Data Science Learning Exchange Software Carpentry Workshops Session I – April 20, 2021

See the exit Poll in the chat box after every session

Workshop Organizers

Frederick National Laboratory for Cancer Research

Special thanks to:

- **Lynn Borkon:** AI and HPC collaboration Development, SDSI, FNL

- **Petrina Hollingsworth:** Community engagement Manager, SDSI, FNL

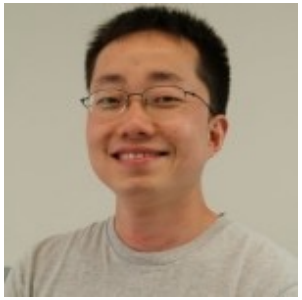
- ***Mike Rinaldi:*** CBIIT AV Expert

Scientific Computing Team

Frederick National Laboratory for Cancer Research

Meet the instructors:

- **Pinyi Lu:** Bioinformatics Analyst, Scientific Data Science Initiative, FNL



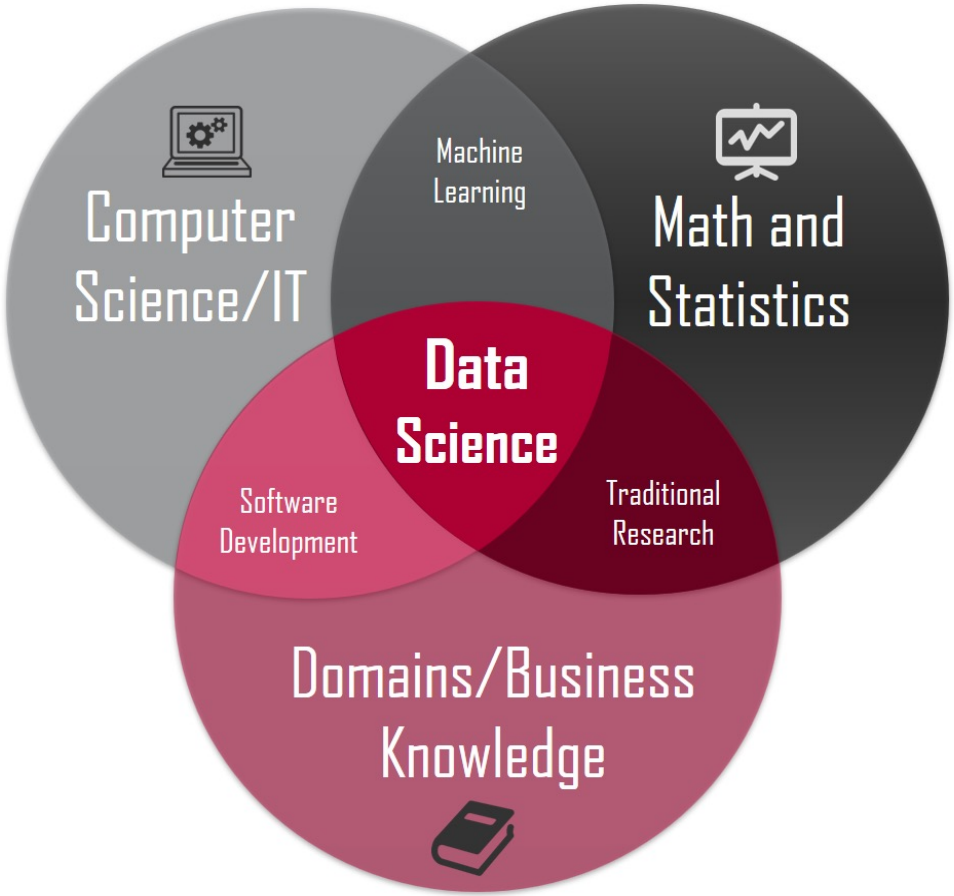
- **Robin Kramer:** Bioinformatics Analyst, Essential Software Inc

- **George Zaki:** Bioinformatics Manager Scientific Data Science Initiative, FNL

Data Science Initiative FNL/CBIIT

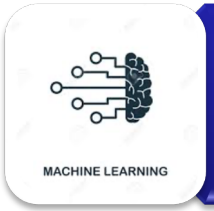
Frederick National Laboratory for Cancer Research

Leverage breakthrough advancements in scientific computing and data science to help NCI scientific staff advance basic research, understanding, and treatments in cancer.



