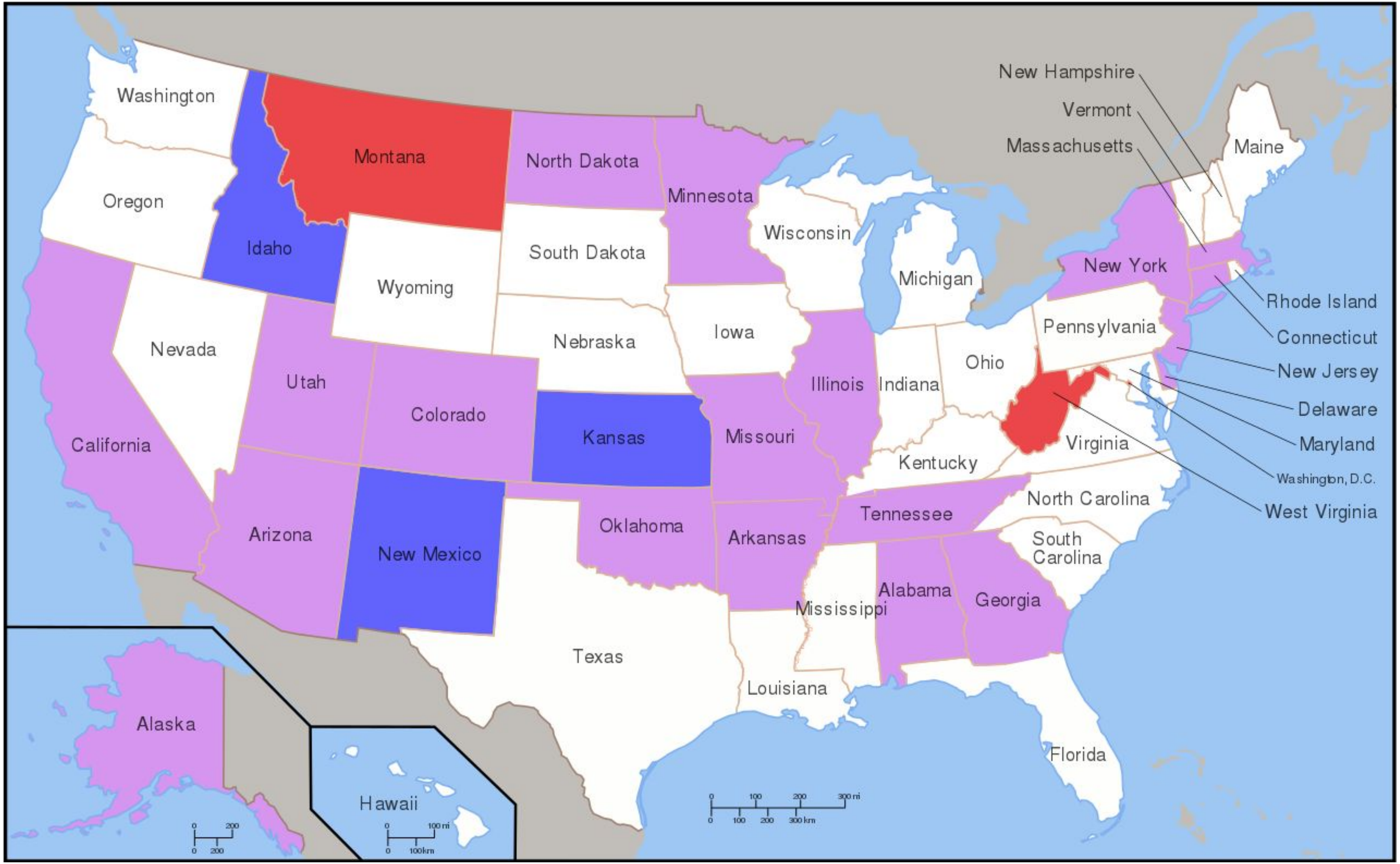
CSCI 285
Scientific Computing



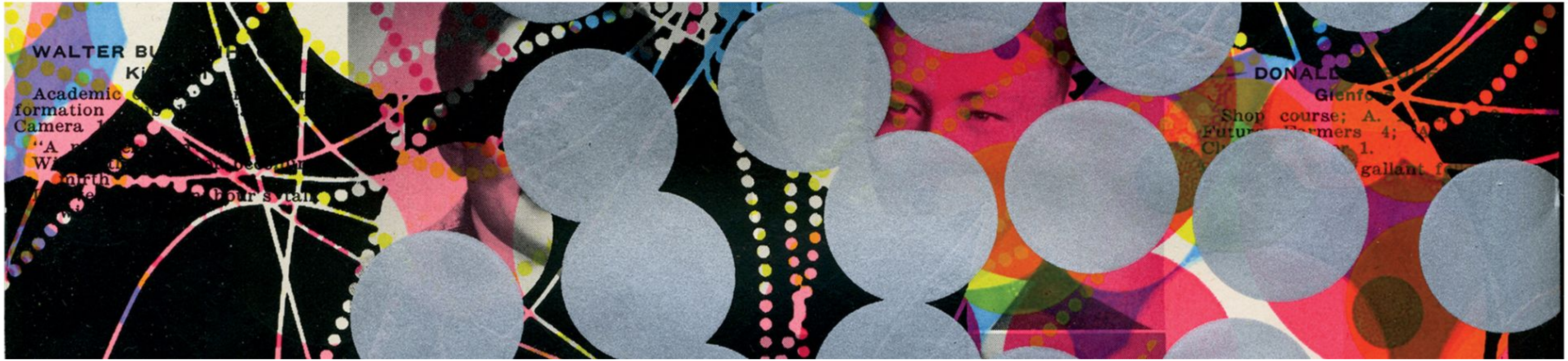
Washington
Montana
North Dakota
Minnesota
Oregon
Idaho
Wyoming
South Dakota
Wisconsin
Michigan
Nevada
Utah
Colorado
Nebraska
Iowa
Illinois
Indiana
Ohio
California
Kansas
Missouri
Kentucky
Arizona
New Mexico
Oklahoma
Arkansas
Tennessee
Texas
Louisiana
Mississippi
Alabama
Georgia
Florida
Alaska
Hawaii
New Hampshire
Vermont
Massachusetts
Maine
New York
Pennsylvania
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
Washington, D.C.
West Virginia
Virginia
North Carolina
South Carolina
0 100 200 300 mi
0 100 200 300 km
0 200
0 200
0 100 mi
0 100 km

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)



Why did you sign up?

