Data Science Pipelines

Christopher Tunnell (Rice University)

Delhi University

21-May-2024

All “HEP results” or “HEP analyses” are conceptually examples of

Data Science Pipelines

- Literature review
- Knowledge gaps

- Get data from infrastructure
- Build or learn data collection tools

- Wrangle
- Data shaping
- Missing data
- Filter and transform

- Visualize/plot data
- Find patterns
- Build exploration tools

- Extract Features
- Fit computational models
- Validate hypotheses

- Communicate findings
- Build data products

Data Science Pipelines

Questions:

1. Where does machine learning typically go?
2. Where do particle physicists spend most of their time?

Frame Problem

- Literature review
- Knowledge gaps

Acquire Data

- Get data from infrastructure
- Build or learn data collection tools

Preprocess

- Wrangle
- Data shaping
- Missing data
- Filter and transform

Explore

- Visualize/plot data
- Find patterns
- Build exploration tools

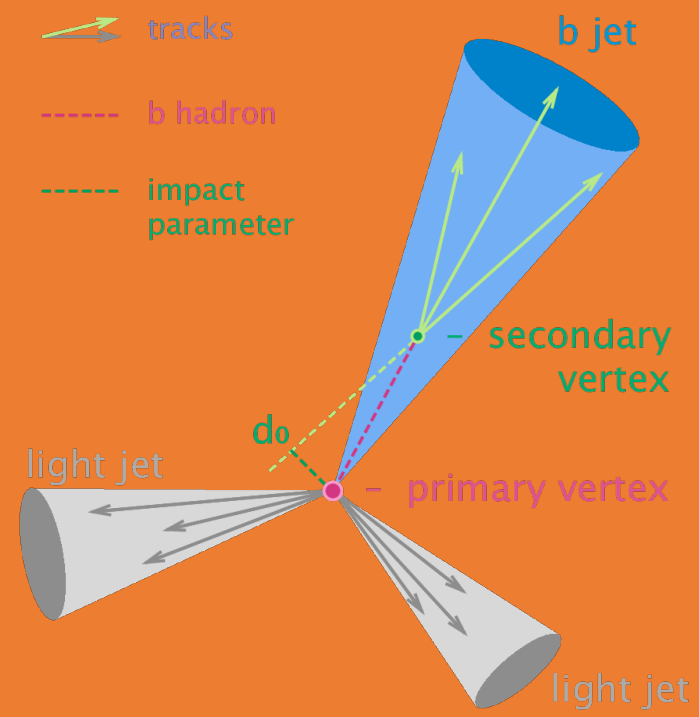
Model

- Extract Features
- Fit computational models
- Validate hypotheses

Deploy

- Communicate findings
- Build data products

- What do you want to learn? (ie what is your thesis topic)
  - Analyze small detector data for R&D task?
  - Understand hadronic jets?
- Who has done similar things before before?
  - If the same thing done, use it.
  - If not, why hasn't it been done before?
- Where are the 'holes' in prior work?



