



Python-Powered Machine Learning

Carrie Gardner, PTOH '23

Special Thanks:
Data Analytics Club



A Little Background on Me

Carrie Gardner

- Member of the Technical Staff @ AI Division, Software Engineering Institute, CMU
 - Cultivating the interdisciplinary engineering discipline of AI Engineering ☺
- Social Science & Information Science Education
- 7+ years using python and working on analytics projects
- Proud Python Evangelist 😊

Carnegie Mellon University

Tepper School of Business

William Larimer Mellon, Founder





Workshop Roadmap

- Logistics
- What is Machine Learning?
- How is ML Performed?
- Why Python?
- Interactive Tutorial
 - 1. Tooling & Data Wrangling
 - 2. Price Prediction w/Scikit-Learn
 - 3. Natural Language Processing

Timing:

~30 Minutes Lecture

~5-10 Minute Break

~70 Minutes

Interactive Coding

Logistics

- This workshop assumes
 - Attendees have little-no experience with machine learning
 - Attendees have little-no experience using python
 - **Everyone** can learn to code (no CS background required!)
- **PLEASE raise your hand and ask questions!**
- Takeaways:
 - **Confidence** to start machine learning with python
 - **Knowledge** of ML and data analytics capabilities with python
 - **Scaffolding** to jump-start a python-powered analytics project

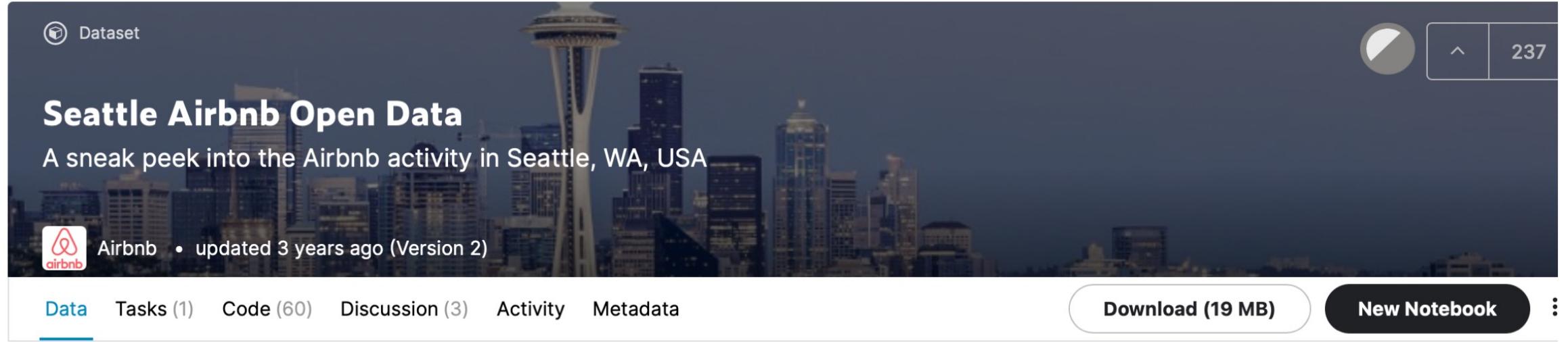




What is Machine Learning?



Let's Frame the Opportunity



A screenshot of a Kaggle dataset page for "Seattle Airbnb Open Data". The page features a dark background with a blurred image of the Seattle skyline at night. At the top left is a "Dataset" icon and at the top right are navigation and search controls. The title "Seattle Airbnb Open Data" is prominently displayed in white, bold text. Below it is a subtitle "A sneak peek into the Airbnb activity in Seattle, WA, USA". The Airbnb logo is visible on the left, followed by the text "Airbnb • updated 3 years ago (Version 2)". Below the title, there are tabs for "Data" (which is underlined in blue), "Tasks (1)", "Code (60)", "Discussion (3)", "Activity", and "Metadata". To the right of these tabs are buttons for "Download (19 MB)" and "New Notebook". A vertical ellipsis icon is also present. At the bottom of the page, there are sections for "Usability 7.1", "License CC0: Public Domain", and "Tags travel, hotels and accommodations, united states".

- Can we **forecast** and **predict** prices?
- Can we **parse** and **glean insights** from reviews?
- Can we **segment** listings into groups?

