

# Data Science Pipelines

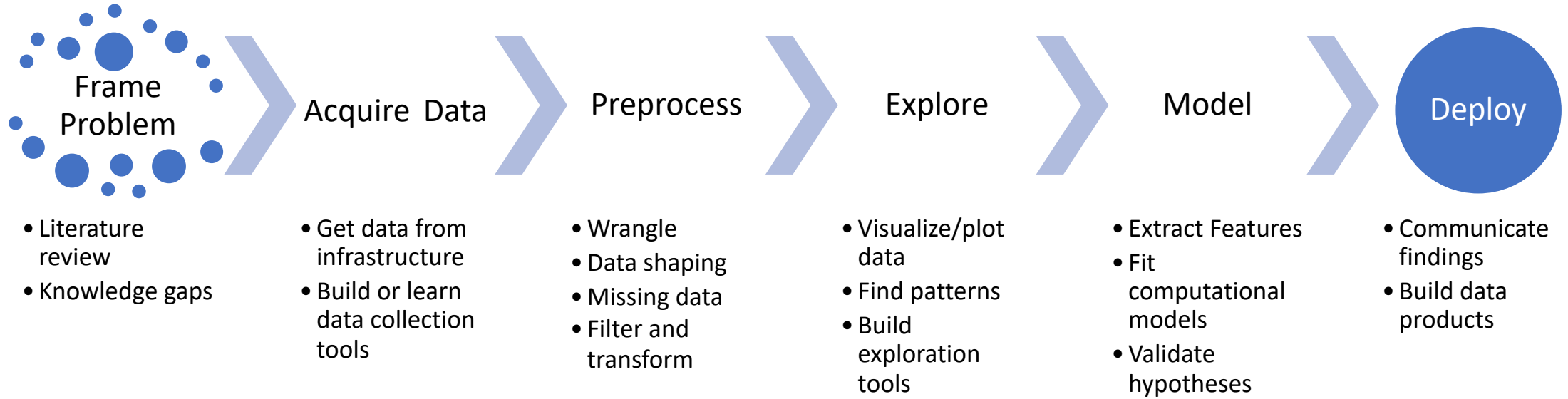
Christopher Tunnell (Rice University)

Delhi University

21-May-2024

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

# Data Science Pipelines



# Data Science Pipelines

## Questions:

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



- Literature review
- Knowledge gaps

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

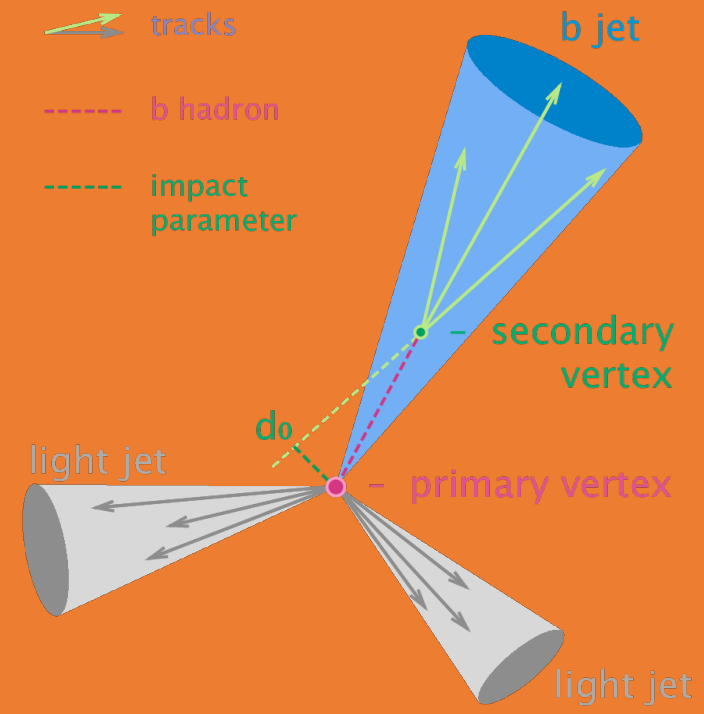
- Wrangle
- Data shaping
- Missing data
- Filter and transform

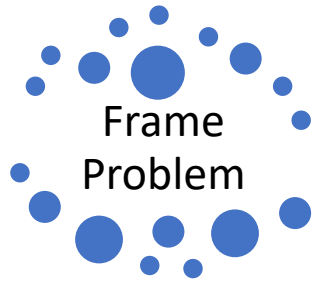
- Visualize/plot data
- Find patterns
- Build exploration tools