Frame Problem

- Literature review
- Knowledge gaps

Acquire Data

- Get data from infrastructure
- Build or learn data collection tools

Preprocess

- Wrangle
- Data shaping
- Missing data
- Filter and transform

Explore

- Visualize/plot data
- Find patterns
- Build exploration tools

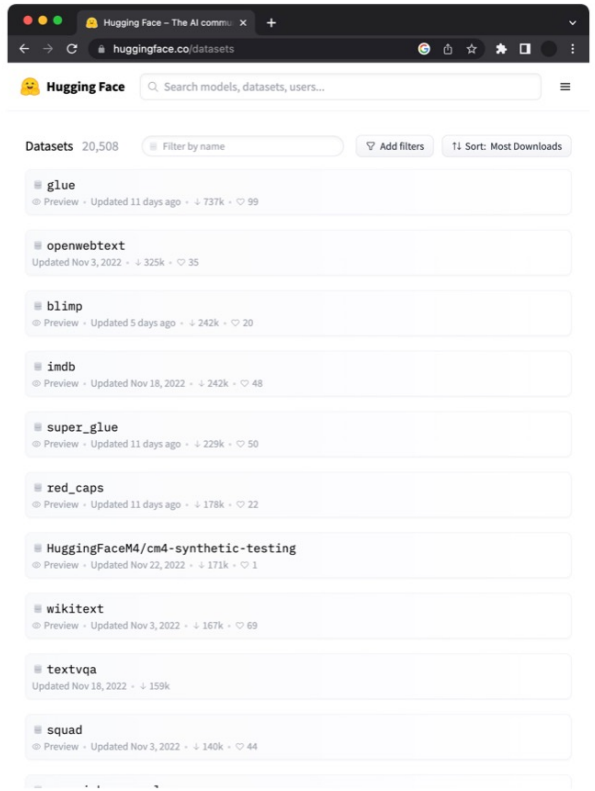
Model

- Extract Features
- Fit computational models
- Validate hypotheses

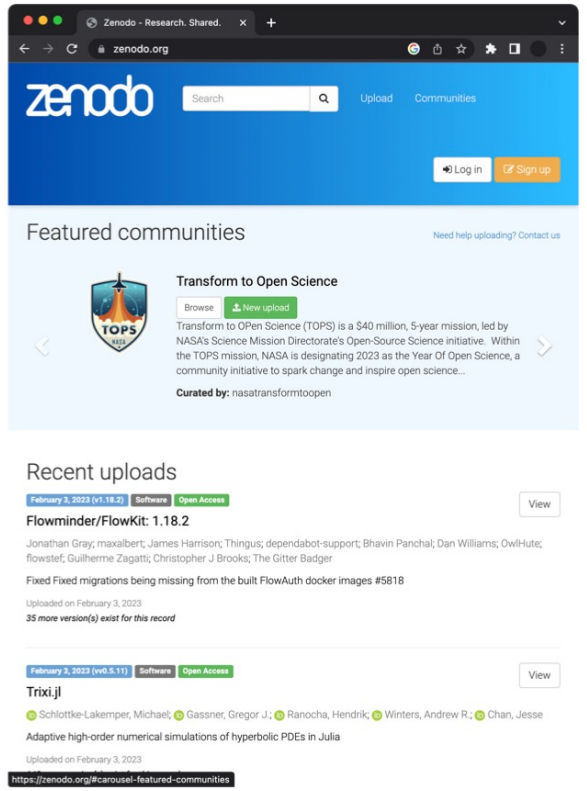
Deploy

- Communicate findings
- Build data products

- "Data" in this context means experimental data or simulations
- In HEP, data is often:
  - Experimental sensor data from the DAQ, often spatiotemporal, or other higher level data product
  - Existing simulation that you run yourself for infinite statistics
- Other datasets exist (Zenodo, Hugging Face)
- Ideally there is an existing simulation or data loading code… though not always.



Hugging Face



Zenodo

Frame Problem

- Literature review
- Knowledge gaps

Acquire Data

- Get data from infrastructure
- Build or learn data collection tools

Preprocess

- Wrangle
- Data shaping
- Missing data
- Filter and transform

Explore

- Visualize/plot data
- Find patterns
- Build exploration tools

Model

- Extract Features
- Fit computational models
- Validate hypotheses

Deploy

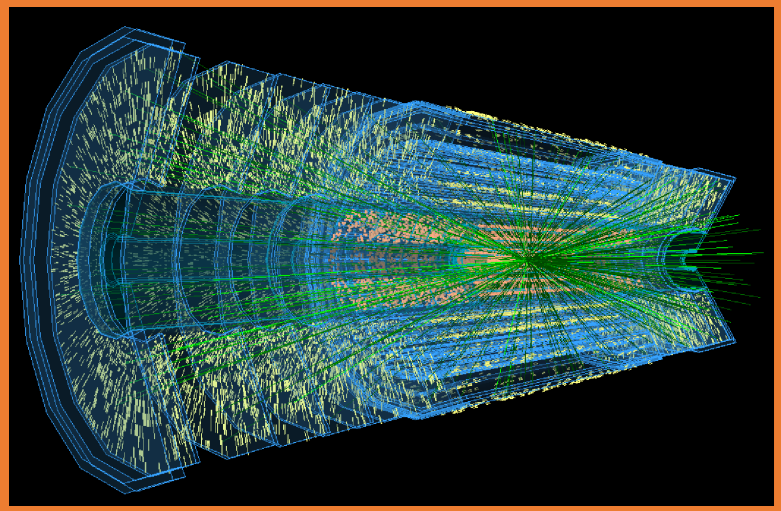
- Communicate findings
- Build data products

Shape: Data ‘wrangling’ or ‘shaping’ is 80%+ of work for most particle ML applications.