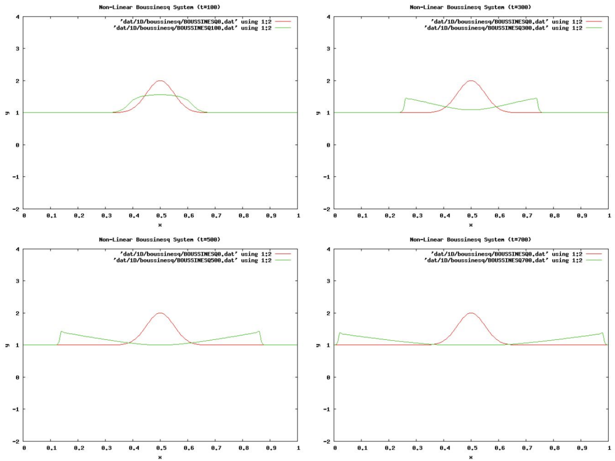
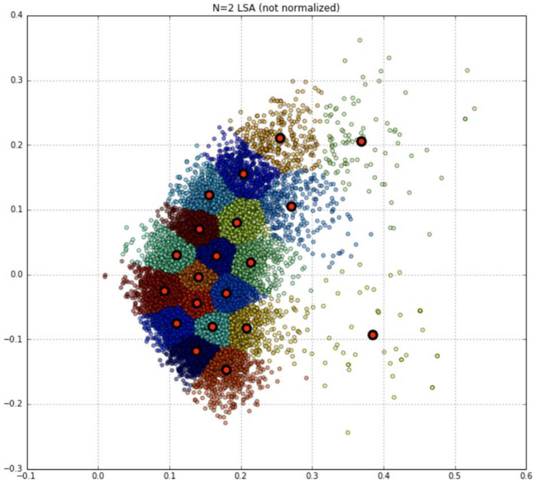
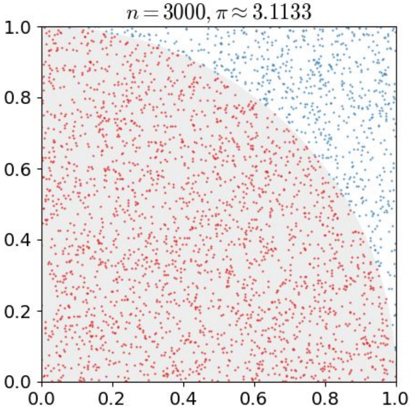
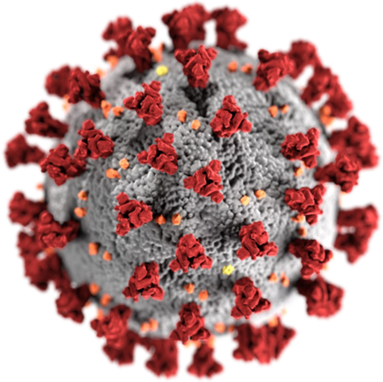


Figure 4: Initial gaussian wave with zero velocity. $\mu = 0.001$, $\varepsilon = 1$, L = 1, N = 256. cfl# = 0.1, $\sigma = 0.05$. Nonlinear shallow water.







CSCI 285 Learning Goals

Module #1: Data Analysis

- Analyze & visualize data sets from a variety of sources.
- Learn several analysis techniques including clustering and regression.

Module #2: Modeling

- Model and solve system dynamics problems.
- Construct a Monte-Carlo simulation model.
- Develop agent-based models for complex simulations.

Module #3: Numerical Techniques

- Approximate the roots of continuous functions.
- Understand the strengths and limitations of numerical techniques.

Write idiomatic python and use scientific python libraries.

CSCI 285 Course Overview

https://hendrix-cs.github.io/csci285/index.html

Policies

- Attendance
- Check ins / Office Hours
- Late Work

Coursework / In-class

- Lecture (36%)
- Labs (27%)
- Exams / Module Review (20%)
- Final Project (17%)

More Info

- Course Calendar / Class Notes / Project Timeline
- W2 Requirement
- Grading scale
- Prerequisites: MATH 130 & CSCI 150
- Teams - comms / submitting assignments

Commitments

- Active Participation
- Constructive Feedback
- Academic Integrity
- Learning Accommodation
- Physical & Mental Health

CSCI 285 Grading Scale

CSCI 285 Development Environment

1. Visit https://www.anaconda.com/

2. Download the open source distribution.

3. Follow the Anaconda3 installer instructions.

4. Launch Anaconda-Navigator (Mac, Windows, Linux)

5. Create new environments, launch processes, surf learning resources, etc.

(Alternatively, check out miniconda if you prefer a more lightweight approach)

Module #1
Data Analysis

import pandas as pd

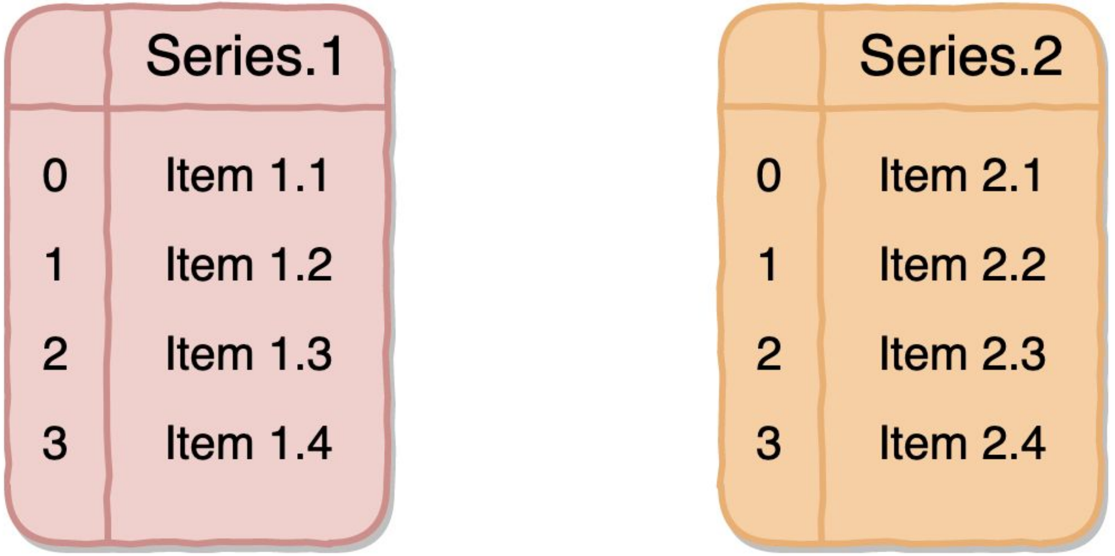
pandas

pandas is an open-source python library built for data manipulation and analysis. It is part of the standard library for many teams of data scientists and engineers. pandas introduces new types that have special syntax for data manipulation that are not shared with python's builtin types (e.g. list, dict). Some of the new syntax can look jarring at first, but is *lingua franca* for many data researchers.

