AFRICA DATA SCHOOL

DATA SCIENCE FRAMEWORK & TOOLS

AGENDA

- Applied Data Science
- CRISP -DM Project Framework
- Data tools

Applied Data Science

Applied data scientists

- Applied scientists use Data science and ML to improve business
  outcomes (e.g., revenue, cost, customer experience).
- The systems they build may be internal (e.g., product classification, fraud detection) or customer-facing (e.g., search,
  recommendations).
- Outside of use-case driven applications, they might also develop internal datasets, tooling, and methodology (e.g., feature stores, package/docker templates, model testing & release checks)
- Applied data scientists have higher and deep technical knowledge of how data science and its methods work.

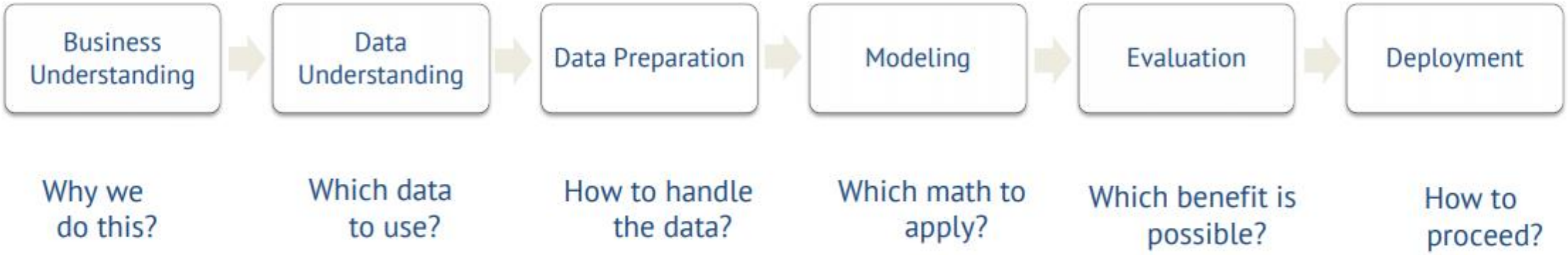
AFRICA DATA SCHOOL

The Cross Industry Standard Process for Data Mining (CRISP-DM)

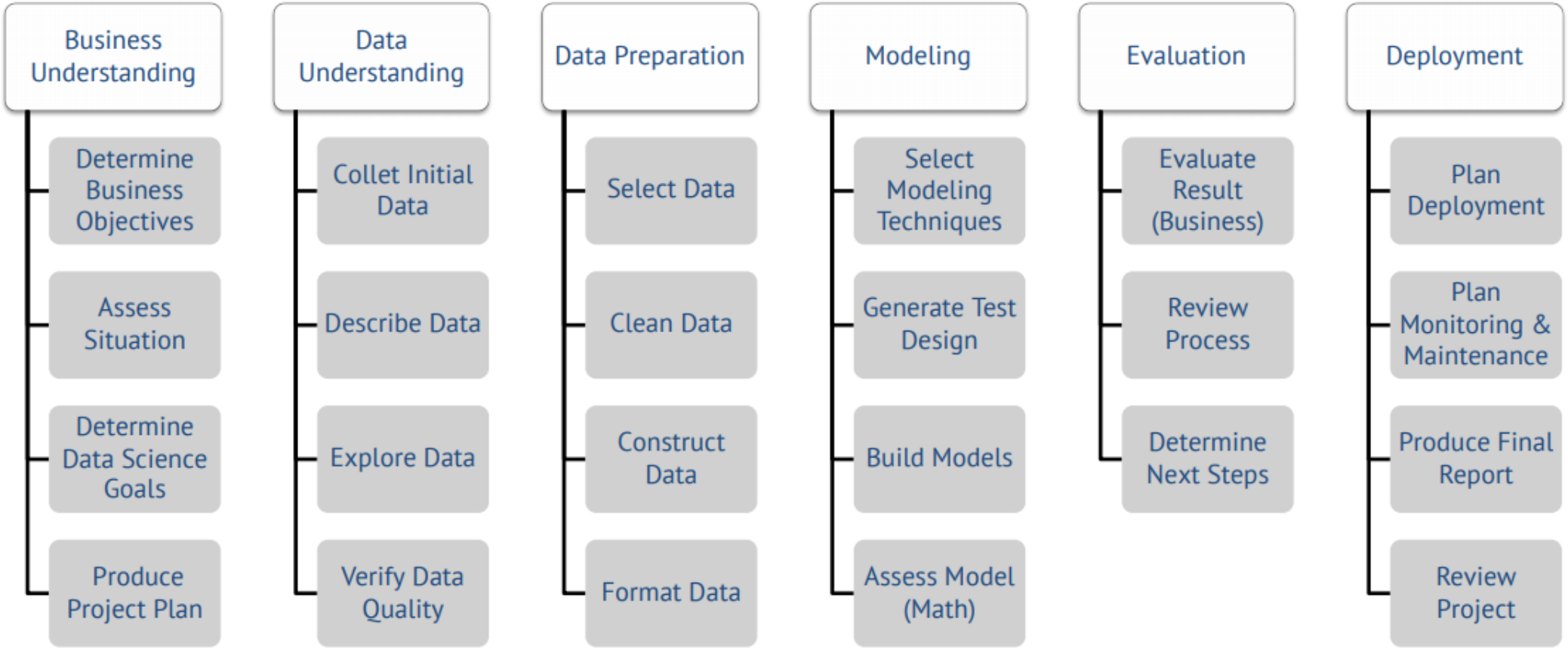
- The Cross Industry Standard Process for Data Mining (CRISP-DM) is an open standard process model that describes common approaches used by data mining experts. (2000)
- It is the most widely-used analytics model and can be used or adapted for data science tasks
- The sequence of phases is highly iterative and not a stringent step by step approach
- It aligns well with agile development principles