Particle physics data is hierarchical and ‘jagged’, which does not map trivially to computer memory. (Ask Jim):

1. Event 1
    1. Jet 1: px, py, pz, E
    2. Jet 2: px, py, pz, E
2. Event 2:
    1. Jet 1: px, py, pz, E
    2. Jet 2: px, py, pz, E
    3. Jet 3: px, py, pz, E

Filter and transform so input isn’t entire e.g. CMS dataset.



Frame Problem

- Literature review
- Knowledge gaps

Acquire Data

- Get data from infrastructure
- Build or learn data collection tools

Preprocess

- Wrangle
- Data shaping
- Missing data
- Filter and transform

Explore

- Visualize/plot data
- Find patterns
- Build exploration tools

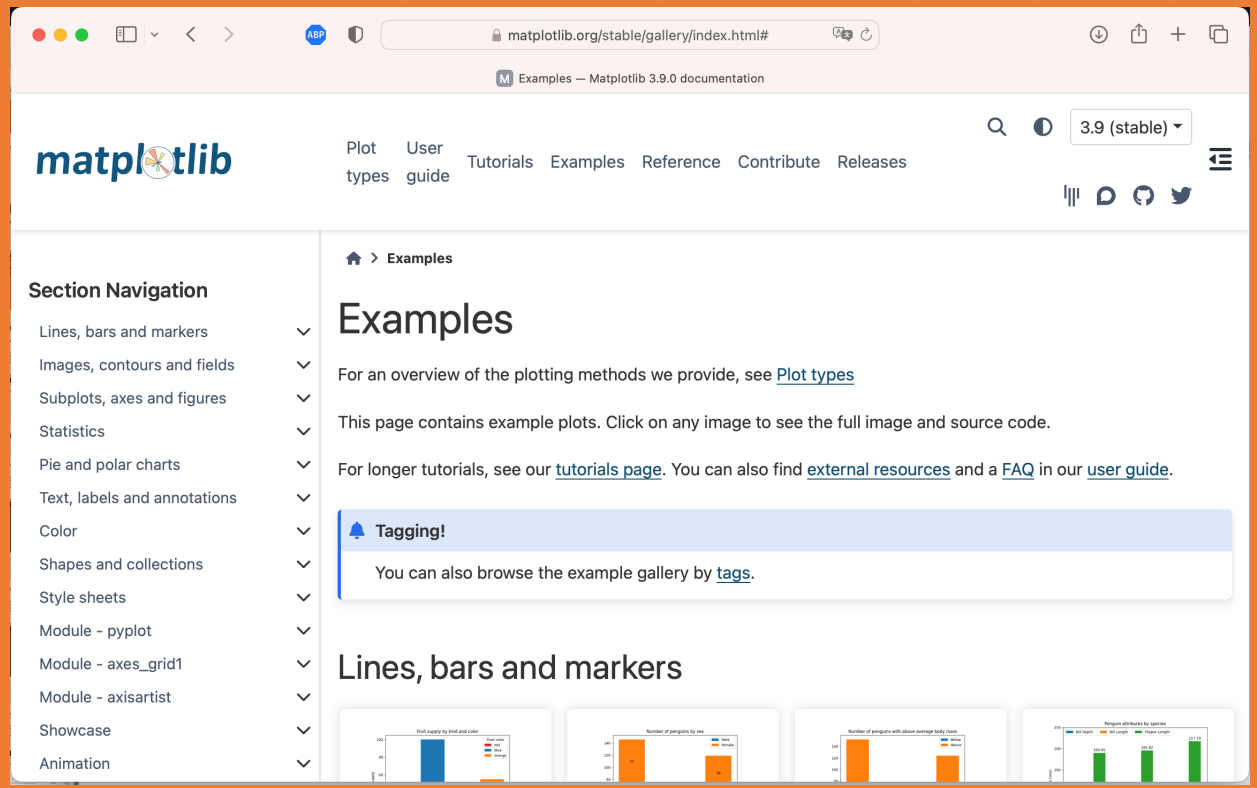
Model

- Extract Features