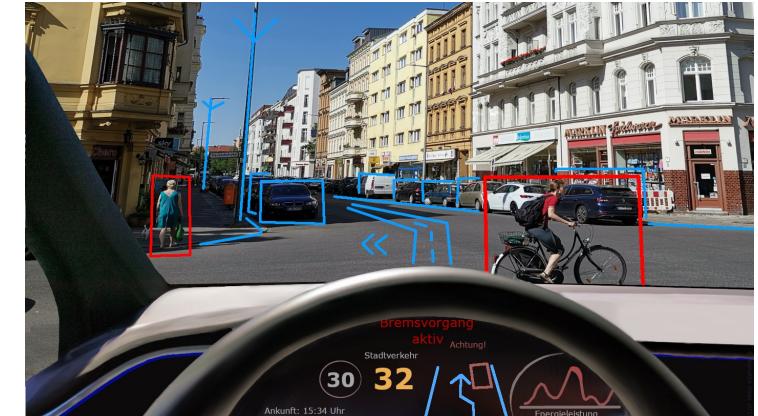
kaggle

What is Machine Learning?

Machine Learning (ML) is a data science technique that learns from existing data to forecast future behaviors, outcomes, and trends

Forms

- **Supervised**
- **Unsupervised**
- Semi-supervised
- Reinforcement/Active

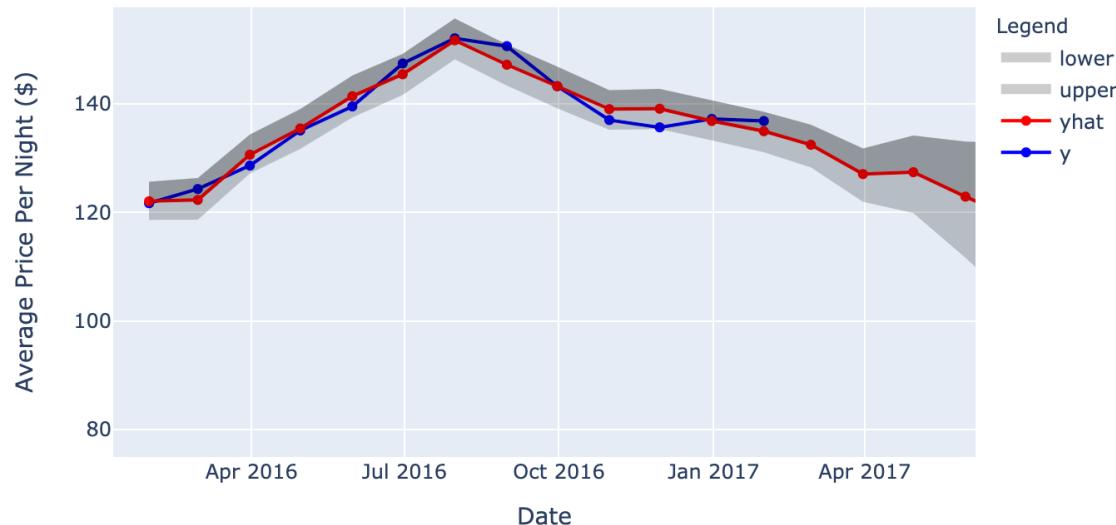


Supervised Learning

Supervised ML models are developed by training algorithms on *labeled* data to understand the relationship between the features and the label

- Classification
- **Regression**

Past, Present, and Predicted Future Monthly Average Prices



Forecast Daily Listing Prices

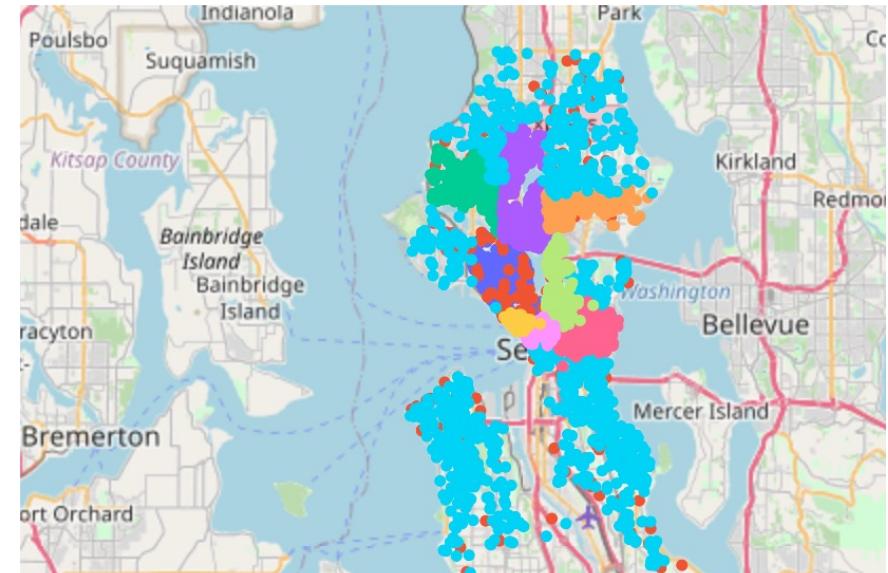
- Modeling using FB's Prophet Library
- Visualization using Plotly

