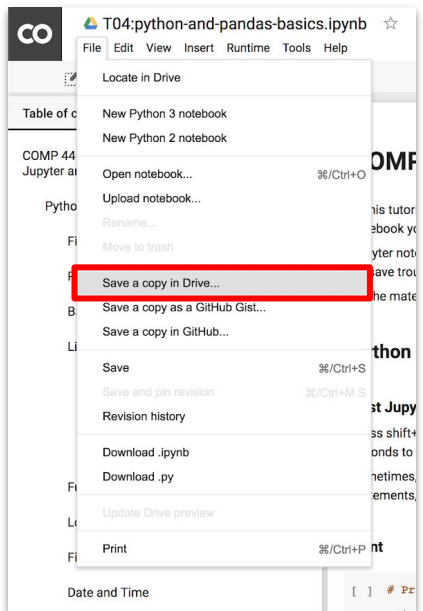
- Altair

- Designed to work with pandas
- API designed with visualization language
  - Marks, encoding channels, data types, scale, interaction idioms, layers, facet
  - Backed by Vega-Lite, which is backed by D3.js
- Web based visualization
  - Passing visualization specification to Jupyter notebook, then visualize by browser with js
  - matplotlib / seaborn are python based with native GUI

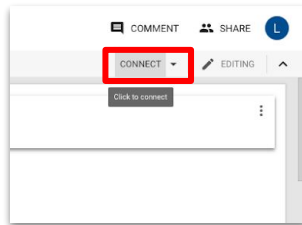
# Google Colab

1. Sign in your Google account
2. Go to the [notebook of this tutorial](#)

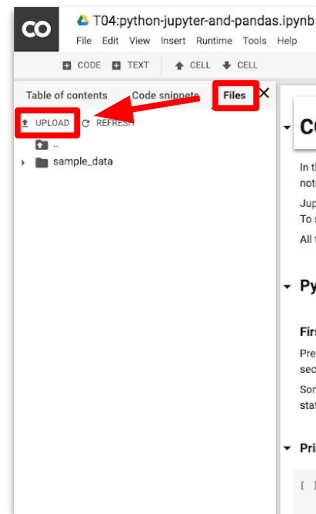
## 3. Make a copy



## 4. Connect



## 5. Upload dataset



# Pandas / Altair

- See the [Jupyter notebook on Google Colab](#)
- Topics on Pandas:
  - Load data (csv, json) / summary / clean data / handle null values
  - Data selection / filter
  - Sort / groupby / aggregate
  - Join dataframes
  - Pivot and melt
  - Rename column / compute new attributes
- Topics on Altair:
  - Marks, encoding, scale
  - Basic charting (bar chart / line chart / scatter plot)
  - Multiplots / juxtapose (side-by-side)
  - Interactions / interactive filtering

# Lab exercise

- Tasks

- Open [this Google Colab notebook](#), make a copy and connect
- Download countries dataset (countries.json) from [GitHub](#)
- Read through the notebook and fill in the “TODO” cells
- Find out the top songs / artists in the world, different continents or regions
- Find out the trends of your favorite songs / artists over the year
- Plot your findings
- Print the whole web page as .pdf and upload to Canvas

- Optional

- Explore the Spotify dataset
  - Find out more about your favorite songs / artists
  - Find out songs you may like
- You may further explore with the [Spotify Song Attributes](#) dataset
  - A lot of interesting attributes: “danceability”, “energy”, “instrumentalness”, etc.
- Do the lab with music!

# More topics on Pandas and Visualization

- More on Pandas
  - **Dataframe:** load data from python objects / feather data format / slicing with selector / dataframe indexing / hierarchical indexing / data join (inner, outer, left, right)
  - **Data cleaning:** fill in missing value / interpolation / discretization and binning / permutation and random sampling / string operation / regex
  - **Grouping operations:** split-apply-combine / customized aggregation functions
  - **With other data science libraries:** scikit-learn / statsmodels / weighted average / correlation / regression
  - **Plotting libraries:** matplotlib / seaborn / bokeh / plotly
- More on Altair
  - **Charts:** stacked (bar, area) charts / streamgraph / histogram / maps
  - **Adjust charts elements:** axis / labels / legends / colors
  - **Data transformation:** aggregate / bin / custom calculation
  - **Interactions:** binding / brush / pan and zoom / selection
  - **Compound Charts:** overlay / horizontal concat / vertical concat / faceted / repeated charts

# Next tutorial

Javascript basics and  
Observable

- Register on [Observable](#)
  - A notebook like environment
    - But with Javascript and solely on browser
    - Interactive
    - Design for visualization
  - Free!
  - No setup