Picture source: CRISP-DM, http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

CRISP-DM breaks the process of data mining into six major phases

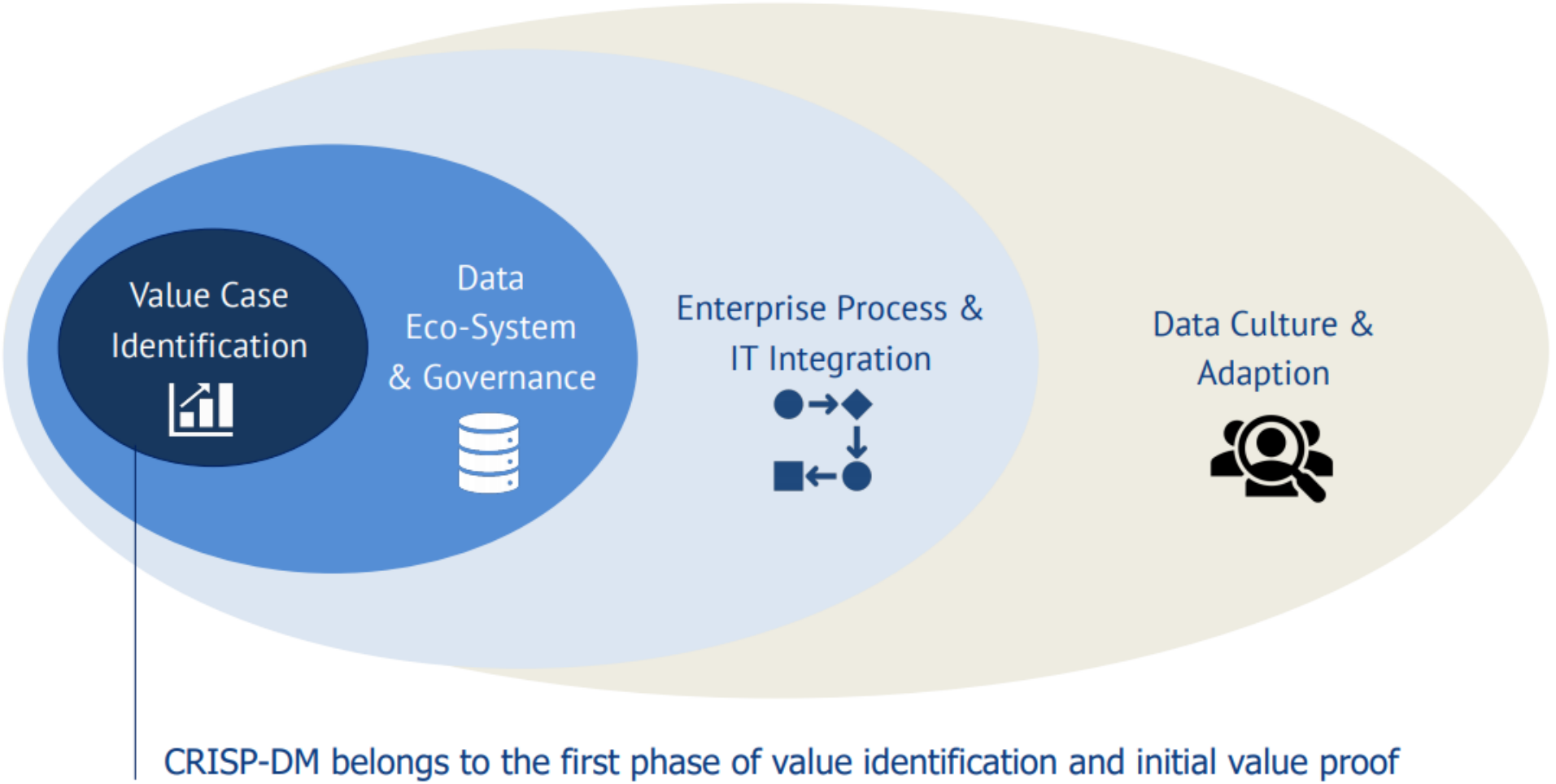


The CRISP-DM describes the high-level tasks of each phase.



Source: CRISP-DM 1.0 Step-by-step data mining guide, 2000, e.g. https://www.the-modeling-agency.com/crisp-dm.pdf

Delivering value through data in an enterprise requires the focus on many different aspects



CRISP-DM belongs to the first phase of value identification and initial value proof

TOOLS

Amazon SageMaker · gradient by Paperspace · FLOYD · DOMINO DATA LAB

"All-in-one"

Data

- Versioning: Pachyderm, DVC, LIQUIDATA
- Labeling: figure eight, scale, Aquarium