- Fit computational models
- Validate hypotheses

Deploy

- Communicate findings
- Build data products

Make plots to understand your data



Frame Problem

- Literature review
- Knowledge gaps

Acquire Data

- Get data from infrastructure
- Build or learn data collection tools

Preprocess

- Wrangle
- Data shaping
- Missing data
- Filter and transform

Explore

- Visualize/plot data
- Find patterns
- Build exploration tools

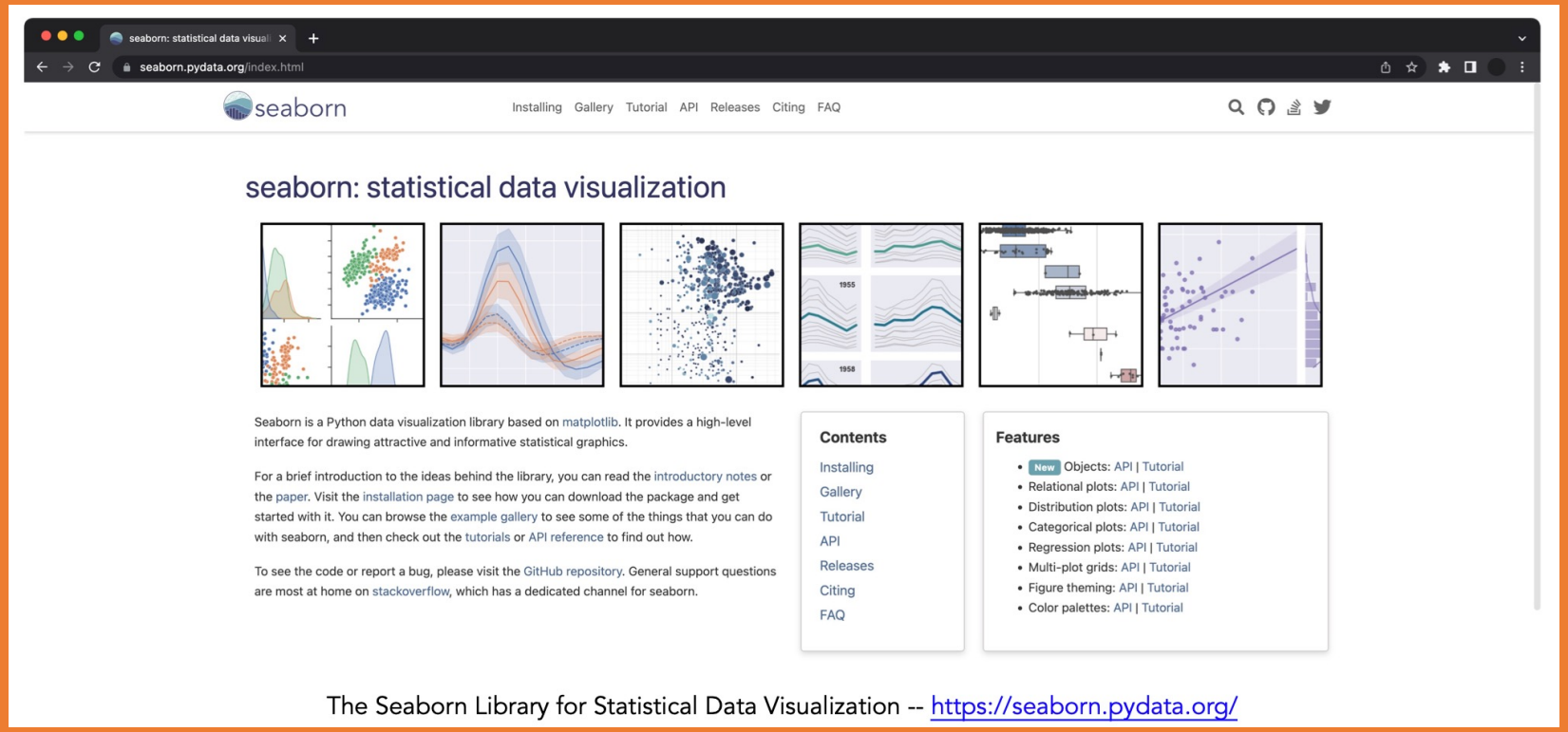
Model

- Extract Features
- Fit computational models
- Validate hypotheses

Deploy

- Communicate findings
- Build data products

Make plots



The Seaborn Library for Statistical Data Visualization -- https://seaborn.pydata.org/