Getting Started with pandas

- https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html
- https://chrisalbon.com/
- https://www.datacamp.com/courses/data-manipulation-with-pandas

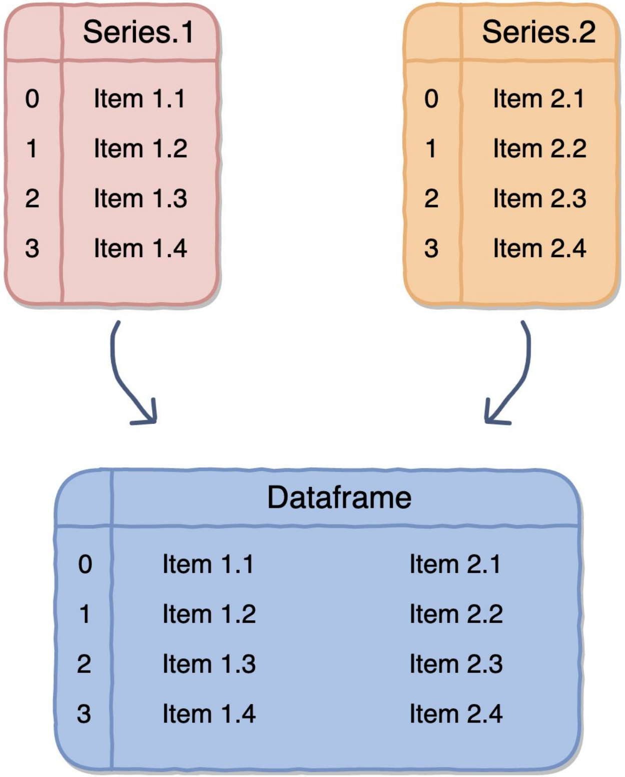
pd.Series



	Series.1
0	Item 1.1
1	Item 1.2
2	Item 1.3
3	Item 1.4

	Series.2
0	Item 2.1
1	Item 2.2
2	Item 2.3
3	Item 2.4

pd.DataFrame



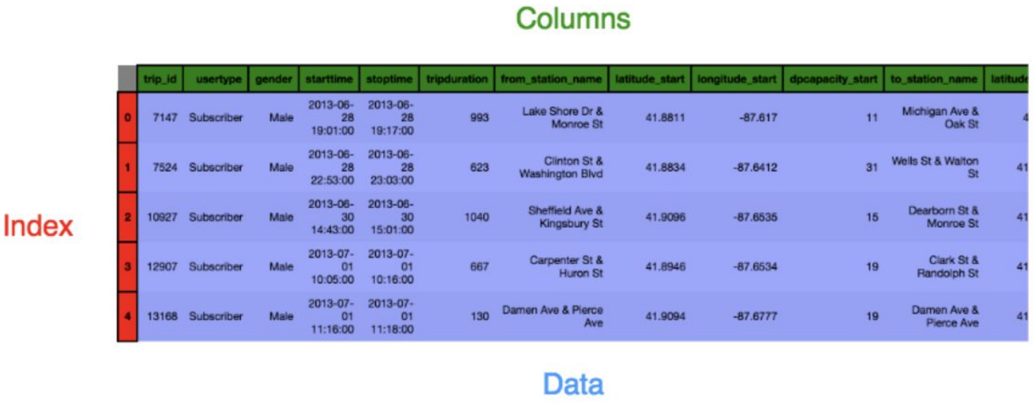
	Series.1
0	Item 1.1
1	Item 1.2
2	Item 1.3
3	Item 1.4

	Series.2
0	Item 2.1
1	Item 2.2
2	Item 2.3
3	Item 2.4

	Dataframe	
0	Item 1.1	Item 2.1
1	Item 1.2	Item 2.2
2	Item 1.3	Item 2.3
3	Item 1.4	Item 2.4

Components of a DataFrame

The Columns, Index, and Data



	trip_id	usertype	gender	starttime	stoptime	tripduration	from_station_name	latitude_start	longitude_start	dpcapacity_start	to_station_name	latitud
0	7147	Subscriber	Male	2013-06-28 19:01:00	2013-06-28 19:17:00	993	Lake Shore Dr & Monroe St	41.8811	-87.617	11	Michigan Ave & Oak St	4
1	7524	Subscriber	Male	2013-06-28 22:53:00	2013-06-28 23:03:00	623	Clinton St & Washington Blvd	41.8834	-87.6412	31	Wells St & Walton St	41
2	10927	Subscriber	Male	2013-06-30 14:43:00	2013-06-30 15:01:00	1040	Sheffield Ave & Kingsbury St	41.9096	-87.6535	15	Dearborn St & Monroe St	41
3	12907	Subscriber	Male	2013-07-01 10:05:00	2013-07-01 10:16:00	667	Carpenter St & Huron St	41.8946	-87.6534	19	Clark St & Randolph St	41
4	13168	Subscriber	Male	2013-07-01 11:16:00	2013-07-01 11:18:00	130	Damen Ave & Pierce Ave	41.9094	-87.6777	19	Damen Ave & Pierce Ave	41

Description

- Columns - label each column
- Index - label each row
- Data - actual values in DataFrame

Alternative Names

- Columns - column names/labels, column index
- Index - index names/labels, row names/labels
- Data - values