Data Science and Scientific Computing Activities

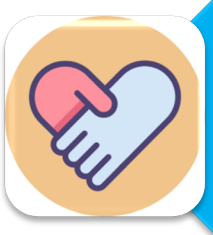
Frederick National Laboratory for Cancer Research



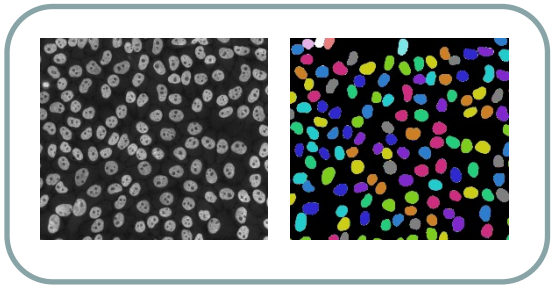
Machine Learning for Image Processing, NGS, Drug, and NLP

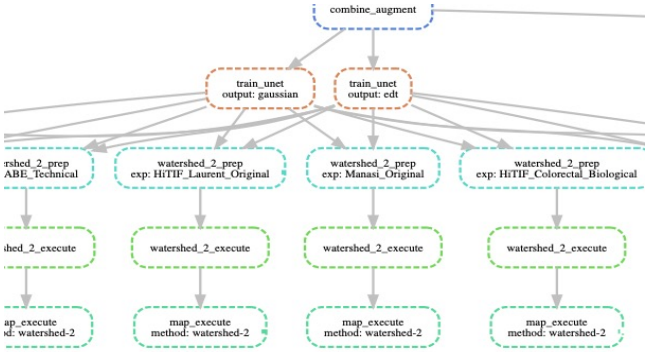
HPC Workflows Development (e.g., Snakemake)

Custom Code Optimization & Accelerated Computing (GPUs/FPGAs)



Education, Outreach/Community Building [Biowulf, FRCE, & others]





NIH.AI

Software Carpentry

- Teaching researcher the computing skills since 1998
- Lessons prepared by volunteers from world wide academic and research institutions
- Software carpentries has lessons in:
  - Python
  - R
  - Unix shell
  - Version control in git
  - Databases and SQL,
  - More….

Software Carpentries Workshops

Frederick National Laboratory for Cancer Research

- **Workshop Materials:**
  - Plotting and programming with Python
  - http://swcarpentry.github.io/python-novice-gapminder/
- **Live coding:**
  - Instructor writes code and explains concepts
  - Learners write code, ask questions
  - Use two screens (one to write your code, one to read instructions/watch instructor)
- **Immediate response**
  - Post question in the chat and get feedback from co-instructors
- **Github repo for materials and recordings:**
  - https://cbiit.github.io/p2p-datasci/2021-03-25-introduction_to_python/