- Processing: Spark, DAGSTER
- Exploration: pandas, RAPIDS, dbt
- Data Lake / Warehouse: snowflake, databricks
- Sources: S3, LOG, Parquet

Training/Evaluation

- Frameworks & Distributed Training: fast.ai, PYTORCH, RAY, K, HOROVOD
- Hyperparameter Tuning: SIGOPT, Determined AI, Weights & Biases, tune
- Experiment Management: Weights & Biases, comet, TensorBoard, Neptune, mlflow
- Resource Management: Determined AI, PY, slurm
- Software Engineering: jupyter, Streamlit, git
- Compute: CoreWeave, Lambda

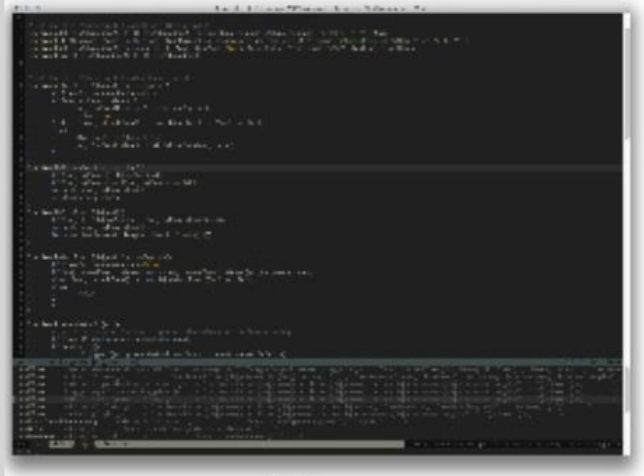
Deployment

- Feature Store: tecton
- Monitoring: fiddler
- Edge: ML, NVIDIA TensorRT, TensorFlow Lite, ONNX
- Web: ALGORITHMIA, SELDON
- CI / Testing: Buildkite, great_expectations
- or or