Frame Problem

- Literature review
- Knowledge gaps

Acquire Data

- Get data from infrastructure
- Build or learn data collection tools

Preprocess

- Wrangle
- Data shaping
- Missing data
- Filter and transform

Explore

- Visualize/plot data
- Find patterns
- Build exploration tools

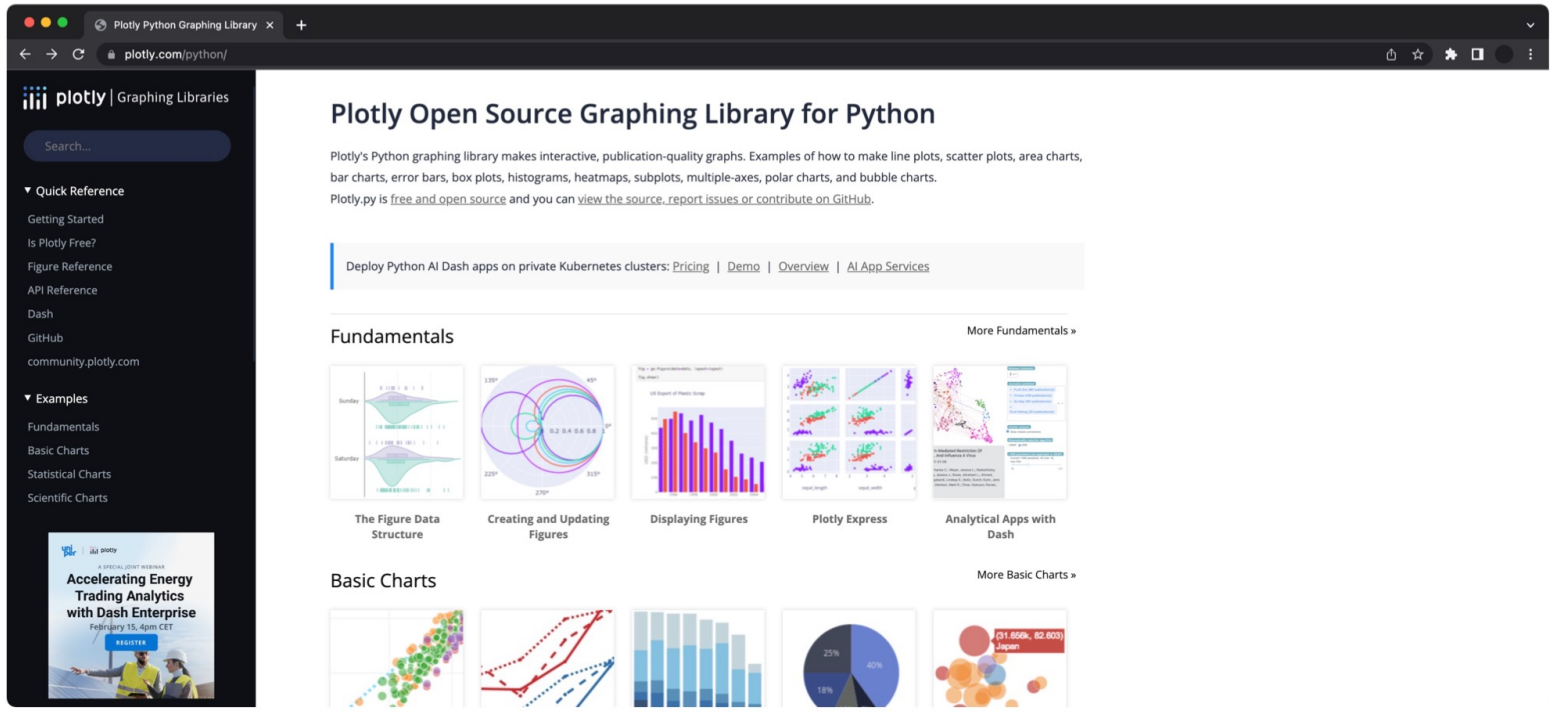
Model

- Extract Features
- Fit computational models
- Validate hypotheses

Deploy

- Communicate findings
- Build data products

Make plots



Plotly Open Source Graphing Library for Python -- https://plotly.com/python/

Frame Problem

Acquire Data

- Literature review