- Extract Features
- Fit computational models
- Validate hypotheses

- Communicate findings
- Build data products

- 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

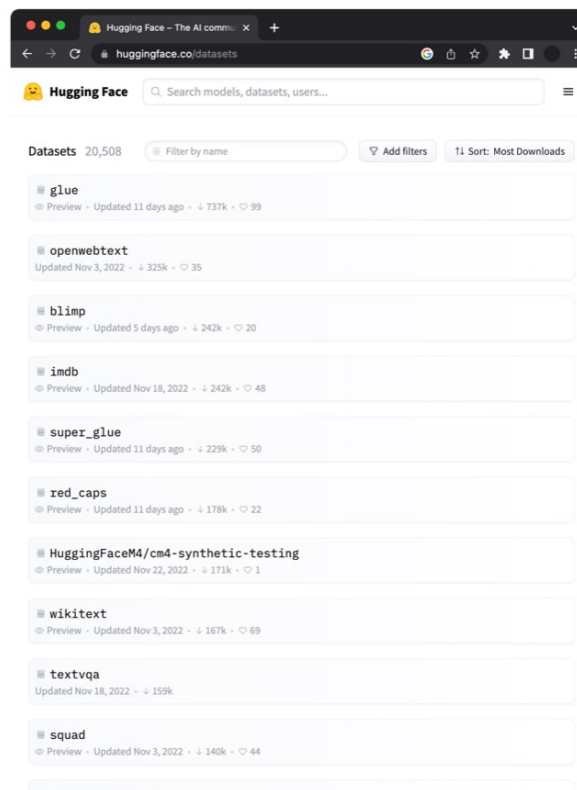
## Model

- Extract Features
- Fit computational models
- Validate hypotheses

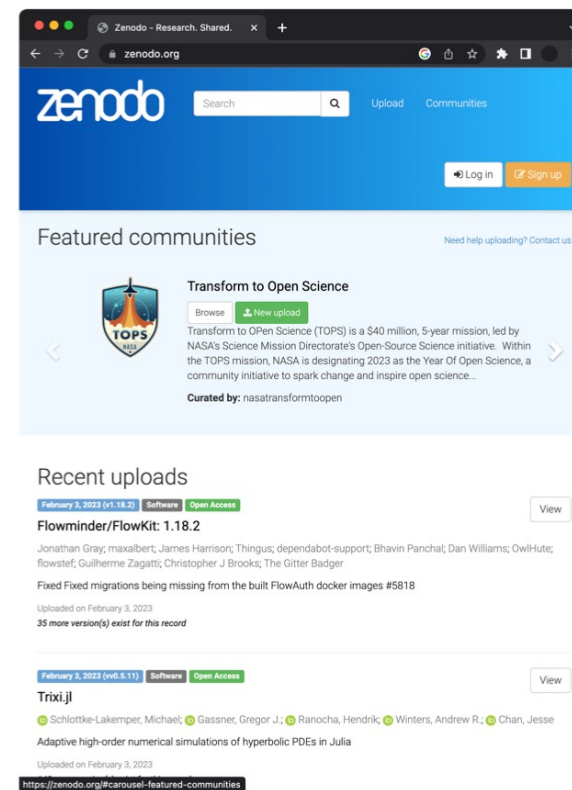
## Deploy

- Communicate findings
- Build data products

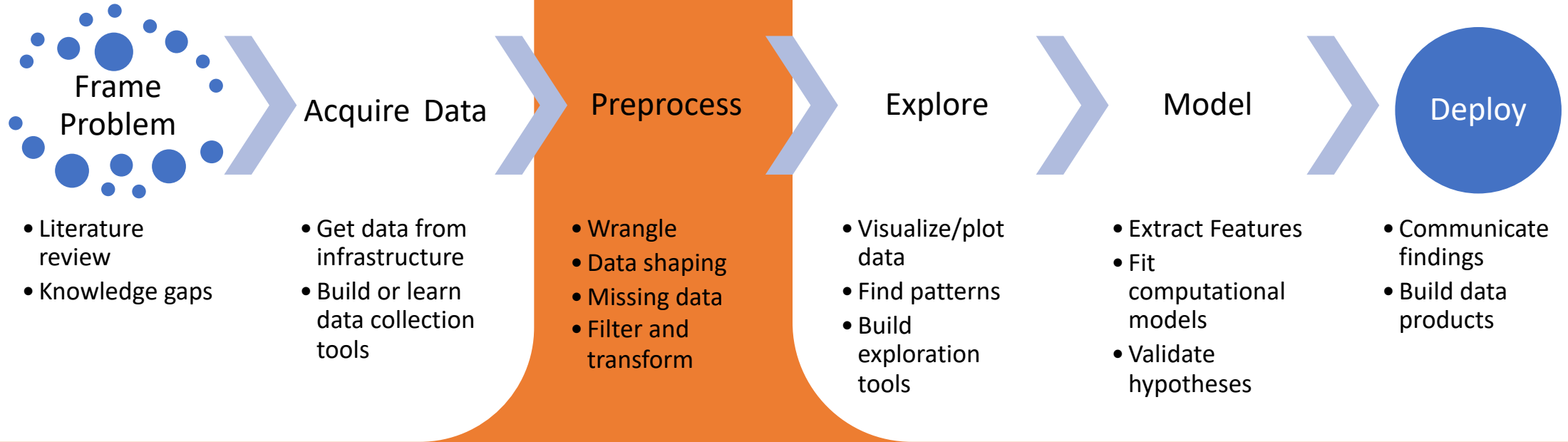
- “Data” in this context means experimental data or simulations
- In HEP, data is often:
  - Experimental sensor data from the DAQ, often spatiotemporal, or other higher level data product
  - Existing simulation that you run yourself for infinite statistics
- Other datasets exist (Zenodo, Hugging Face)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