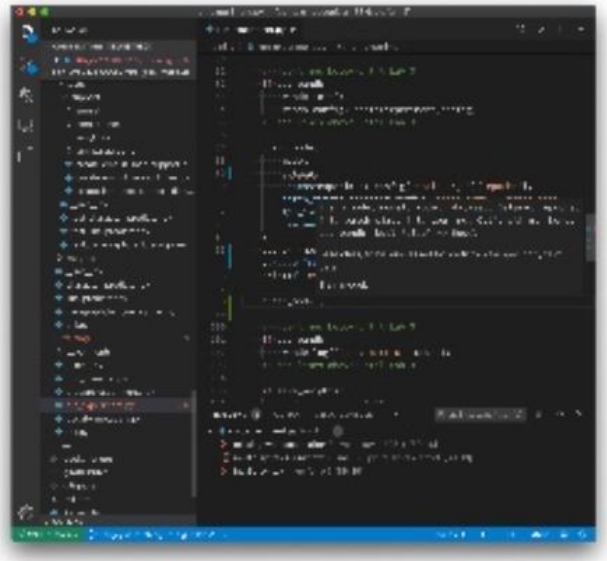
Programming Language

- Python, because of the libraries
  - Clear winner in scientific and data computing

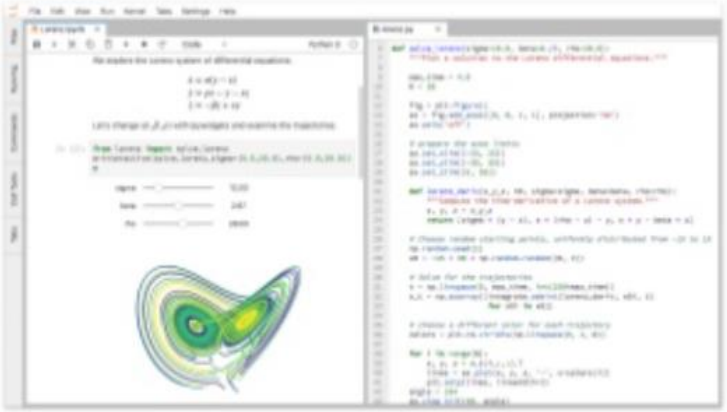
Editors