Axis Number

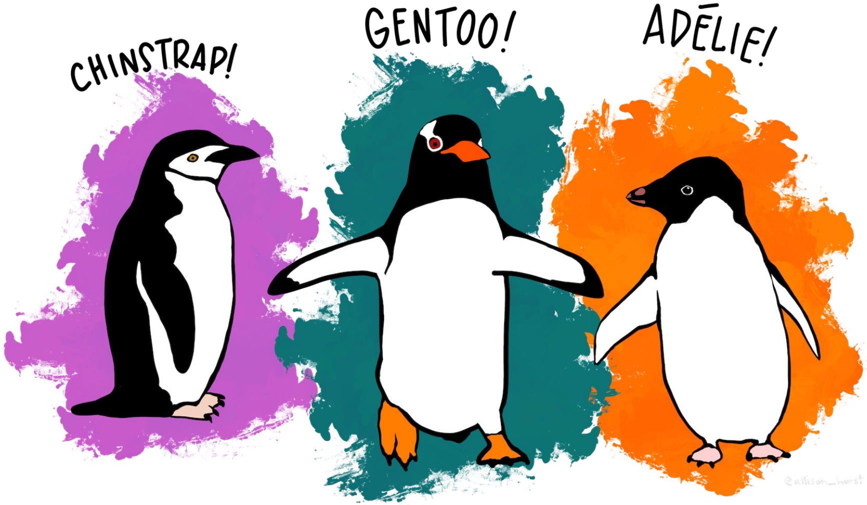
- Columns: 1
- Index: 0

Application: Pandas Intro

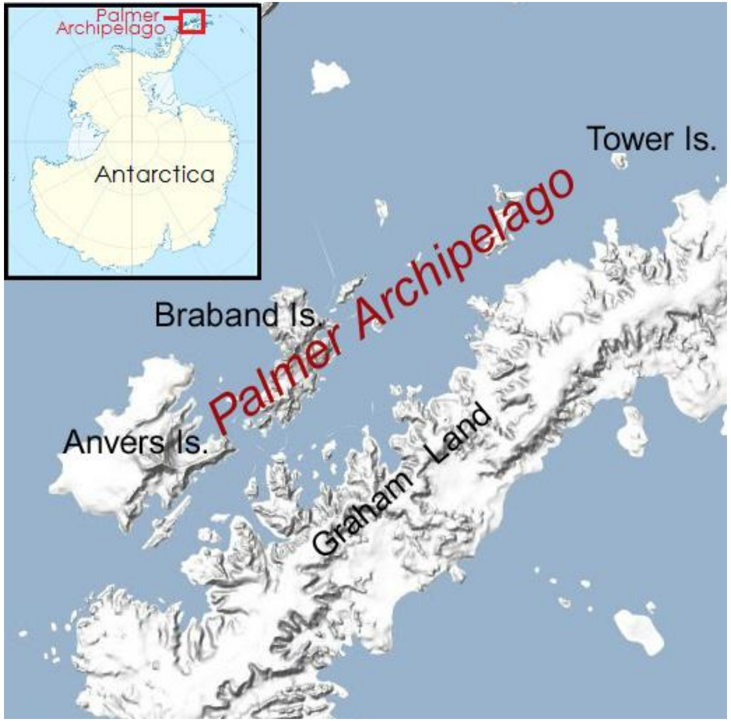
pandas

Application: Palmer Penguins

pandas



Artwork by @allison_horst”.



Lab #1: Lake Trout

Module #1
Data Visualization

Lab Report Format

- Palmer Penguins notebook (reference)
- Professor Wilson’s trout lab (reference, next week)
- Mixture of Markdown, Code, and Figures
- Submitted via Teams (zip file)
- Must Include
    a. Any input data (data used to produce the report)
    b. Any output data (e.g. CSV files)
    c. Notebook with relative paths to load the data

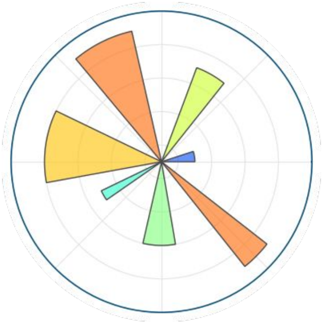
Any Questions about Lab #1?

import seaborn as sns

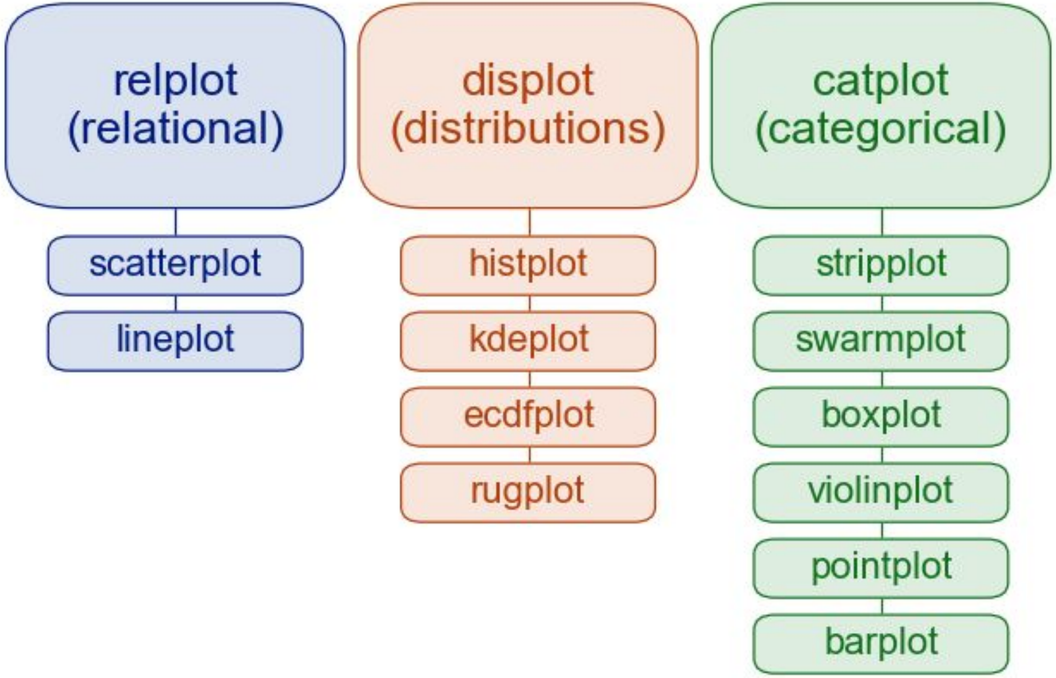
Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics

Getting Started with seaborn

- https://seaborn.pydata.org/introduction.html
- https://seaborn.pydata.org/tutorial/function_overview.html
- https://chrisalbon.com/
- https://www.datacamp.com/courses/intermediate-data-visualization-with-seaborn



import seaborn as sns



relplot (relational)

scatterplot

lineplot

displot (distributions)

histplot

kdeplot

ecdfplot

rugplot

catplot (categorical)