[Hugging Face](https://huggingface.co/datasets)



[Zenodo](https://zenodo.org)



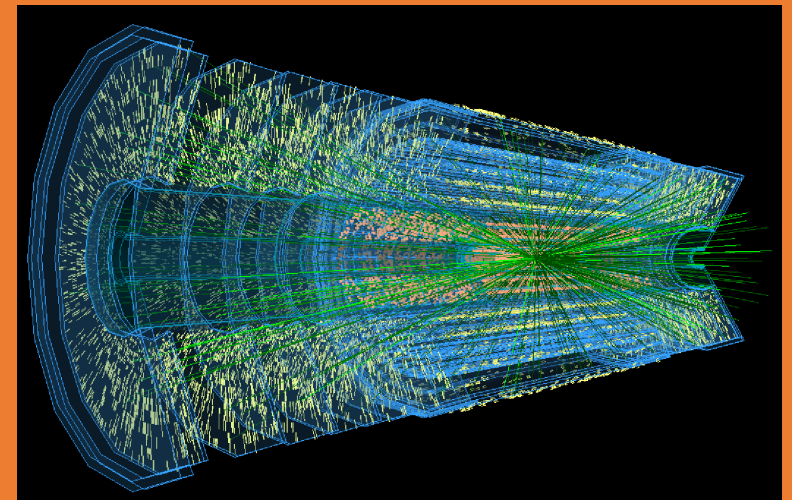
**Shape:** Data ‘wrangling’ or ‘shaping’ is 80%+ of work for most particle ML applications.

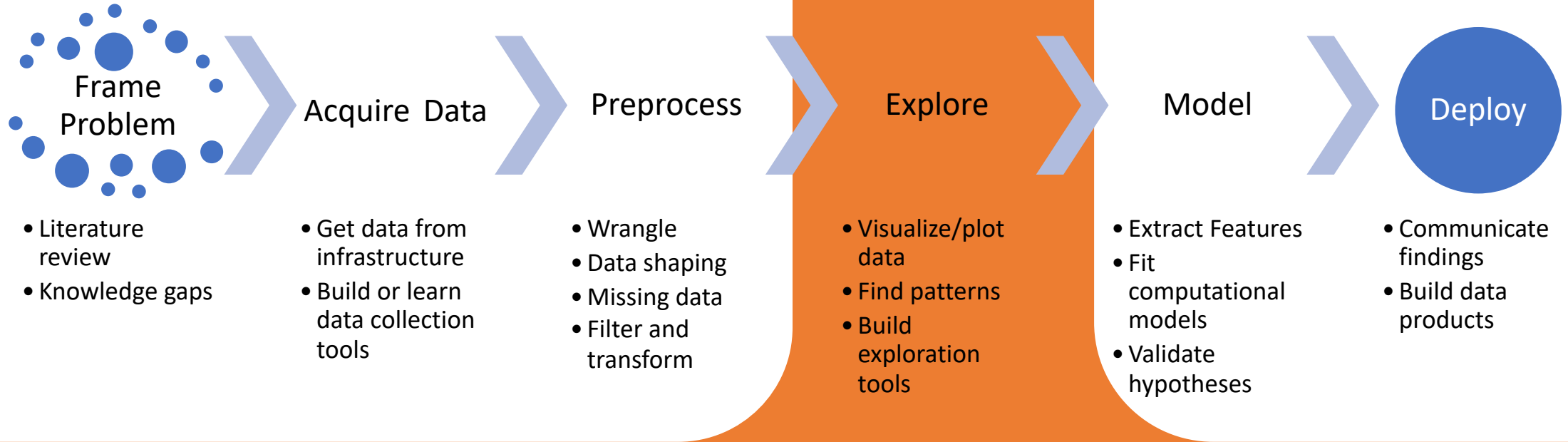
Particle physics data is hierarchical and ‘jagged’, which does not map trivially to computer memory. (Ask Jim):