Unsupervised Learning

Unsupervised ML models are developed by training algorithms on *unlabeled* data to find relationships amongst variables

- **Clustering**
- Association

K-Means Clustering of AirBnB Spaces



Cluster Listings into Neighborhoods

- Modeling using Scikit-Learn
- Visualization using Plotly

Remember:
a machine
learning model is
designed to
represent the
problem you are
studying, and it is
not a perfect
reflection of the
problem

*All models are wrong
but some are useful*



George E.P. Box



More Business Use Cases

Predictive Maintenance

Optimal Price Selection

Personalized Ads

Fraud Detection

Market Forecasting

Sentiment Analysis

Document Workflow Automation

Supplier Selection

Route Optimization

Credit Scoring

Algorithmic Trading

...



How is Machine Learning Performed?



ML Workflow: Simple

BLUF

- What is the goal?
- What's the expected operational context?
- What's the expected deployment state?
- What's our operational plan to fulfill this?

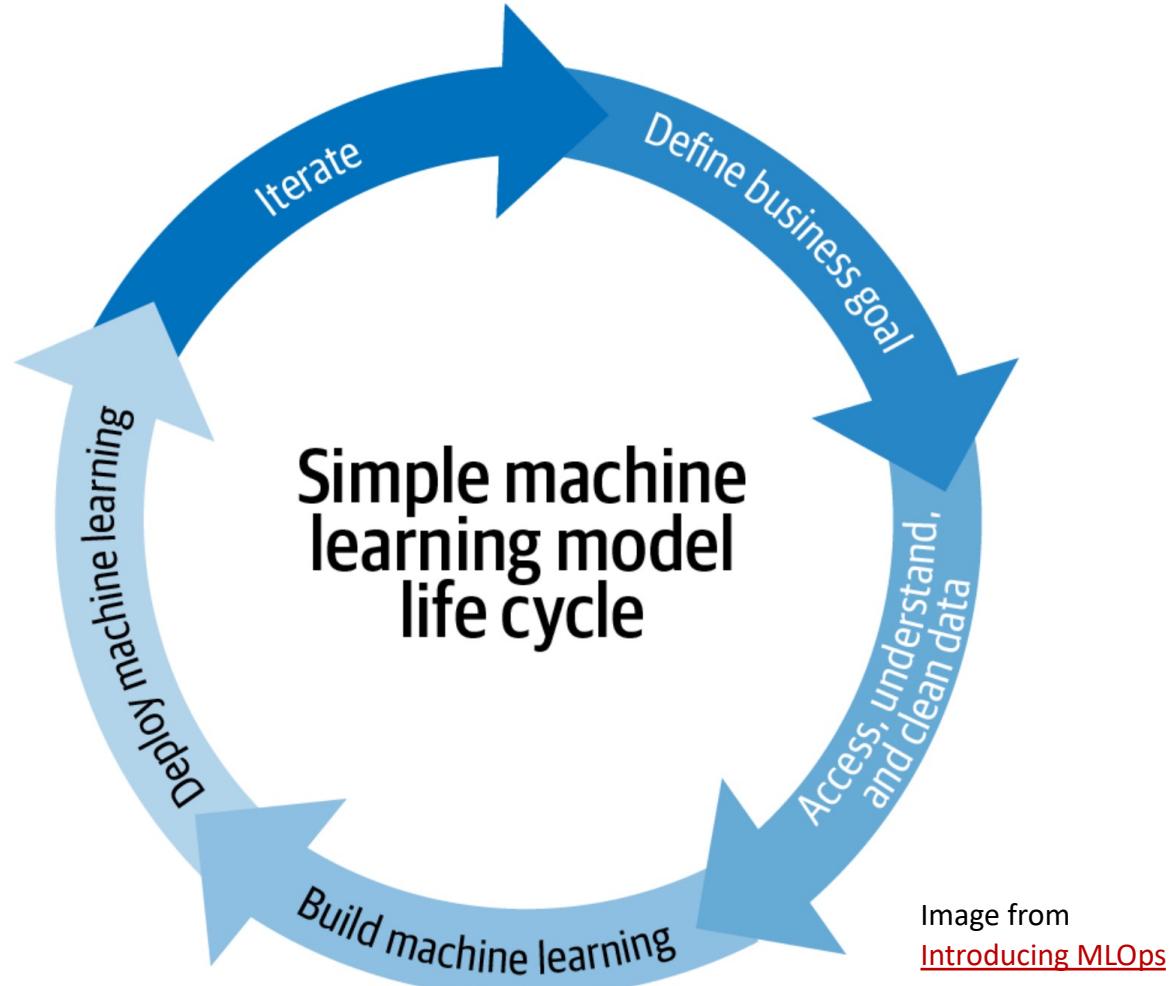


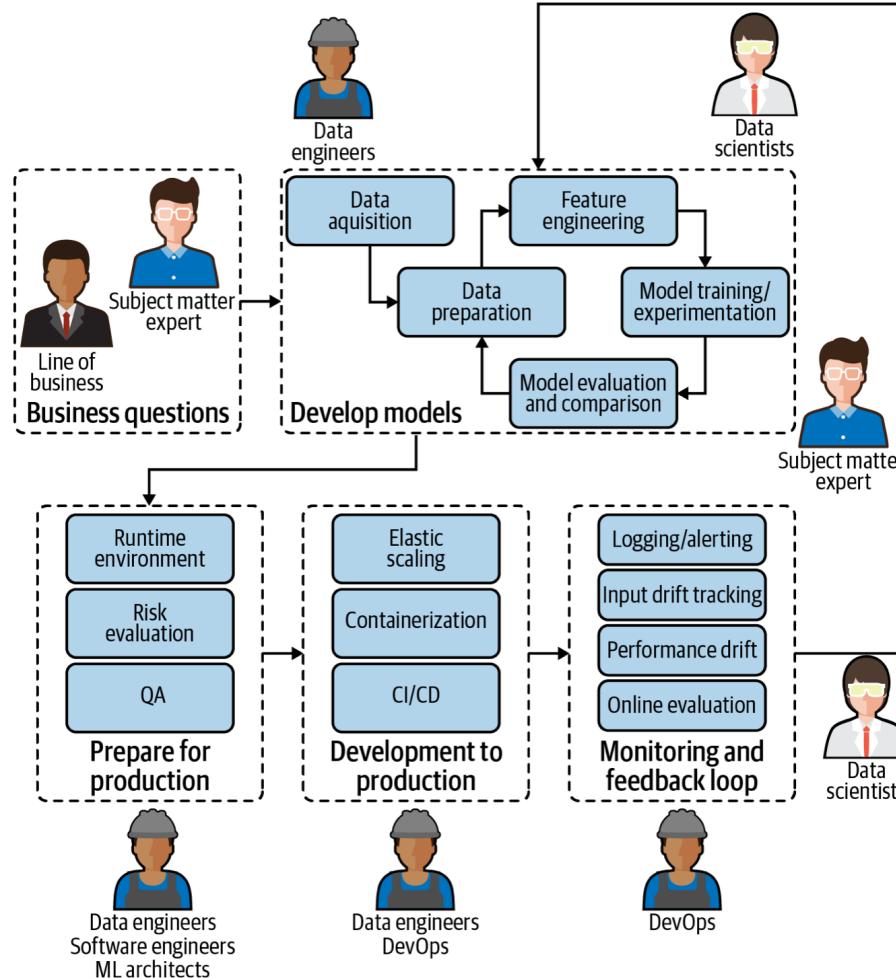
Image from
[Introducing MLOps](#)

Figure 1-2. A simple representation of the machine learning model life cycle, which often underplays the need for MLOps, compared to Figure 1-3

ML Workflow: Complex

Questions

- What are product integration needs?
- How might this need to scale?
- What's the post-deployment maintenance plan?



An “MLOps” perspective

Image from
[Introducing MLOps](#)

Figure 1-3. The realistic picture of a machine learning model life cycle inside an average organization today, which involves many different people with completely different skill sets and who are often using entirely different tools.

ML Workflow: Roles & Responsibilities



	Role	Responsibility
	Subject Matter Experts	Provide business use case and KPIs Provide expertise on data and inference context Manage ML project performance
ML Engineer	Data Scientists	Design and prototype models that address business use case Iteratively test modeling approaches (e.g., neural nets versus decision trees) Conduct exploratory data analysis (EDA) Identify ML risk concerns
	Data Engineers	Conduct exploratory data analysis (EDA) Perform Extend Transform Load (ETL) activities to prepare data Design and develop data “pipelines” and storage capabilities Deploy “dataOps” capabilities to manage state of data
	DevOps	Design and develop microservice architecture (e.g., using AWS services) Design and develop microservice containers Develop CI/CD infrastructure for automated deployment (e.g., automated logging and testing)
	Software Engineers	Translate and develop production software code from the prototype model Select and design software libraries

High-Level ML Workflow

BLUF

- What is the purpose?
- What is the end state goal?
- What are the requirements vs. the requests?

Problem Formation

- What is the goal?
- What are the requirements?