stripplot

swarmplot

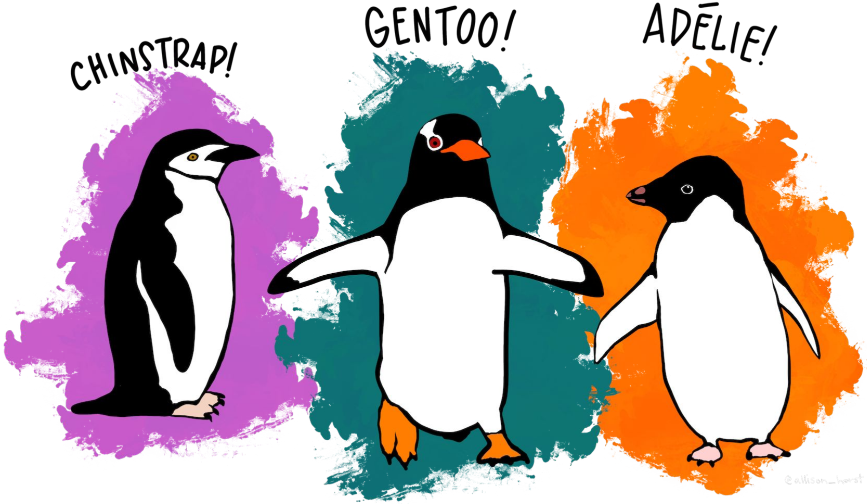
boxplot

violinplot

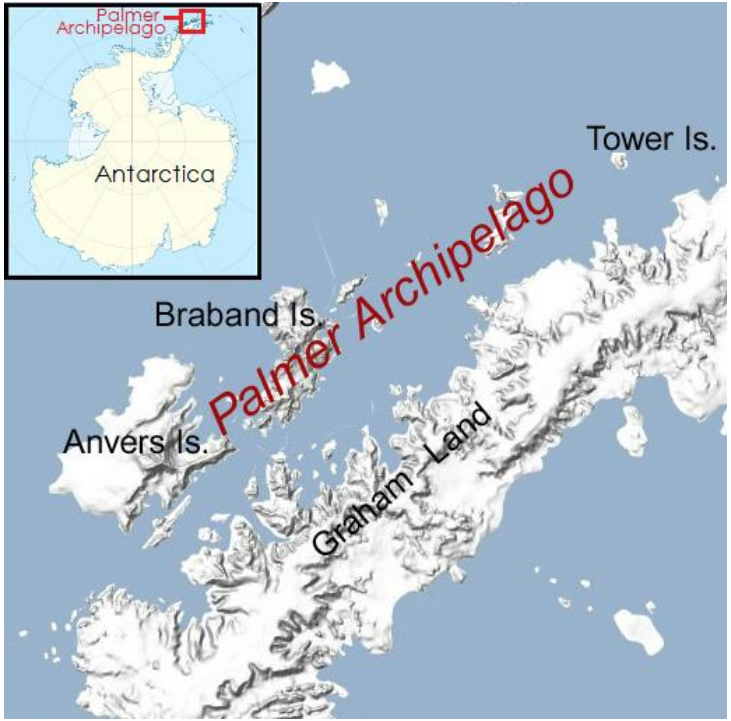
pointplot

barplot

Application: Palmer Penguins



Artwork by @allison_horst”.



Final Project

- Project Description
- COMAP
- Markdown (https://dillinger.io/)
- LaTeX / Overleaf

Lab #2: Data Visualization

seaborn

- Due Date: 9/12 (midnight).
- FEV notebook
- Linear Regression

Module #1
Machine Learning

What is Machine Learning?

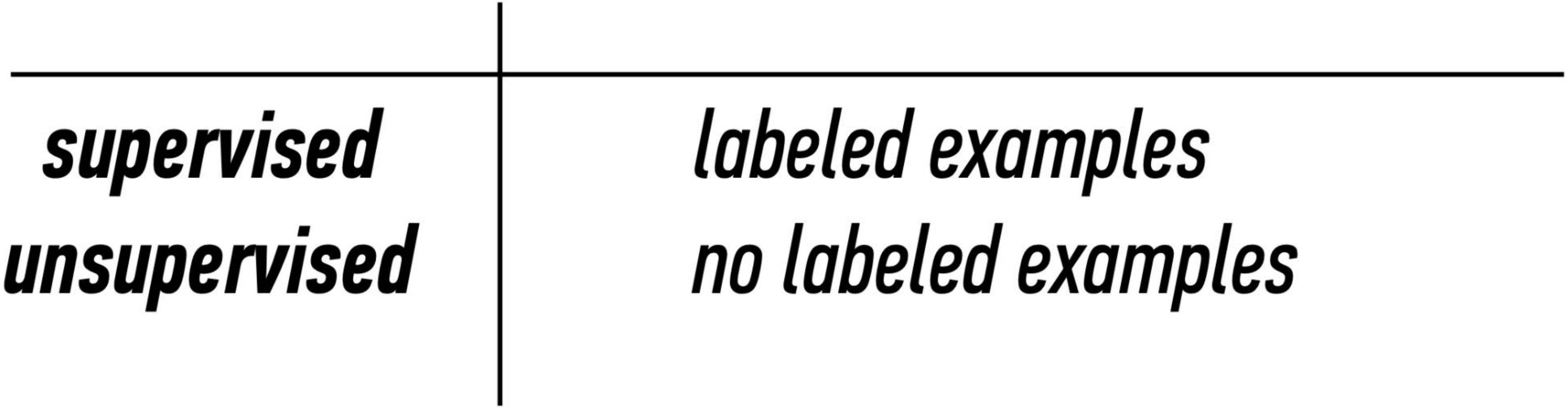
From Wikipedia:

Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data…

The core of machine learning deals with ***representation*** *and* ***generalization***

- Representation == extracting structure from data
- Generalization == making predictions from data