1. Event 1
  1. Jet 1:  $p_x$ ,  $p_y$ ,  $p_z$ ,  $E$
  2. Jet 2:  $p_x$ ,  $p_y$ ,  $p_z$ ,  $E$
2. Event 2:
  1. Jet 1:  $p_x$ ,  $p_y$ ,  $p_z$ ,  $E$
  2. Jet 2:  $p_x$ ,  $p_y$ ,  $p_z$ ,  $E$
  3. Jet 3:  $p_x$ ,  $p_y$ ,  $p_z$ ,  $E$

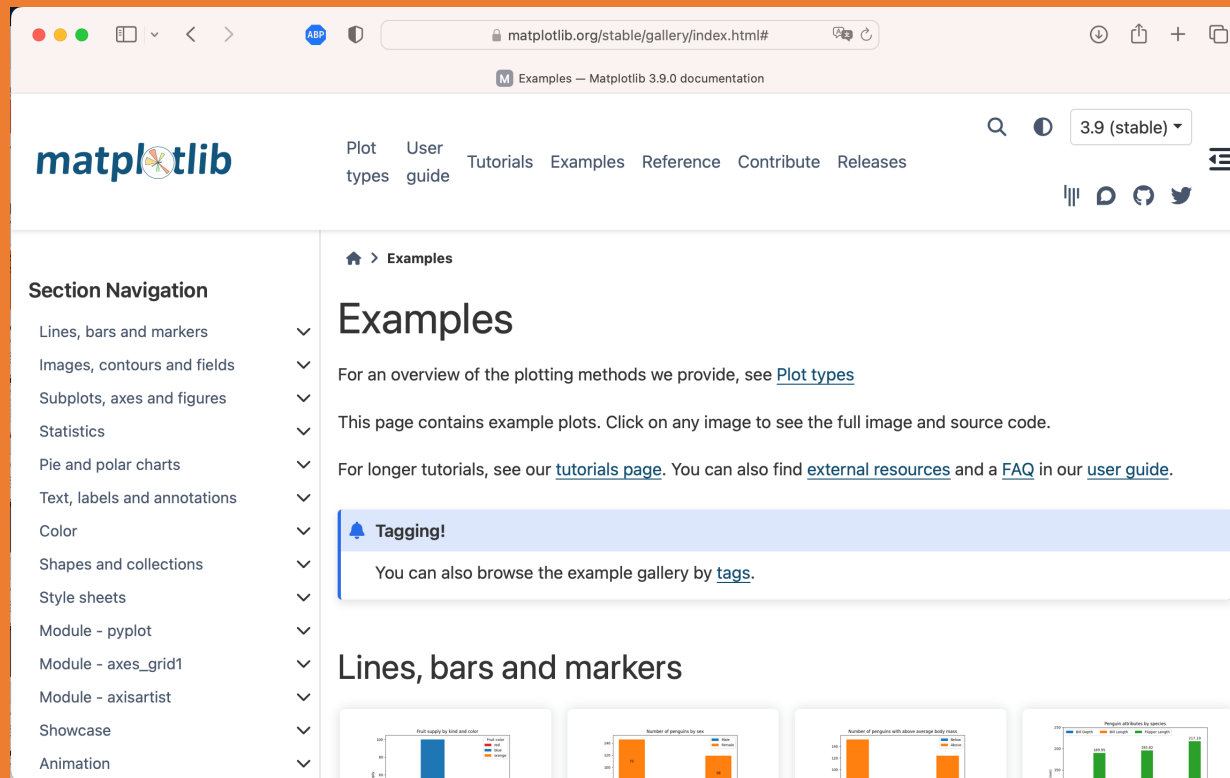


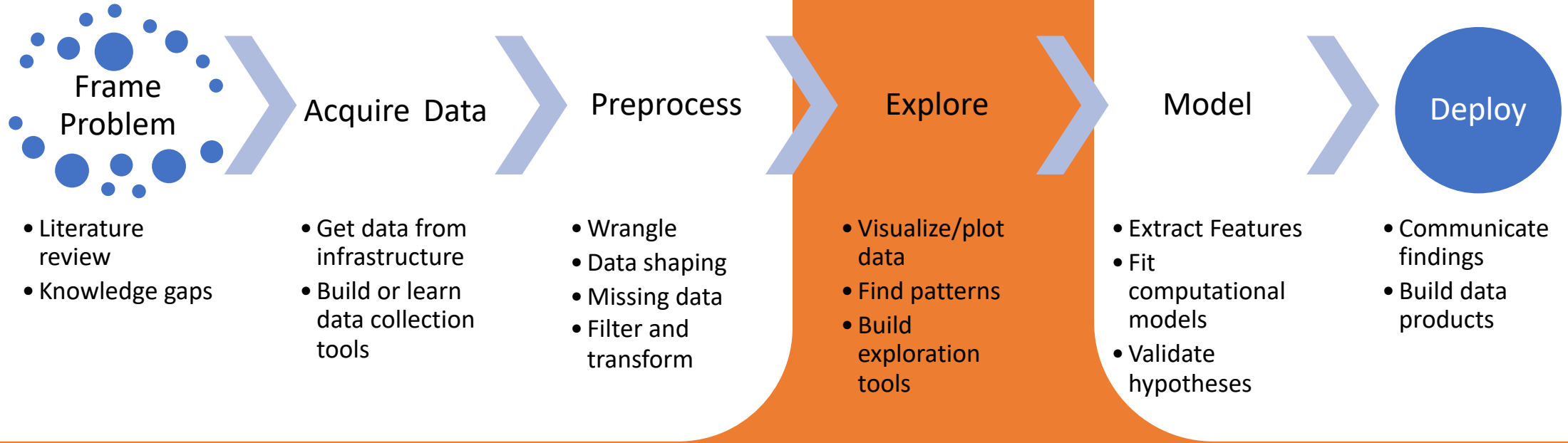
**Filter** and transform so input isn’t entire e.g. CMS dataset.





**Make plots to understand your data**





## Make plots

seaborn: statistical data visualization

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

For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see some of the things that you can do with seaborn, and then check out the tutorials or API reference to find out how.