Data Curation

- Collect Data
- Preprocess

Exploratory Analysis

- Learn basic features
- Visualize and discern patterns

Modeling

- Select techniques per Problem Formation
- Train and fit ML pipeline model

Testing & Evaluation

- Iteratively test and evaluate ML model for appropriateness of fit / tune parameters

Integration

- Validate model meets specifications
- Deploy ML tool into production

Considerations

- Product Integration?
- Decision Support Tool or Fully Autonomous?
- Risk Tolerance Requirements ☺

Building Model

Train

Inference

It is often easiest to visualize a pipeline as a flowchart, such as the one depicted for Figure 2-5.

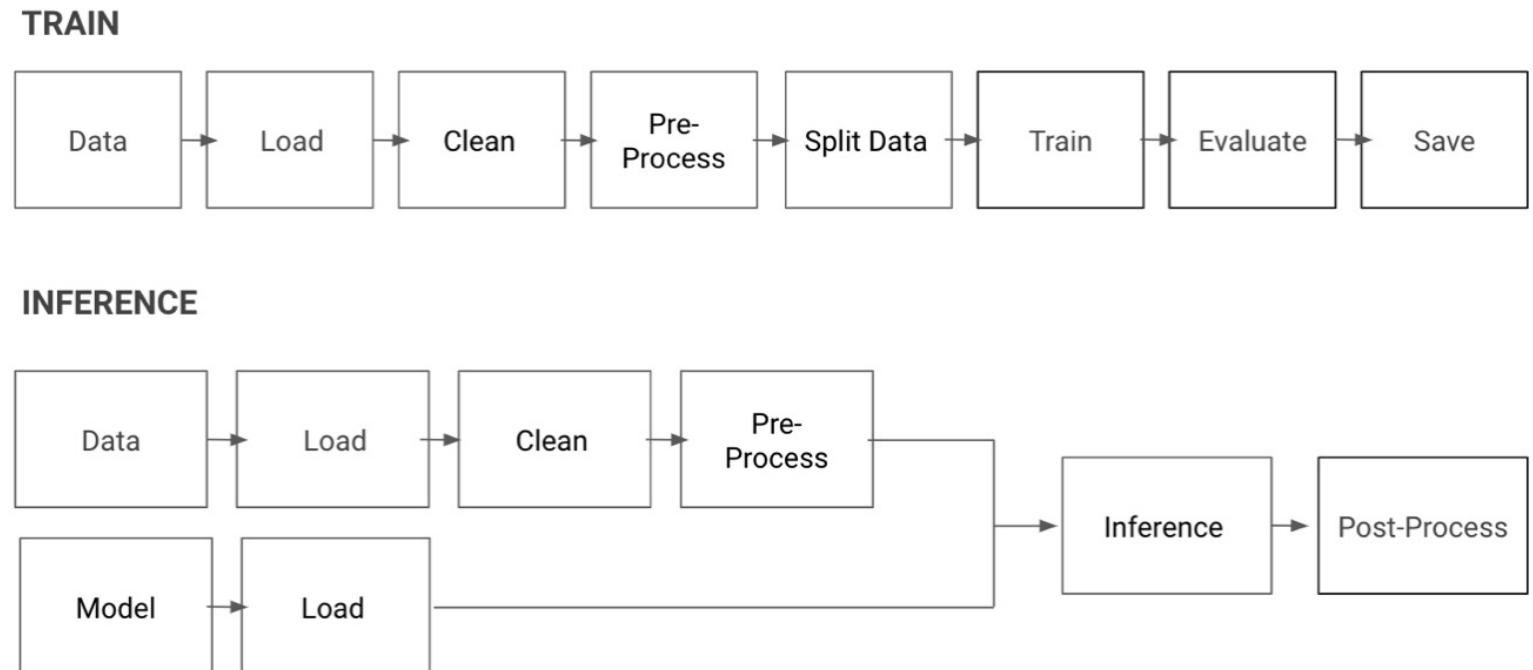


Figure 2-5. Pipelines for the editor

Image from
[Building Machine Learning Powered Applications](#)