Types of Learning Problems

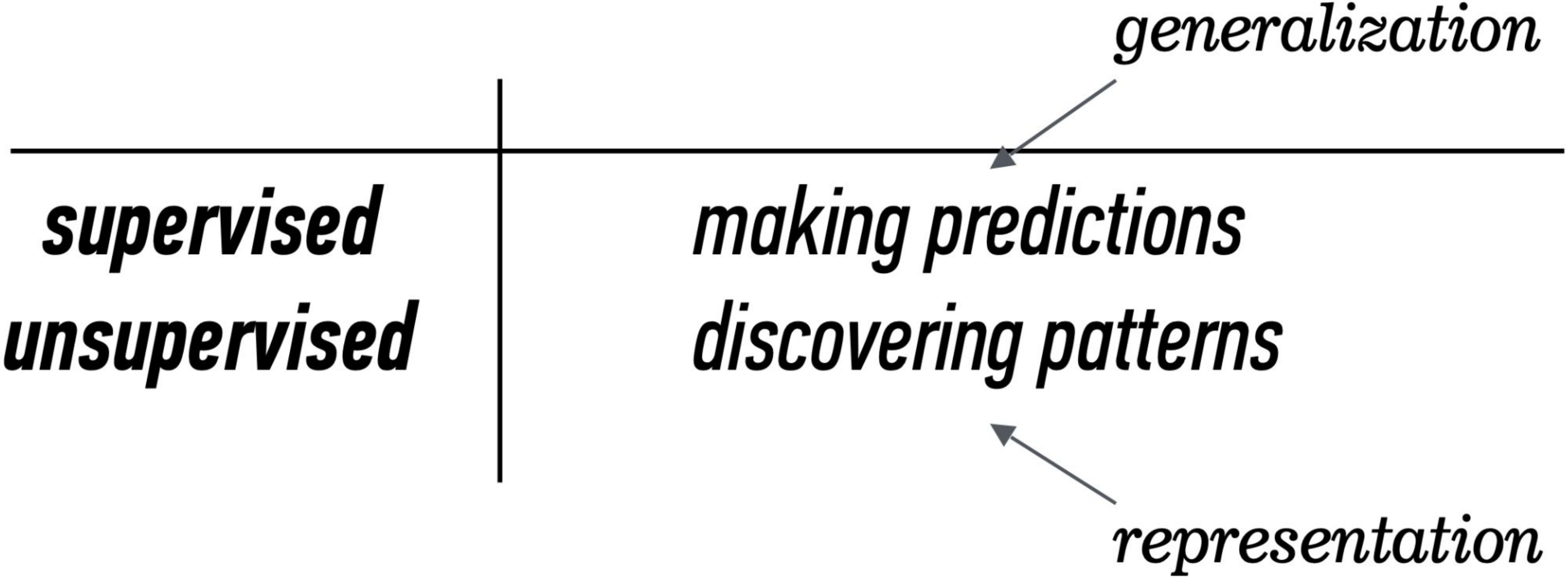


supervised	*labeled examples*
unsupervised	*no labeled examples*

Types of Learning Problems

supervised	making predictions
unsupervised	discovering patterns

Types of Learning Problems



generalization

supervised

unsupervised

making predictions

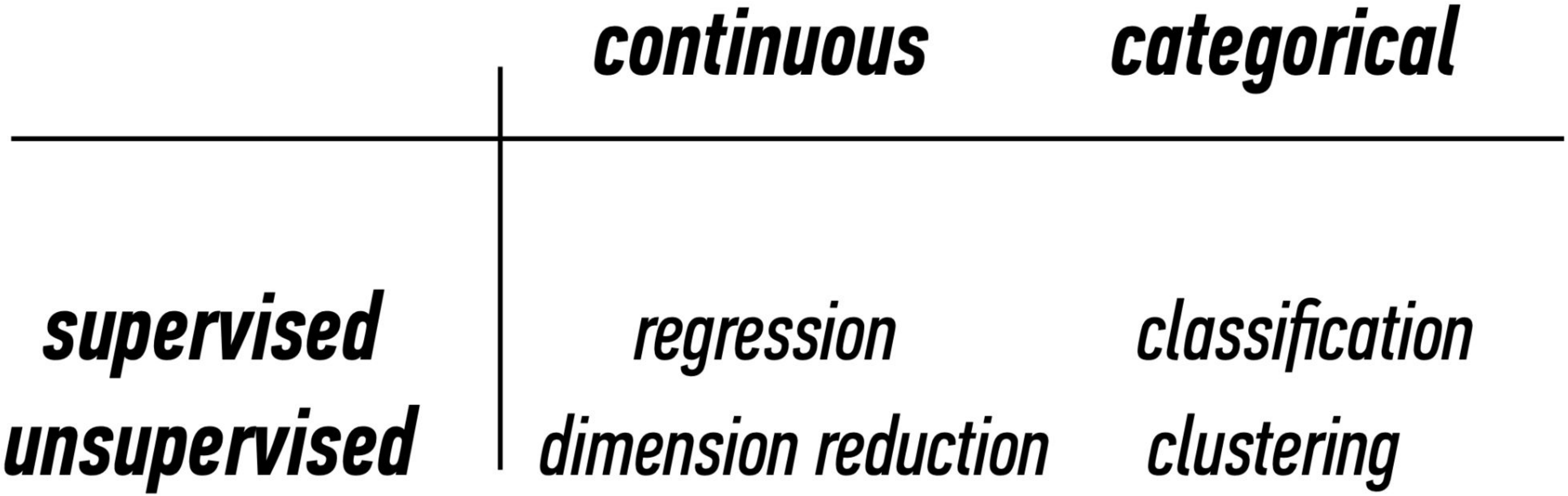
discovering patterns

representation

Types of Data

	continuous	categorical
	quantitative age, salary, height, etc.	qualitative city, yes/no, vote, etc.

Types of ML



	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

from sklearn import

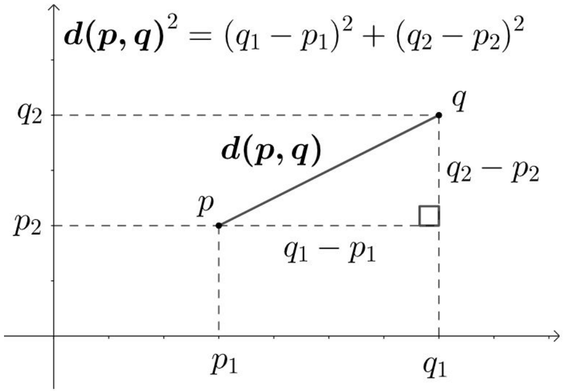
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

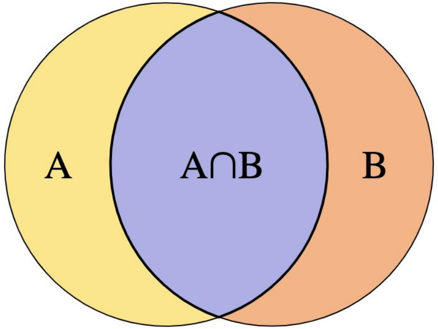
Getting Started with scikit-learn

- https://scikit-learn.org/stable/index.html
- https://www.datacamp.com/courses/machine-learning-with-scikit-learn
- https://chrisalbon.com/

Distance Measures (sklearn.metrics)

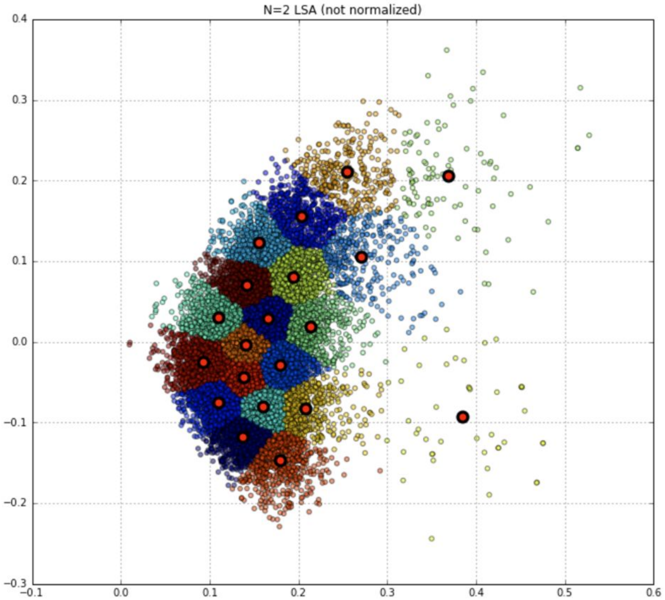
- Euclidean Space
  a. L1 Dist (manhattan)
  b. L2 Dist (pythagorean)
  c. LR Dist (general formulation)
- Non-Euclidean Space
  a. Jaccard Dist (sets)
  b. Edit Dist (strings)


$\boldsymbol{d}(\boldsymbol{p},\boldsymbol{q})^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2$
q
q_2
$\boldsymbol{d}(\boldsymbol{p},\boldsymbol{q})$
$q_2 - p_2$
p
p_2
$q_1 - p_1$
p_1
q_1