Vim



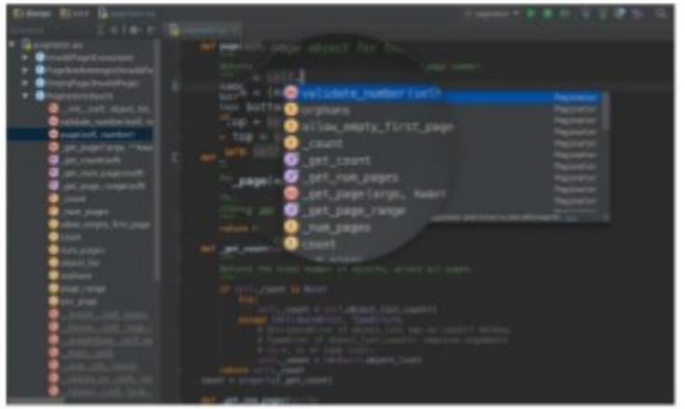
VS Code



Jupyter

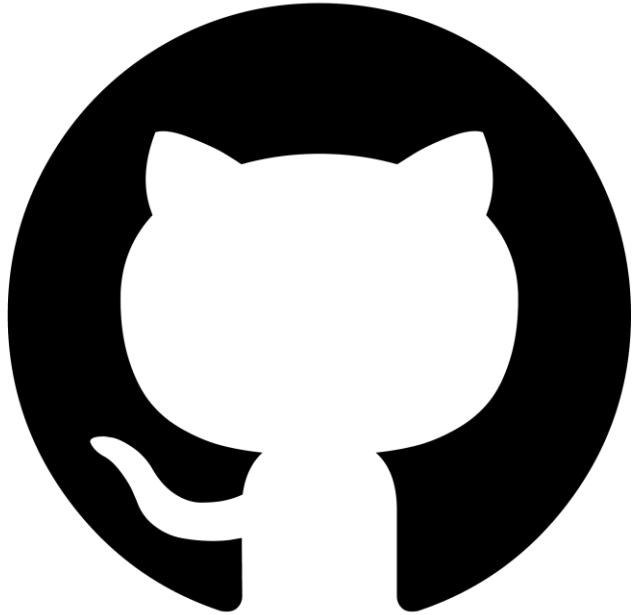
Emacs

replit

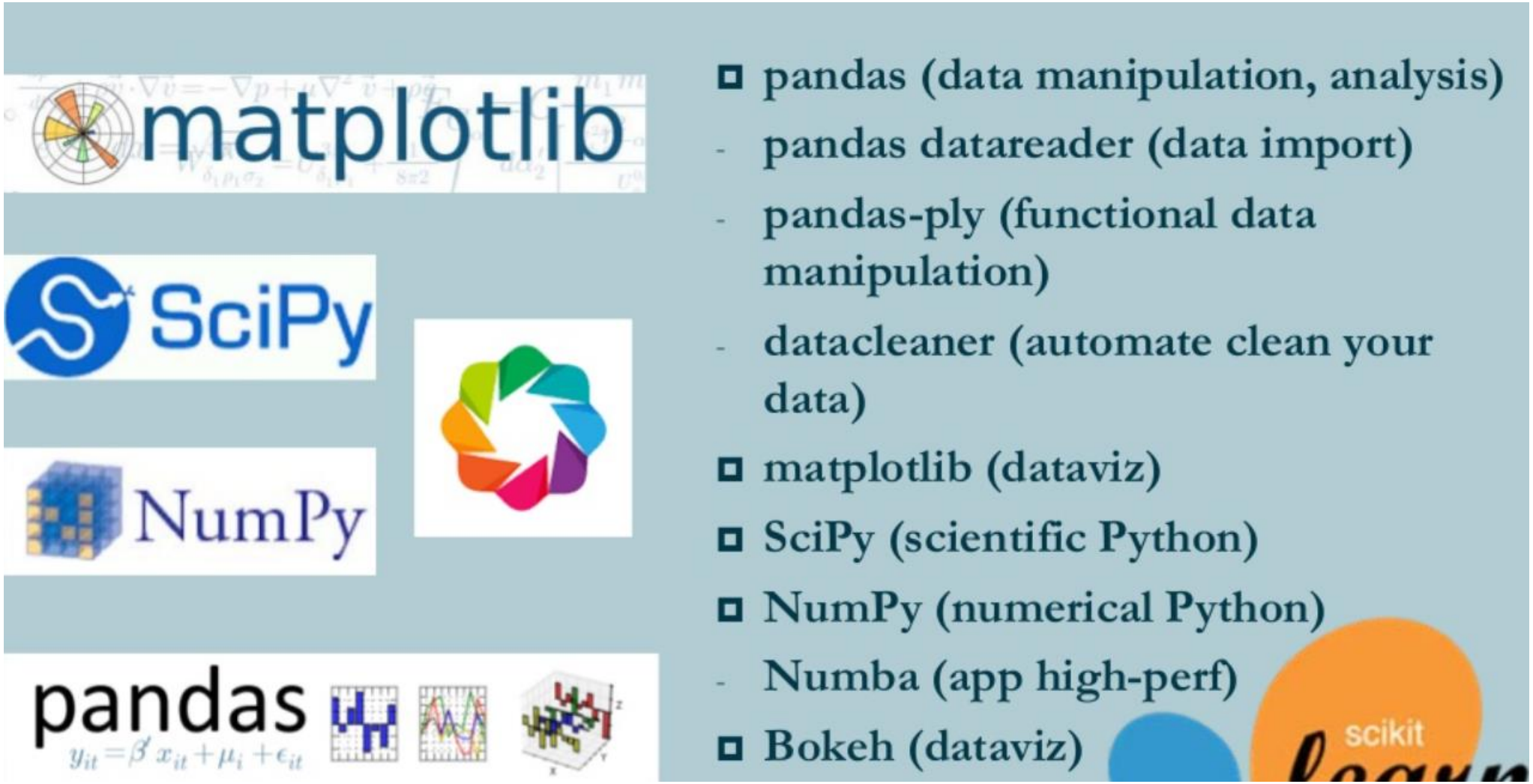


PyCharm