Deploying Model

- Deploy Streaming Model Service

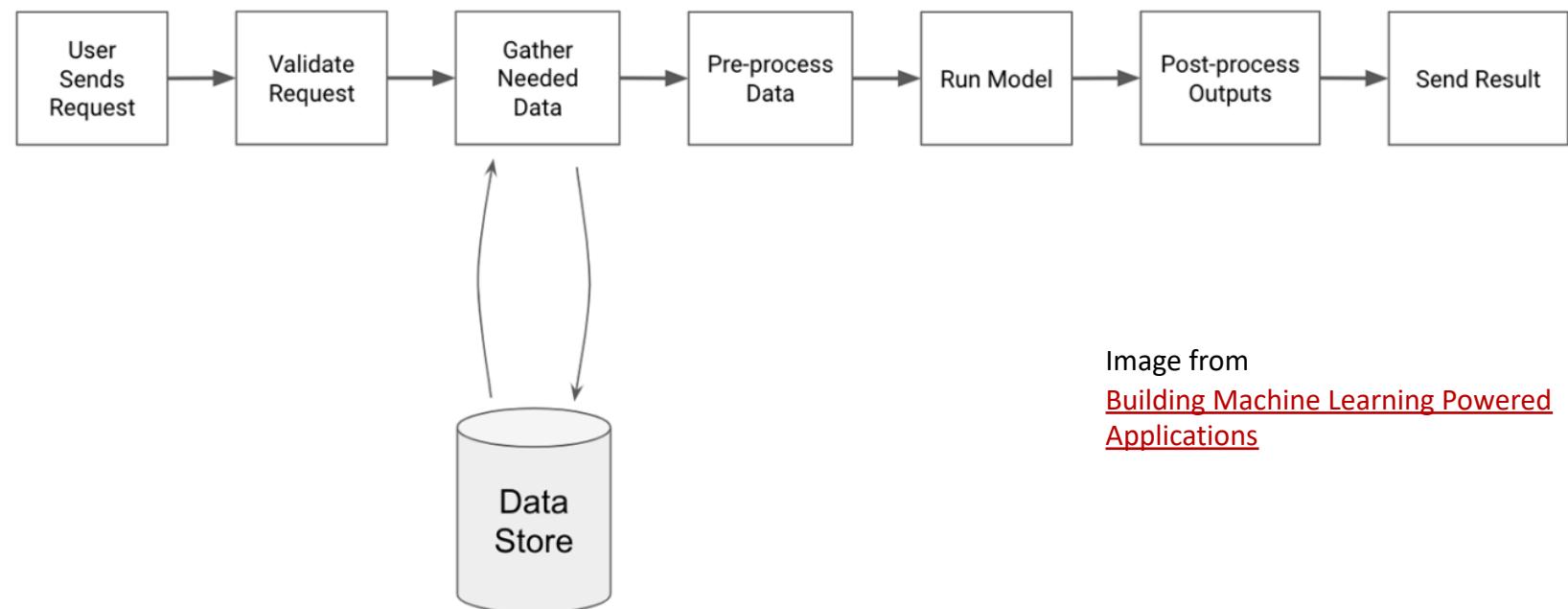


Image from
[Building Machine Learning Powered Applications](#)

Figure 9-1. Streaming API workflow



Why Python?

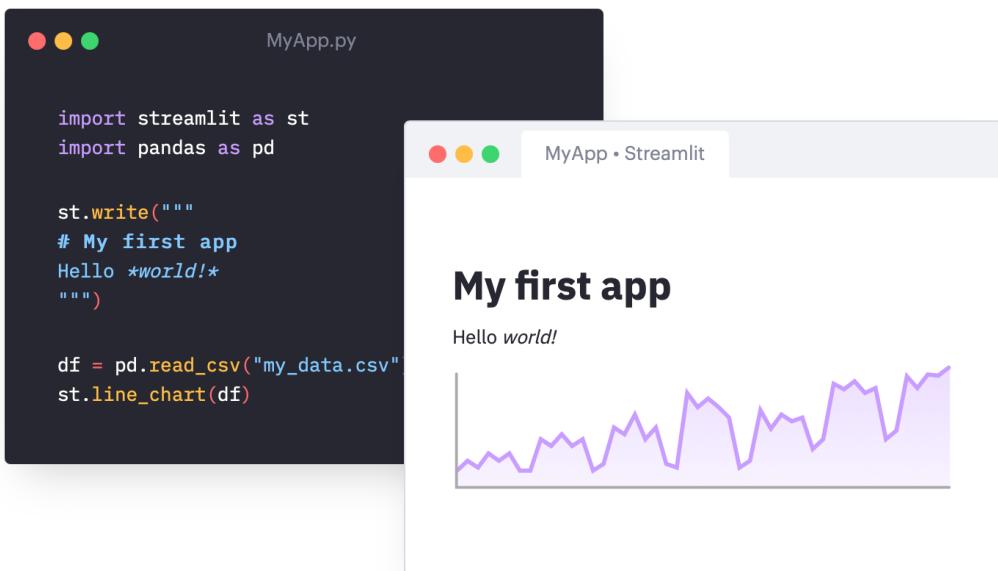




Why Python?