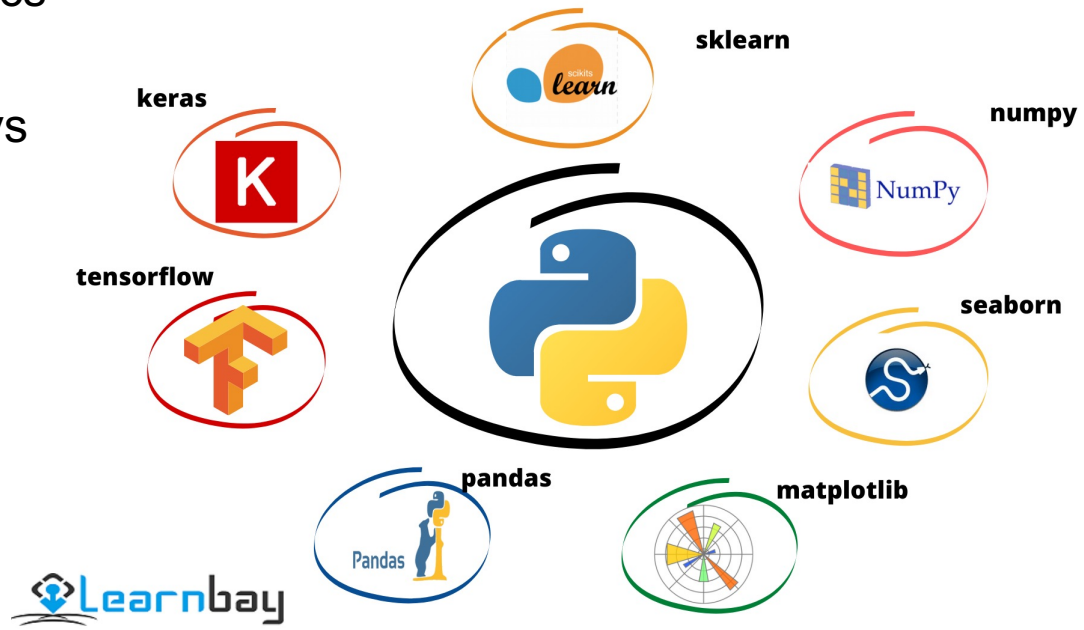
Tentative Course Schedule

Frederick National Laboratory for Cancer Research

- **Week 1:** Introduction to Python and Colab, Running and Quitting, Variables and Assignment
  - **NOTE**: A one-hour help session will be offered on **April 23, 11 AM – 12 PM**: *Getting Started with Google Colab*
- **Week 2**: Data Types and Type Conversion, Built-in Functions and Help, Libraries
- **Week 3:** Reading Tabular Data into DataFrames, Pandas DataFrames, Plotting 1
- **Week 4**: Plotting 2, Lists, For Loops
- **Week 5**: Conditionals, Looping Over Data Sets, Writing Functions
- **Week 6**: Variable Scope, Programming Style, Wrap-Up

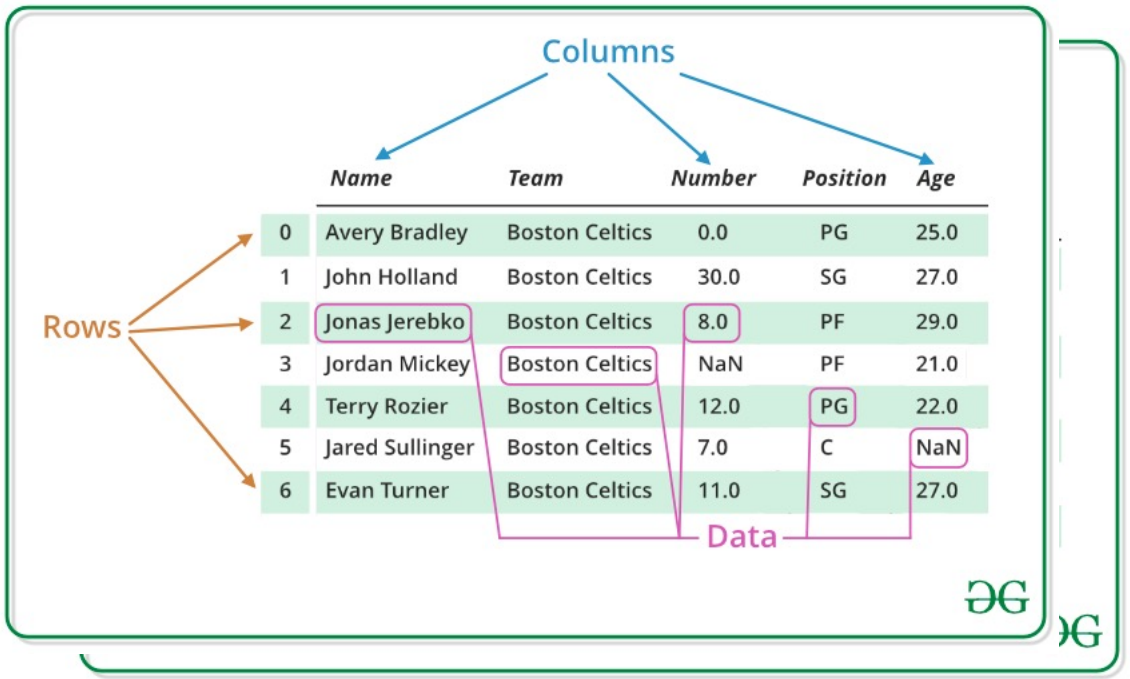
Python for Data Science

- **Open source, Interpreted (executes what you type)**
- **Libraries for:**
  - Mathematics & Statistics (Scipy)
  - Multidimensional arrays (Numpy)
  - Data frames (Pandas)
  - Visualization (Matplotlib, Seaborn)
  - Machine Learning (sklearn)
  - Deep Learning (Keras, TensorFlow, PyTorch)