To see the code or report a bug, please visit the GitHub repository. General support questions are most at home on stackoverflow, which has a dedicated channel for seaborn.

**Contents**

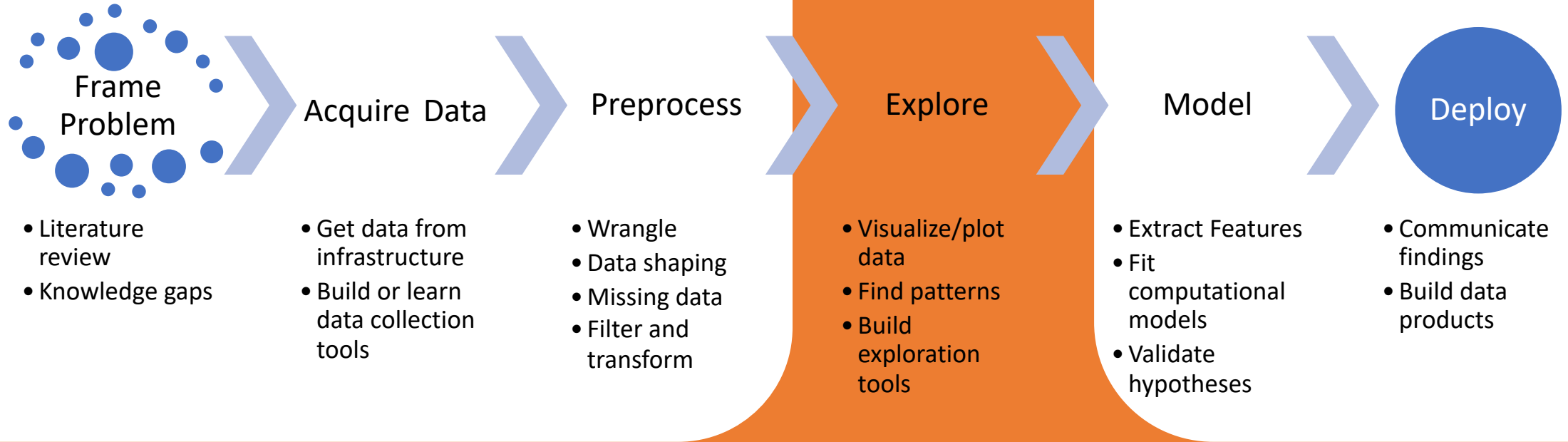
- Installing
- Gallery
- Tutorial
- API
- Releases
- Citing
- FAQ

**Features**

- **New** Objects: API | Tutorial
- Relational plots: API | Tutorial
- Distribution plots: API | Tutorial
- Categorical plots: API | Tutorial
- Regression plots: API | Tutorial
- Multi-plot grids: API | Tutorial
- Figure theming: API | Tutorial
- Color palettes: API | Tutorial

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





## Make plots

Plotly Python Open Source Graphing Library for Python

Plotly's Python graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts. Plotly.py is free and open source and you can view the source, report issues or contribute on GitHub.