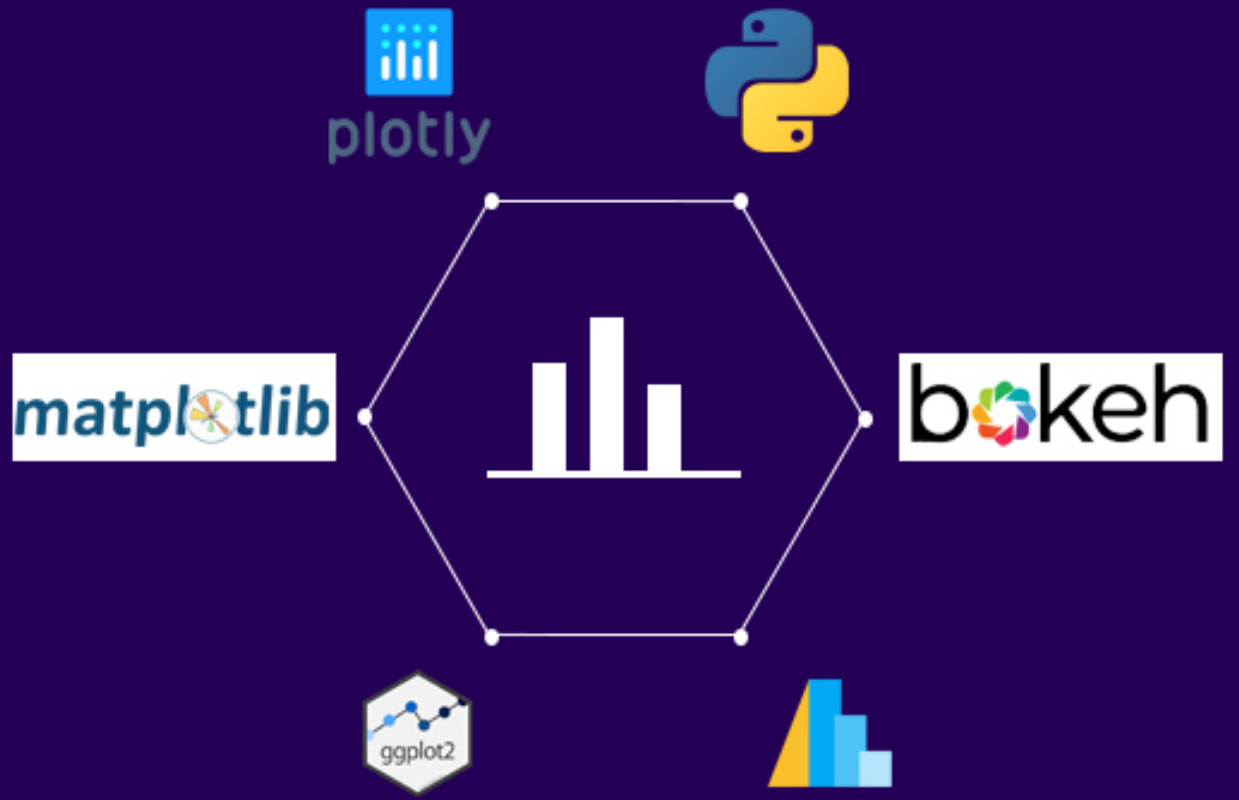
Version control (GIT & GITHUB)



DATA ANALYSIS (Pandas & NumPy)



DATA VISUALIZATION (Matplotlib & Plotly)



Maps & location (Mapbox and Folium)

DASHBOARD (STREAMLIT)