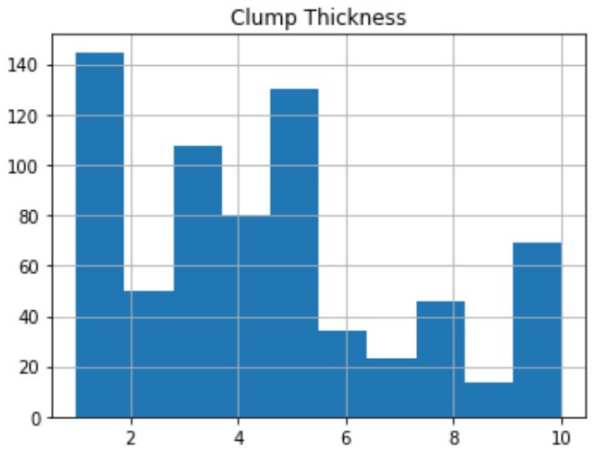
Data Ingestion

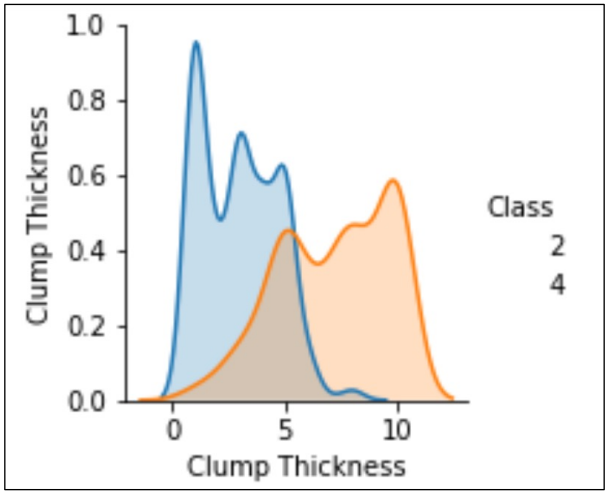
- Data sources: own experiments, online portal, data repositories
- Original data might need to be processed/cleaned. Generate gene expression count, remove artifact from processing tools, etc.
- For example, tabular data is in a form of: Samples * features, then it can be loaded in memory as **dataframe** (e.g. using Pandas)

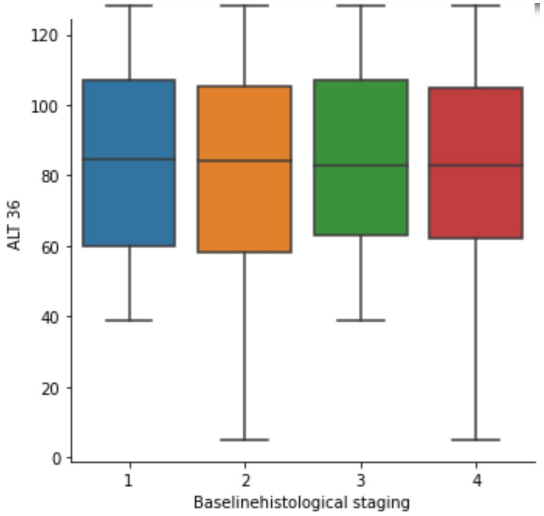


Univariate plots

- These are methods to show the distribution of a single variable.
- Popular methods are histograms, dot plots, box plots, and kernel density plots: **`df.hist`**, **`seaborn.pairplot`**, **`seaborn.catplot`**