Deploy Python AI Dash apps on private Kubernetes clusters: [Pricing](#) | [Demo](#) | [Overview](#) | [AI App Services](#)

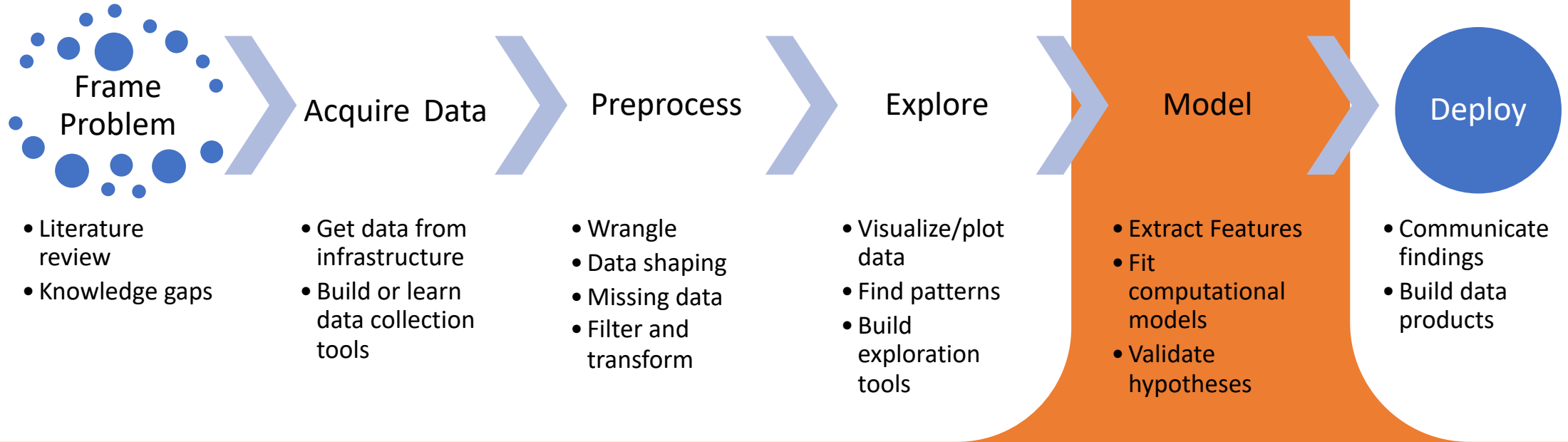
**Fundamentals**

- The Figure Data Structure
- Creating and Updating Figures
- Displaying Figures
- Plotly Express
- Analytical Apps with Dash