- Knowledge gaps

Preprocess

- Get data from infrastructure
- Build or learn data collection tools

Explore

- Wrangle
- Data shaping
- Missing data
- Filter and transform

Model

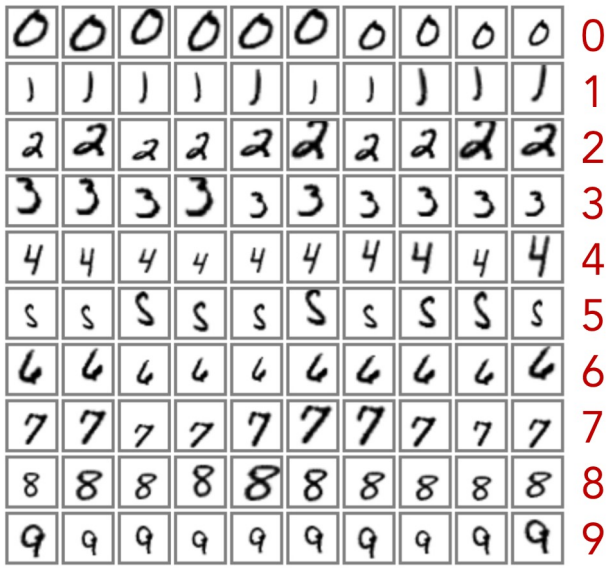
- Visualize/plot data
- Find patterns
- Build exploration tools

Deploy

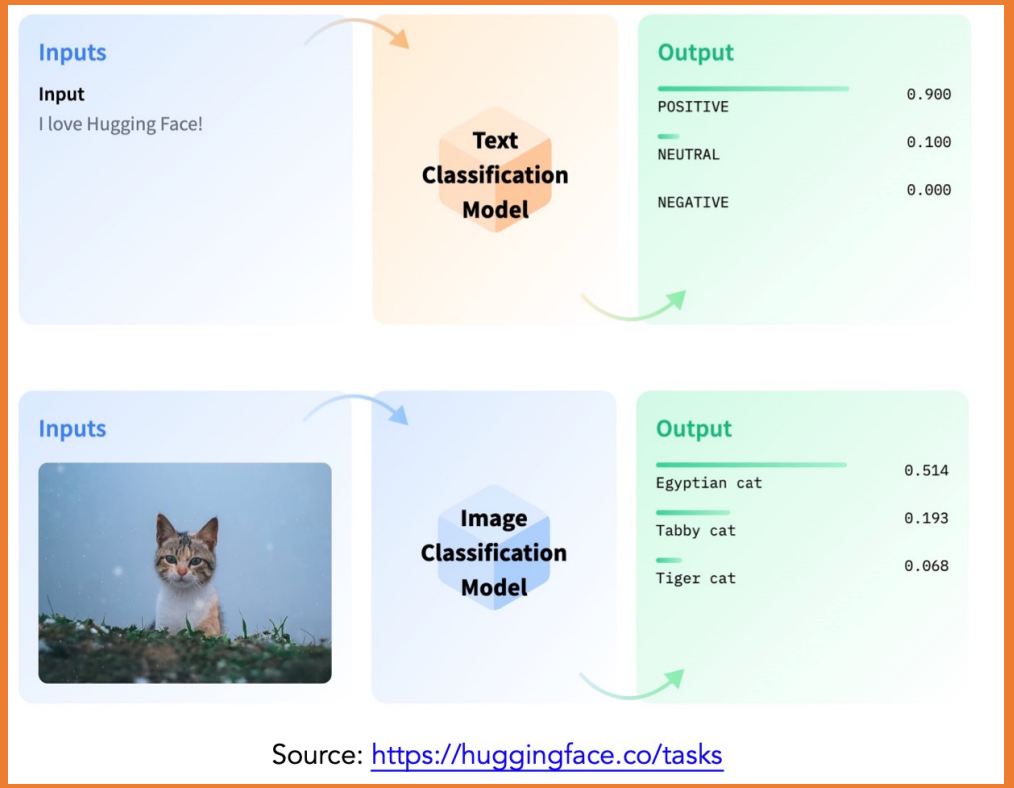
- Extract Features
- Fit computational models
- Validate hypotheses

- Communicate findings
- Build data products

Statistical modeling and machine learning



LeCun et al. 1998. Gradient-based learning applied to document recognition. IEEE.