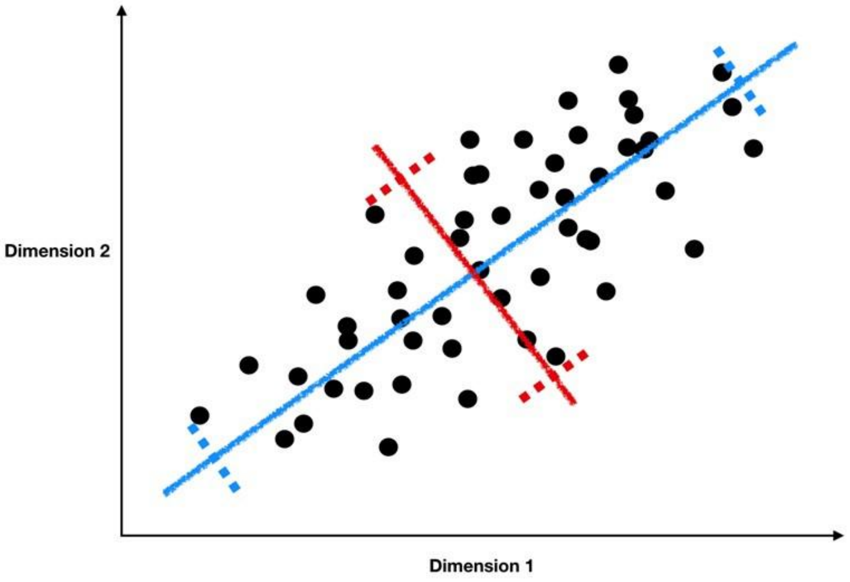
Clustering (sklearn.cluster)

- K-Means Algorithm (step-by-step)
- scikit-learn K-Means module
- `from sklearn.datasets import make_blobs`
- Picking the right value for *k*
- Clustering Palmer Penguins
- Comparing predictions to ground truth
  - Confusion matrix
  - Seaborn heatmaps

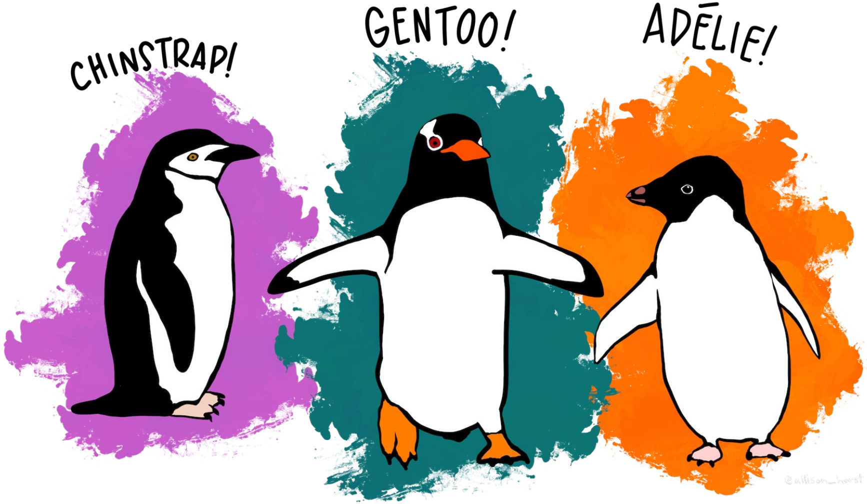


Decomposition & Preprocessing

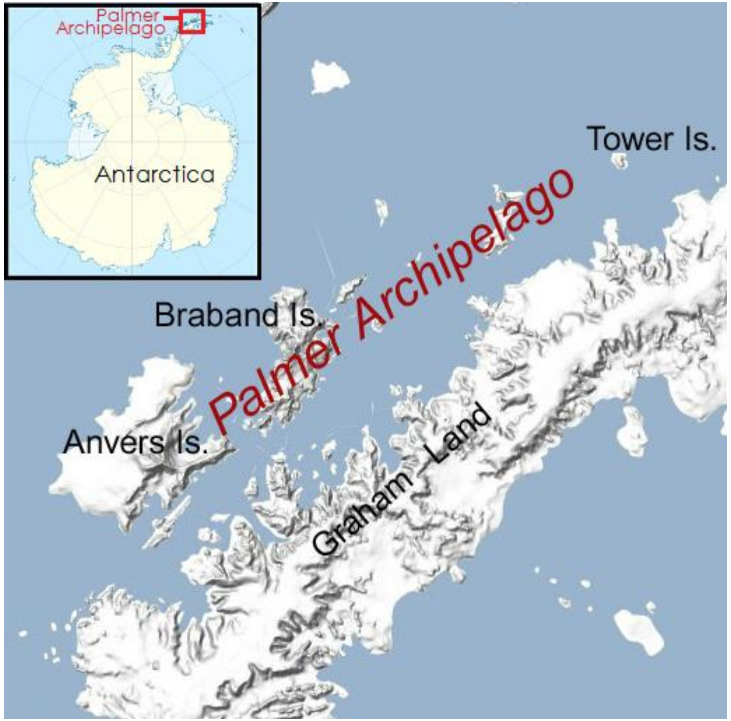
- Feature Scaling
  - StandardScaler
- Curse of dimensionality
- PCA (principal component analysis)



Application: Palmer Penguins



Artwork by @allison_horst”.