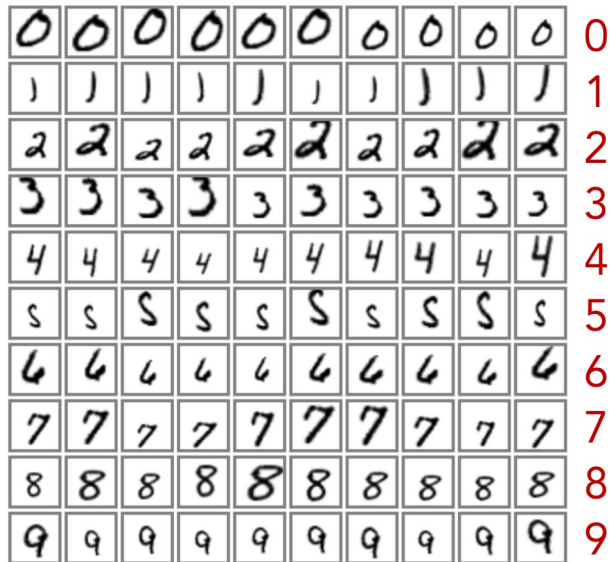
**Basic Charts**

- Scatter plots
- Line plots
- Bar charts
- Pie charts
- Bubble charts

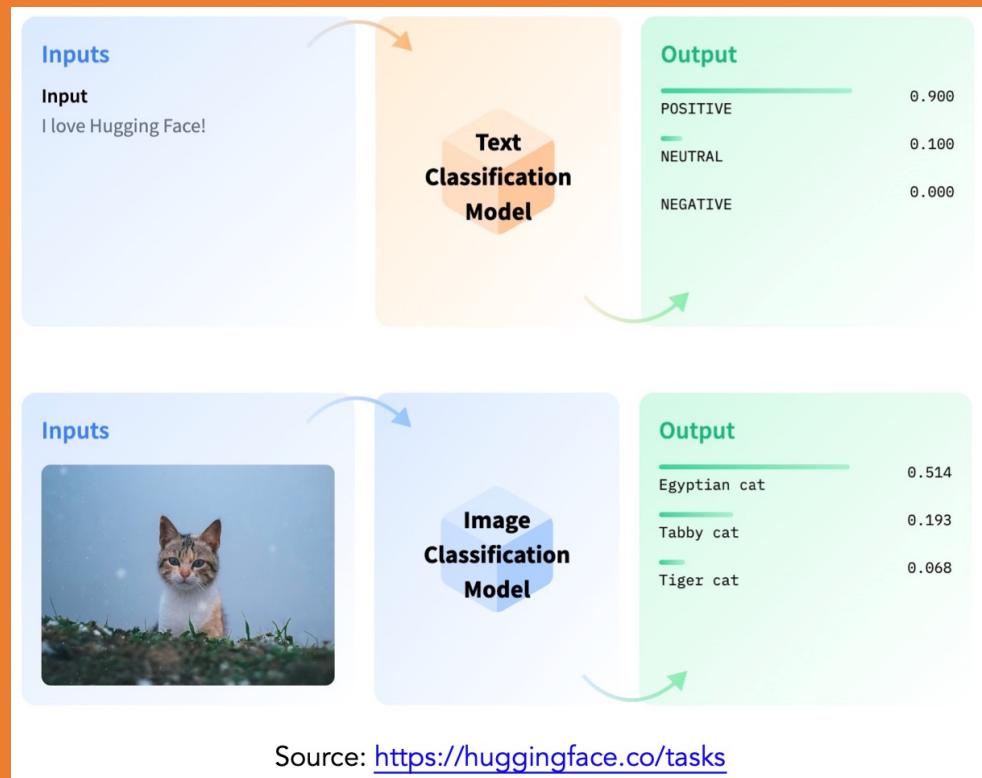
Plotly Open Source Graphing Library for Python -- <https://plotly.com/python/>