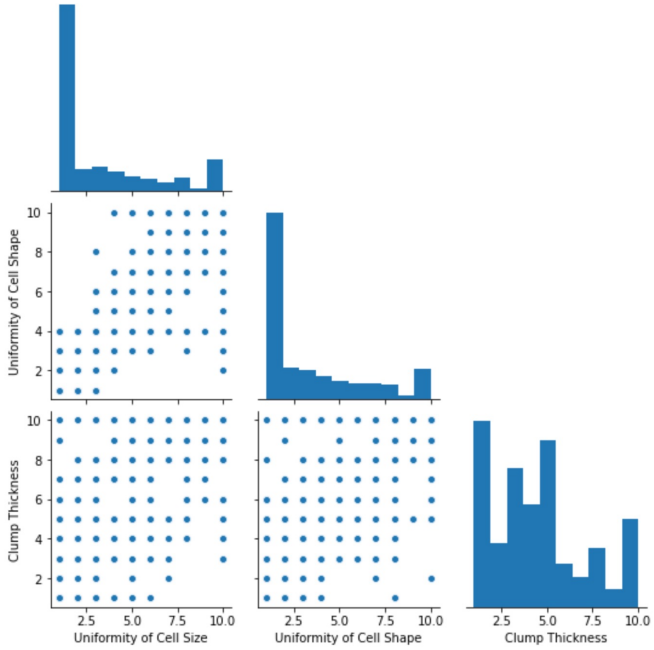


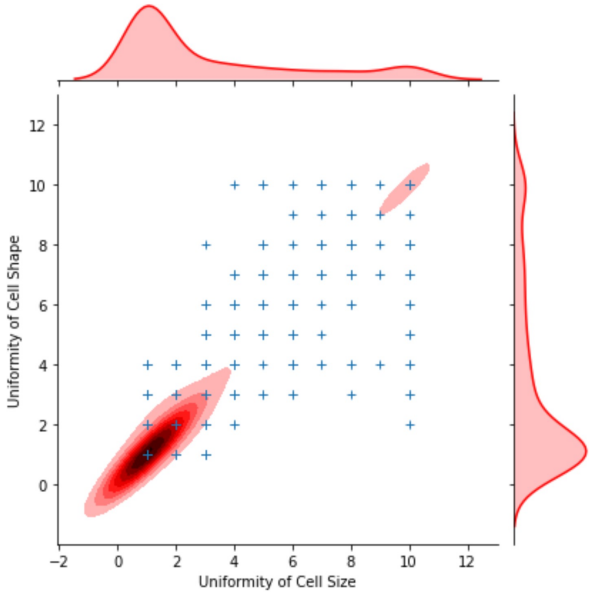


- These plots helps in understanding the assumptions in a model (e.g., normal probability plot) and check the limitations where a model may not fit well the data.

Bivariate plots

- Scatter plots can highlight the relationship between two variables and possible trends. **`seaborn.jointplot, seaborn.pairplot`**
- The components of the trend are: (a) ***direction*** (positive or negative), (b) ***form*** (linear or curvilinear), and (c) ***strength*** (degree of variability around the trend). `