Total Outstanding Handovers

31

↑ 1 Compared to 1 hour ago

Current Handover Average

24 Mins

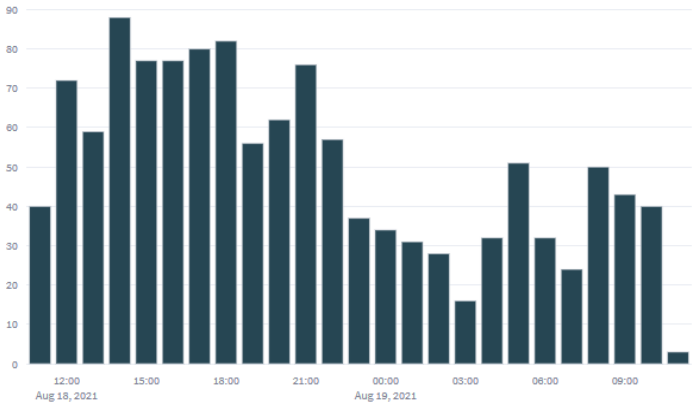
↓ -6 Compared to 1 hour ago

Time Lost today (Above 15 mins)

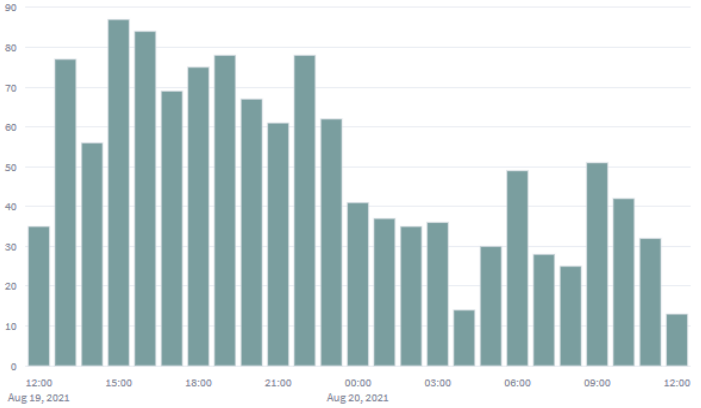
306 Hours

↓ -69 Compared to yesterday

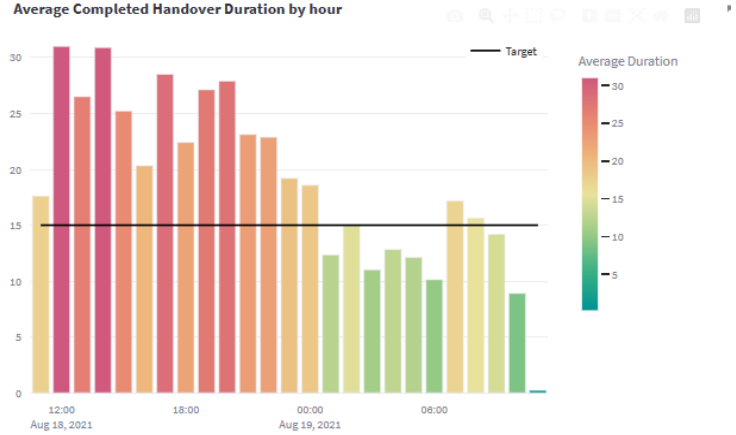
Number of Completed Handovers by Hour



Predicted Number of Arrivals



Average Completed Handover Duration by hour



Current Waiting Handovers

Hospital Attended	Expected	Inbound	Arrived	Waiting	0 - 15 Mins	15 - 30 Mins	30 - 60 Mins	60 - 90 Mins	90 + Mins
Ant NHS Foundation Trust	1	0	1	4	4	0	0	0	0
Bear NHS Foundation Trust	5	0	0	0	0	0	0	0	0
Bumblebee NHS Foundation Trust	3	4	1	2	1	0	0	0	1
Camel NHS Foundation Trust	1	0	0	2	0	1	0	0	1
Cat NHS Foundation Trust	4	0	0	1	0	1	0	0	0
Dog NHS Foundation Trust	0	3	1	1	1	0	0	0	0
Elephant NHS Foundation Trust	1	1	0	4	1	2	0	1	0
Fox NHS Foundation Trust	2	2	0	2	0	1	0	0	1
Gecko NHS Foundation Trust	2	1	0	1	0	0	0	1	0
Goldfish NHS Foundation Trust	3	0	2	0	0	0	0	0	0
Goose NHS Foundation Trust	2	3	0	0	0	0	0	0	0
Hedgehog NHS Foundation Trust	5	0	0	3	0	3	0	0	0
Mole NHS Foundation Trust	2	1	0	0	0	0	0	0	0
Moose NHS Foundation Trust	1	3	0	3	2	0	0	1	0
Ostrich NHS Foundation Trust	4	0	0	1	0	1	0	0	0
Parrot NHS Foundation Trust	6	2	0	2	1	1	0	0	0
Shark NHS Foundation Trust	2	1	1	1	0	1	0	0	0
Squirrel NHS Foundation Trust	1	3	0	2	0	2	0	0	0
Starling NHS Foundation Trust	4	1	1	0	0	0	0	0	0
Whale NHS Foundation Trust	2	0	1	2	0	0	0	2	0
Grand Total	51	25	8	31	10	13	0	5	3

Current Waiting Callsigns

Callsign	Hospital Attended	Arrived	Duration
7981	Fox NHS Foundation Trust	19/08/2021 08:46	02:33:32
6364	Camel NHS Foundation Trust	19/08/2021 08:53	02:26:40
120	Bumblebee NHS Foundation Trust	19/08/2021 09:40	01:39:39
9040	Elephant NHS Foundation Trust	19/08/2021 09:56	01:23:42
3279	Moose NHS Foundation Trust	19/08/2021 10:06	01:13:17