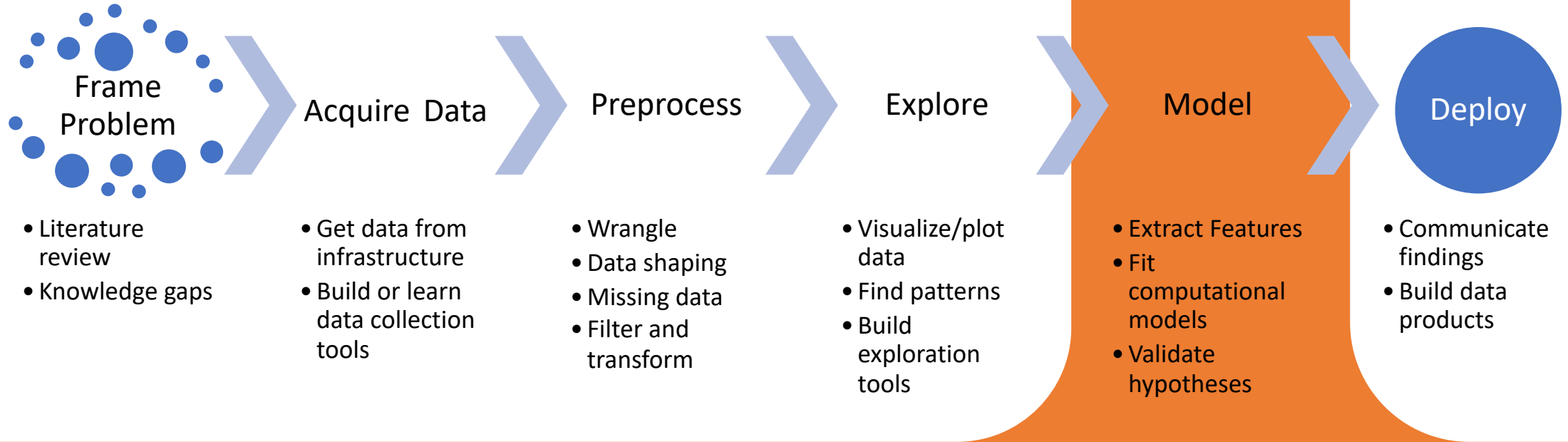


## Statistical modeling and machine learning

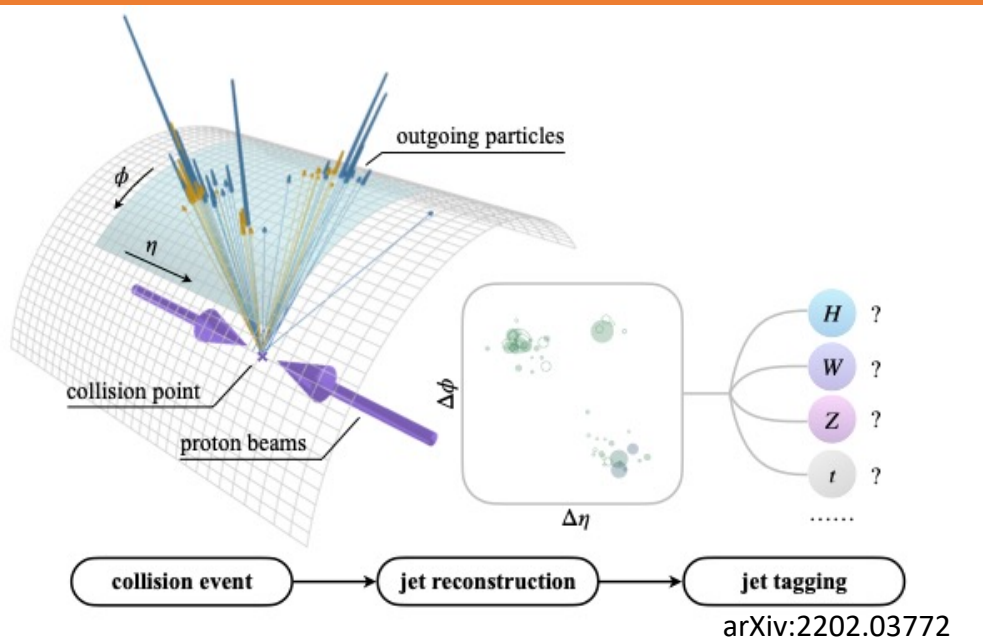


LeCun et al. 1998. Gradient-based learning applied to document recognition. IEEE.





## Statistical modeling and machine learning



Barred Owl



American Robin



American Crow



Rufous Hummingbird

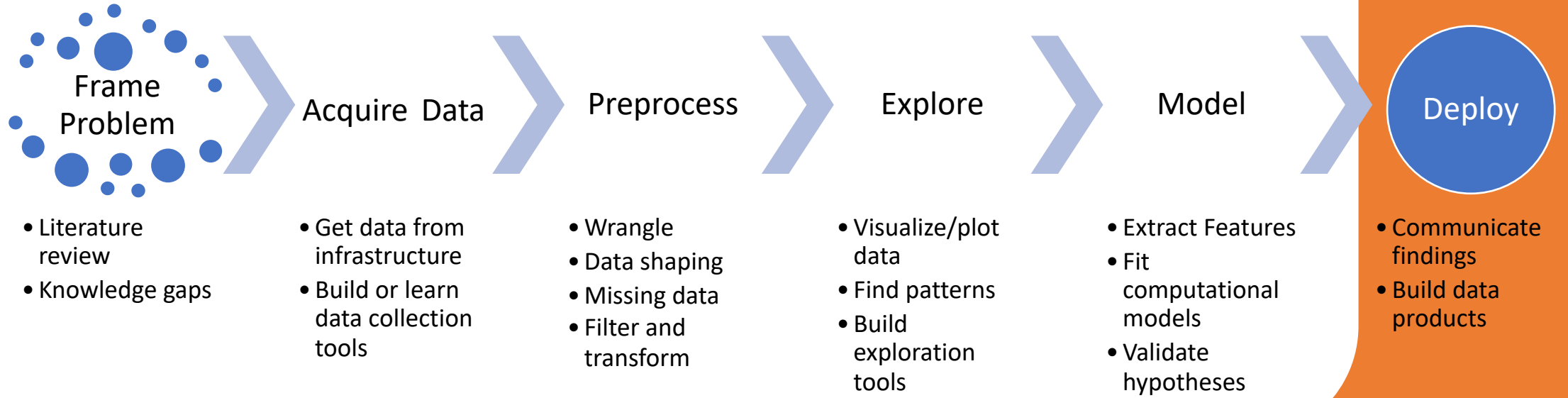


Rock Pigeon



Canada Goose

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 and 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 Satellite 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)