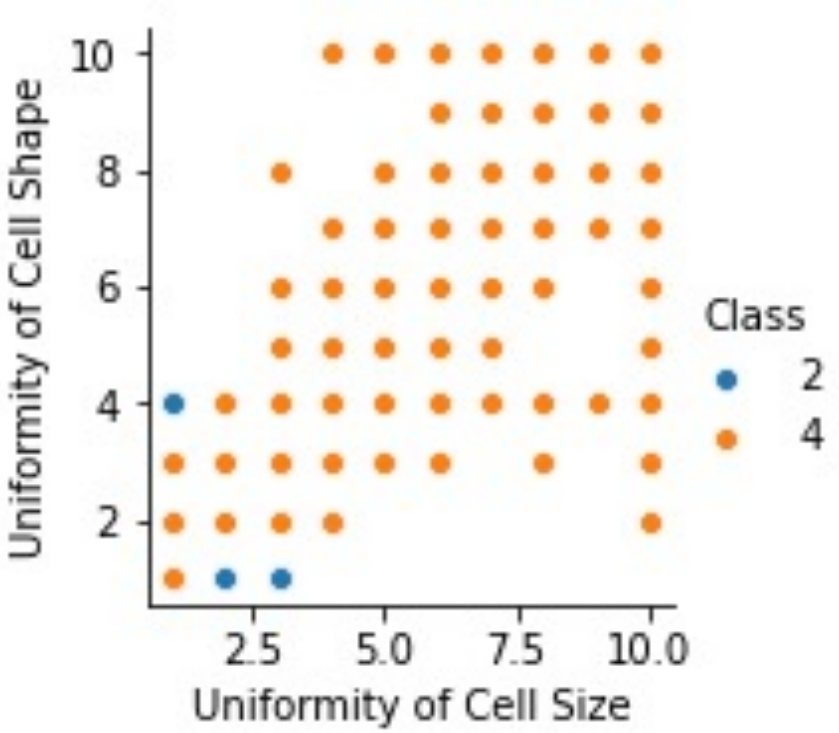




- Existence of ***clusters*** can also be identified in a scatter plot.

Multivariate

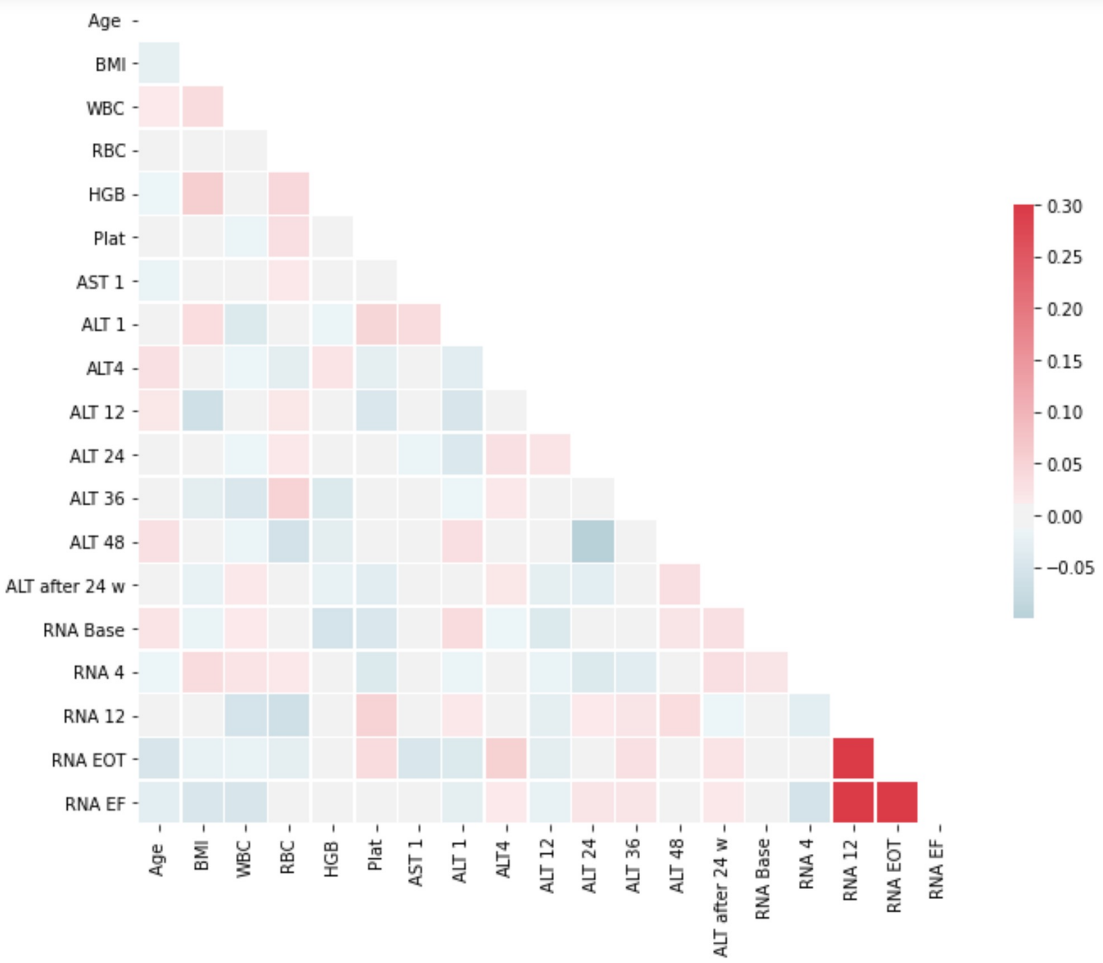
- Exponential number of plots: 3 variables: N^3, 4 variables: N^4
- To limit the number of plots, use insights from the bivariate plots and select few candidates for multivariate you will investigate.
- In Seaborn, we can use 2D plots + the semantics of **hue, size, and style** to add up to three more variables.