Python is “an interpreted, high-level and general-purpose programming language”

- Versatility
- Code readability
- General and specialized libraries
- Object-oriented design
- **Full data lifecycle support – from Exploratory Analysis to Data Product Development**



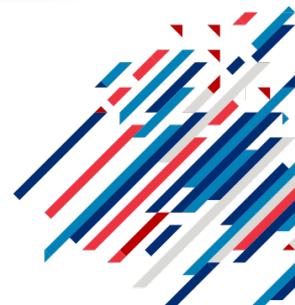
The image shows a comparison between a Python code file and its resulting Streamlit application. On the left, a dark-themed code editor window titled 'MyApp.py' displays the following Python code:

```
import streamlit as st
import pandas as pd

st.write("""
# My first app
Hello *world!*
""")

df = pd.read_csv("my_data.csv")
st.line_chart(df)
```

On the right, a light-themed Streamlit application window titled 'MyApp • Streamlit' shows the output: a title 'My first app' followed by the text 'Hello world!' and a line chart visual.





Python-Powered Applications

- Web Development
- Embedded Systems
- Desktop Applications
- Command Line Interfaces (CLI)
- Data Analytics & Machine Learning
- Scripting



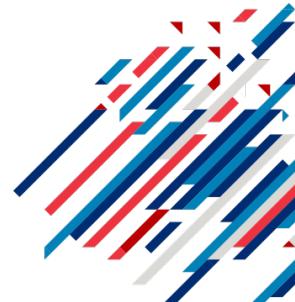
TensorFlow



K Keras

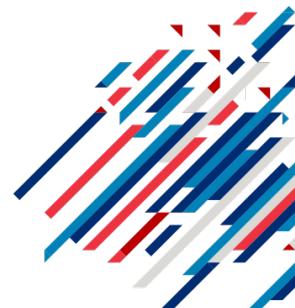


statsmodels



Companies that Rely Upon Python

- Google
 - "Python has been an important part of Google since the beginning, and remains so as the system grows and evolves. Today dozens of Google engineers use Python, and we're looking for more people with skills in this language." - Peter Norvig, director of search quality at Google, Inc.
- Facebook
- Instagram (a Django app!)
- Netflix (super Jupyter proponents)
- Dropbox
- Pinterest (Django)
- Goldman Sachs
- Paypal
- DHS
- DoD



JOBs

Appreciation of Python + MBA is a powerful combo...

- Business Intel Analyst
- Analytics Lead
- Data Scientist
- People Analytics
- Product Manager
- ...

Carnegie Mellon University

Tepper School of Business

William Larimer Mellon, Founder

The screenshot shows a LinkedIn job search interface. The search bar at the top contains 'python mba' with a location filter set to 'United States'. Below the search bar are several filters: 'Jobs' (selected), 'Date Posted', 'Experience Level', and 'Company'. The main search results are displayed under the heading 'Python mba in United States' with '1,703 results'. A 'Job Alert Off' toggle switch is visible next to the results count. Three job listings are shown:

- Analytical Lead, Careers & Education** at Microsoft in Washington, DC. Posted 2 hours ago. The listing includes a Microsoft logo, three profile pictures, and the text '4 connections work here'.
- Fraud Analyst - Apple Pay - San Diego** at Apple in San Diego, CA. Posted 2 days ago. The listing includes an Apple logo, one profile picture, and the text '1 connection works here'.
- Sr. Business Analyst** at Meritor in Florence, IN. The listing includes a small logo featuring a bull's head and the text 'MERITOR'.



Python-Powered ML Tech Stack

Analytics

- Scikit-Learn
- Plotly
- spaCy, keras, tensorflow, torch, etc...



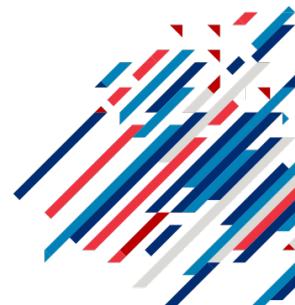
Data Manipulation

- Pandas
- Numpy



Code & Documentation

- Colab
- Jupyter



Generating Insights for AirBnB

BLUF

- Purpose: Clean Data, Visualize Data, Predict Price, Cluster Homes, Parse Reviews

Problem Formation

- Predict Price
- What are the requirements?

Data Curation

- Kaggle AirBnB Data
- Scaling, Imputing, Indicator Variable

Exploratory Analysis

- Plots
- Descriptive Stats

Modeling

- Select techniques per Problem Formation
- Train and fit ML pipeline model

Testing & Evaluation

- Fit Scikit-Learn Model & Statsmodels
- Tune Parameters

Integration

- Cross-Validate



Let's Apply a Little Python + ML to the AirBnB Dataset with an *Interactive Tutorial*

Time to Code!