Source: https://huggingface.co/tasks

Frame Problem

Acquire Data

Preprocess

Explore

Model

Deploy

- Literature review
- Knowledge gaps

- Get data from infrastructure
- Build or learn data collection tools

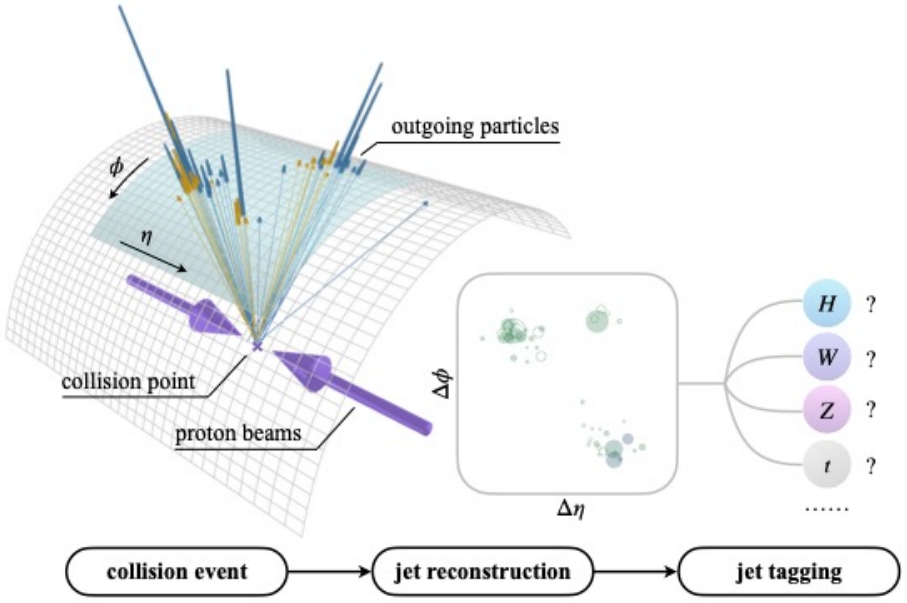
- Wrangle
- Data shaping
- Missing data
- Filter and transform

- Visualize/plot data
- Find patterns
- Build exploration tools

- Extract Features
- Fit computational models
- Validate hypotheses

- Communicate findings
- Build data products

Statistical modeling and machine learning



arXiv:2202.03772

Retrieved from https://ebird.org

Communicating:

After wrangling, hardest part is communicating results to physicists and ML practitioners since both have extreme levels of jargon and cultural considerations.

Production:

Depending on experiment, getting your algorithm into the software stack cnd used could potentially be a major work.

- Literature review
- Knowledge gaps

- Get data from infrastructure
- Build or learn data collection tools

- Wrangle
- Data shaping
- Missing data
- Filter and transform

- Visualize/plot data
- Find patterns
- Build exploration tools

- Extract Features
- Fit computational models
- Validate hypotheses

- Communicate findings
- Build data products

- Additional Resources:
  - Examples:
    - Read the Smell Pittsburgh paper for example of pipeline
    - Machine Learning Pipelines with Modern Big Data Tools for High Energy Physics
  - The paper below studies various data science pipelines at different scale, which can give you a good understanding of common data science practices:
    - The Art and Practice of Data Science Pipelines
  - Below are website for data visualization inspirations:
    - Seaborn: Statistical Data Visualization
    - Exploratory Data Analysis by the US EPA
    - Examples of Data Exploration by the Statistics Netherlands
    - Examples of Data Visualization
  - Below are interesting data science case studies:
    - Case Studies of Satelite Image Analysis
    - Case Studies of Machine Learning and Design
  - The textbook below contains more information about how to select models:
    - Section 11.8 Comparing Different Models in book: Introduction to Statistics and Data Analysis
- Slides – or at least this list -- loosely adapted from Yen-Chia Hsu (UvA)