Correlation

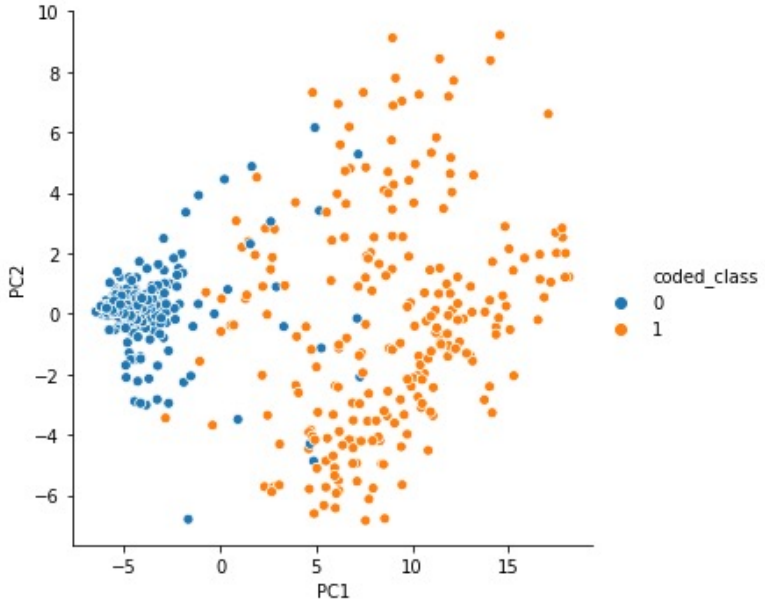
- One common step in feature selection is to remove correlated variables.

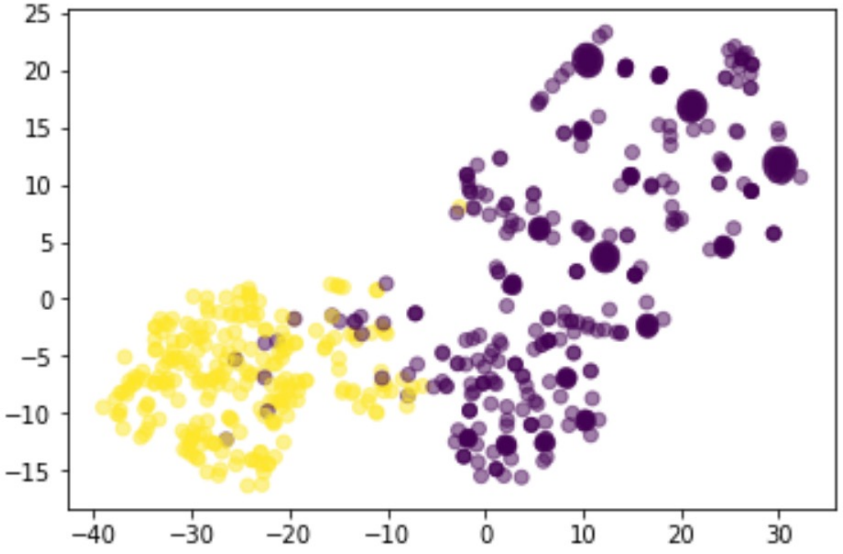
- Correlation can be computed using **`df.corr`** and visualized using **`seaborn.diverging_palette`**



Dimensionality Reduction

- **To visualize raw numeric data on a 2D plot, we need to reduce the dimensions.**
- **Two popular techniques: Principal Component Analysis (PCA) (linear), and TSNE (non linear)**





`sklearn.decomposition.PCA`

`sklearn.manifold.TSNE`

Image Visualization

Frederick National Laboratory for Cancer Research