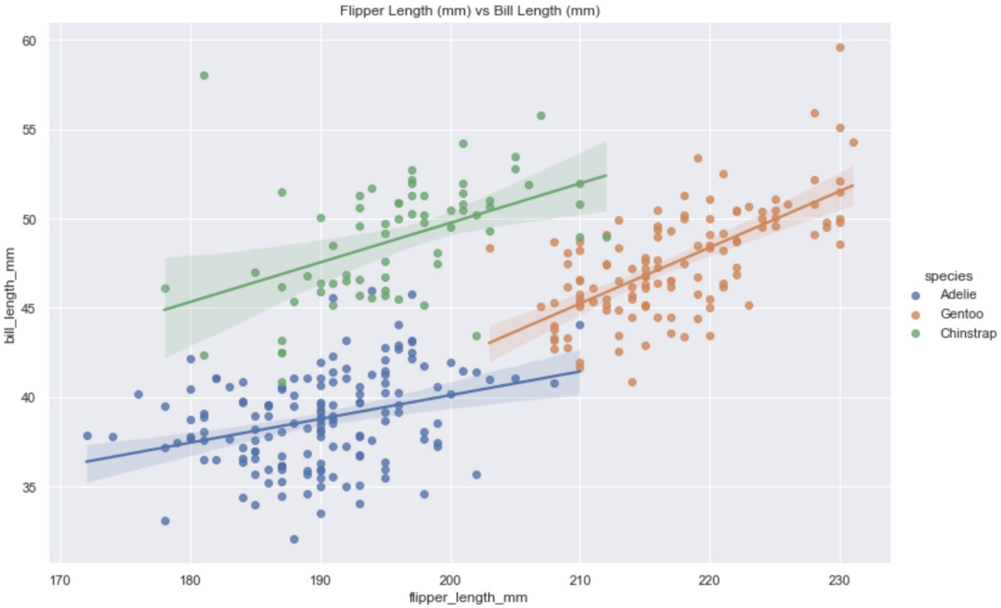


from sklearn.datasets

- sklearn.datasets

from sklearn.linear_model

- sklearn.linear_model



Module #1
Choropleths & More Data Viz

Module #1 Timeline

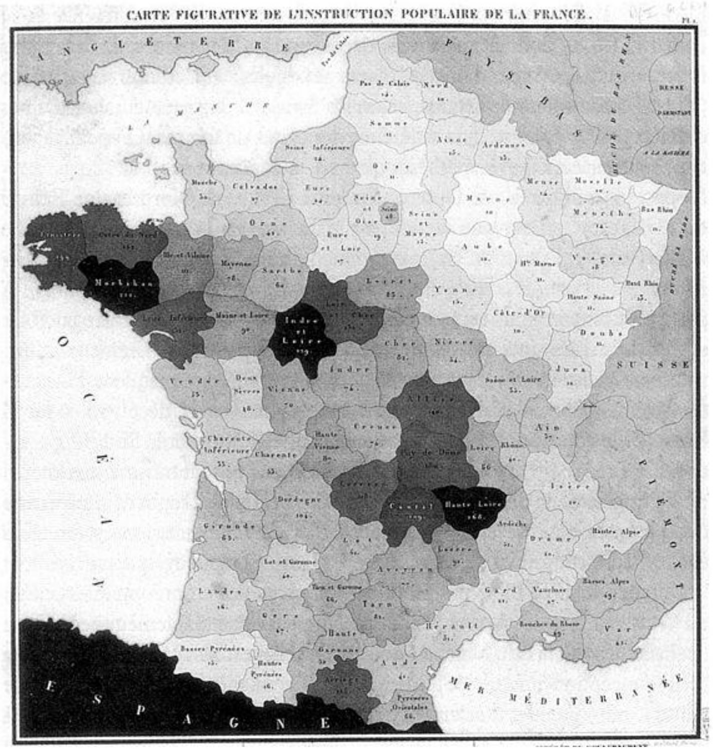
- More Data Visualizations (today!)

- Module #1 Review / Exam Prep (9/20)
- Class Cancelled (9/22)

- Exam #1 (9/27)
- Lab #4 (9/29)

- System Dynamics Modeling (10/4)

Choropleths



import geopandas as gpd

GeoPandas, as the name suggests, extends the popular data science library pandas by adding support for geospatial data.

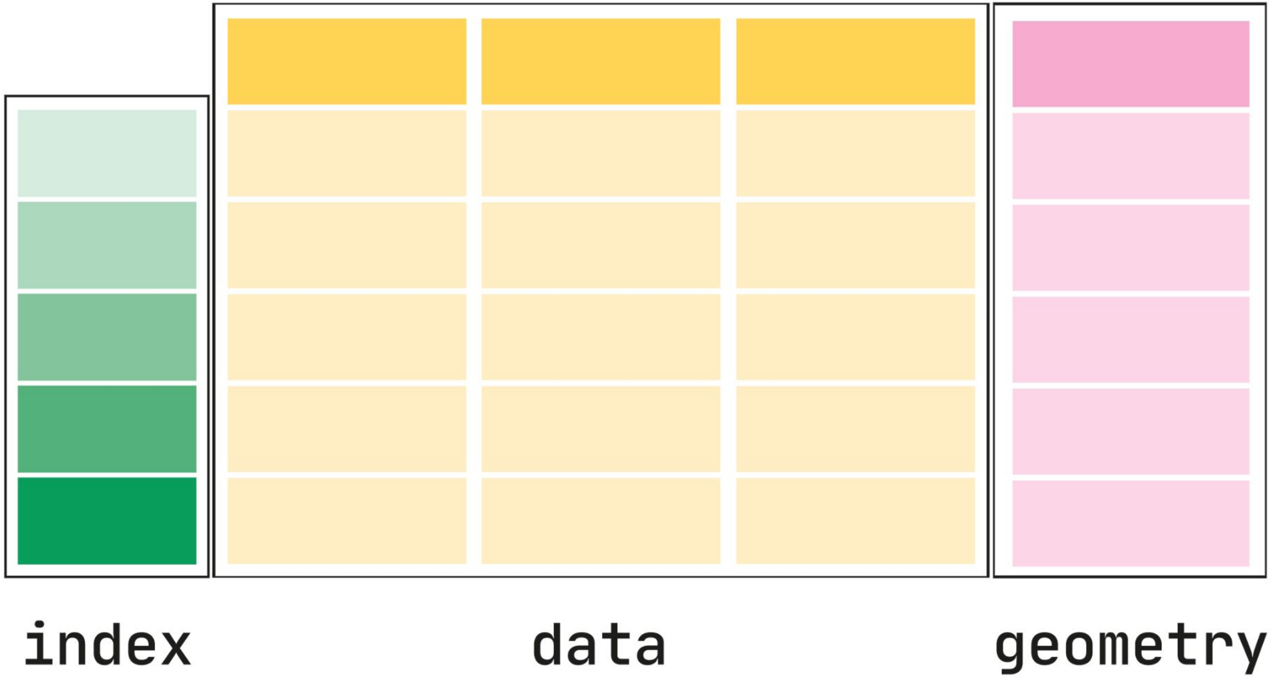
- pandas.DataFrame → geopandas.GeoDataFrame
- pandas.Series → geopandas.GeoSeries

Getting Started with geopandas

- https://geopandas.org/en/stable/getting_started/introduction.html
- https://www.kaggle.com/

import geopandas as gpd

GeoPandas



Barcharts & Heatmaps

seaborn

Exam #1: 9/27 (Next Tuesday)

- A: **Complete** two **Exams** and at least **Partially Complete** the remaining **Exam**.
- B: **Complete** one **Exam** and **Partially Complete** the remaining **Exams**.
- C: **Partially Complete** all three **Exams**.
- D: **Partially Complete** two **Exams**.

1. Tuesday, 9/27 Exam #1 will be “opened”
2. Due by beginning of class on Thursday (9/29)

CSCI 285 Learning Goals

Module #1: Data Analysis

- Analyze & visualize data sets from a variety of sources.
- Learn several analysis techniques including clustering and regression.

Module #2: Modeling

- Model and solve system dynamics problems.
- Construct a Monte-Carlo simulation model.
- Develop agent-based models for complex simulations.

Module #3: Numerical Techniques

- Approximate the roots of continuous functions.
- Understand the strengths and limitations of numerical techniques.

Write idiomatic python and use scientific python libraries.

Module #2
System Dynamics