7837	Whale NHS Foundation Trust	19/08/2021 10:12	01:07:19
7054	Gecko NHS Foundation Trust	19/08/2021 10:12	01:07:16
1449	Whale NHS Foundation Trust	19/08/2021 10:17	01:02:21
1808	Hedgehog NHS Foundation Trust	19/08/2021 10:52	00:28:13
8780	Elephant NHS Foundation Trust	19/08/2021 10:52	00:27:34
6425	Ostrich NHS Foundation Trust	19/08/2021 10:54	00:25:49
1387	Fox NHS Foundation Trust	19/08/2021 10:55	00:25:08
5905	Hedgehog NHS Foundation Trust	19/08/2021 10:57	00:23:11
3285	Squirrel NHS Foundation Trust	19/08/2021 10:57	00:22:51
2300	Elephant NHS Foundation Trust	19/08/2021 10:57	00:22:43
5321	Shark NHS Foundation Trust	19/08/2021 10:57	00:22:38
2256	Cat NHS Foundation Trust	19/08/2021 11:00	00:19:48
9119	Hedgehog NHS Foundation Trust	19/08/2021 11:01	00:18:46
2738	Parrot NHS Foundation Trust	19/08/2021 11:02	00:17:34
1649	Squirrel NHS Foundation Trust	19/08/2021 11:03	00:16:57
2725	Camel NHS Foundation Trust	19/08/2021 11:03	00:16:18

MACHINE LEARNING (Sklearn)

DEEP LEARNING

Natural language processing

Natural
Language
ToolKit

Set up

1. Google Colabs

   https://www.kdnuggets.com/2020/06/google-colab-deep-learning.html

2. Anaconda

   https://docs.anaconda.com/anaconda/install/

AFRICA DATA SCHOOL

Read more

1)CRISP-DM:

https://www.the-modeling-agency.com/crisp-dm.pdf

2)What’s the Difference Between a Data Scientist, Research Scientist, and an Applied Scientist?

https://towardsdatascience.com/whats-the-difference-between-a-data-scientist-research-scientist-and-an-applied-scientist-30c04190c1fa

3)How the World’s Biggest Companies Design Machine Learning-Powered Applications

https://towardsdatascience.com/how-the-worlds-biggest-companies-design-machine-learning-powered-applications-701f4114e089

4) Beyond Interactive: Notebook Innovation at Netflix

https://netflixtechblog.com/notebook-innovation-591ee3221233