GitHub

Github Project Link:

<https://github.com/carriegardner428/ML-with-Python-Tepper-CY22-Mini4>

colab

Colab Link:

<https://colab.research.google.com/github/carriegardner428/ML-with-Python-Tepper-CY22-Mini4/blob/main/>



Parting Thoughts

- Thank you for your attention!
 - If you liked this workshop, please endorse one of my skills on LinkedIn 
- Please reach out if you want to...
 - Learn more on these topics
 - Get more hands-on python and analytics experience (e.g., collaborative coding events)
 - Share your knowledge and experience on these topics
- Again: You can do this! 





Resources

- Hands-on Machine Learning with Scikit-Learn, Keras, & TensorFlow (Book)
- ML Design Patterns (Book)
- Introducing MLOps (Book)
- Building Machine Learning Powered Applications (Book)





Backup Slides

Carnegie Mellon University

Tepper School of Business

William Larimer Mellon, Founder

Slide 29

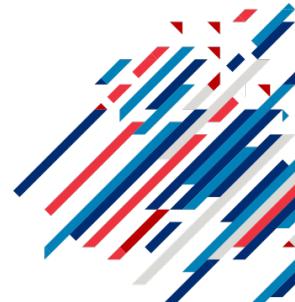
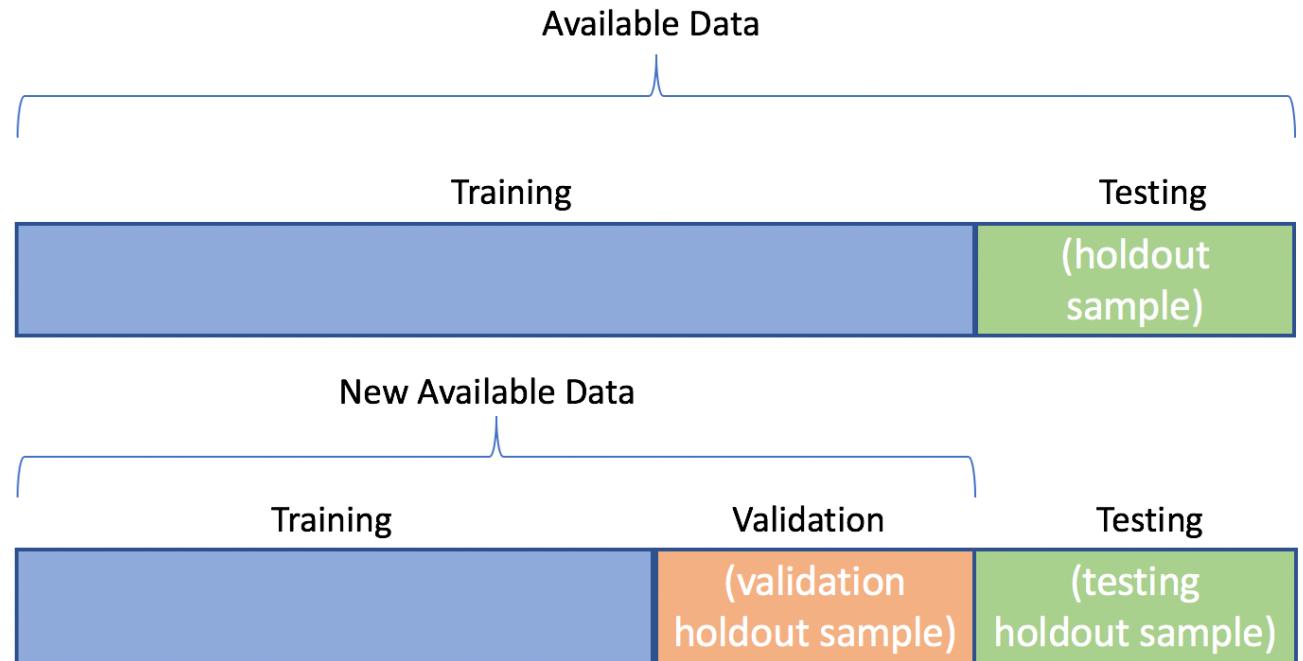


Model Training & Evaluation



Training & Evaluation (T&E)

- Training: data used to fit (train) the model (typically accounts for 70-90% of all data)
- Evaluation (Test): data used to at the end of fitting to measure how robust the model is to handling other data
- Holdout (Validation): data used to assess the performance of the model after the model has been fitted, tested, and is ready for deployment





General Tasks

- Clean Data
- Visualize Data
- Predict Labels
- Presence Detection
- Predict Numerical Values
- Detect Topics in Text
- Detect Sentiment
- Summarize Text
- Infer Absolute and Relative Risk
- Find Clusters
- Real-Time Decisions
- Automate Processes
- ...

