AFRICA DATA SCHOOL

DATA SCIENCE FRAMEWORK & TOOLS

AGENDA

- Applied Data Science
- CRISP -DM Project Framework
- How the World’s Biggest Companies Design Machine Learning-Powered Applications
- 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.

Applied Data Science

Applied data scientists

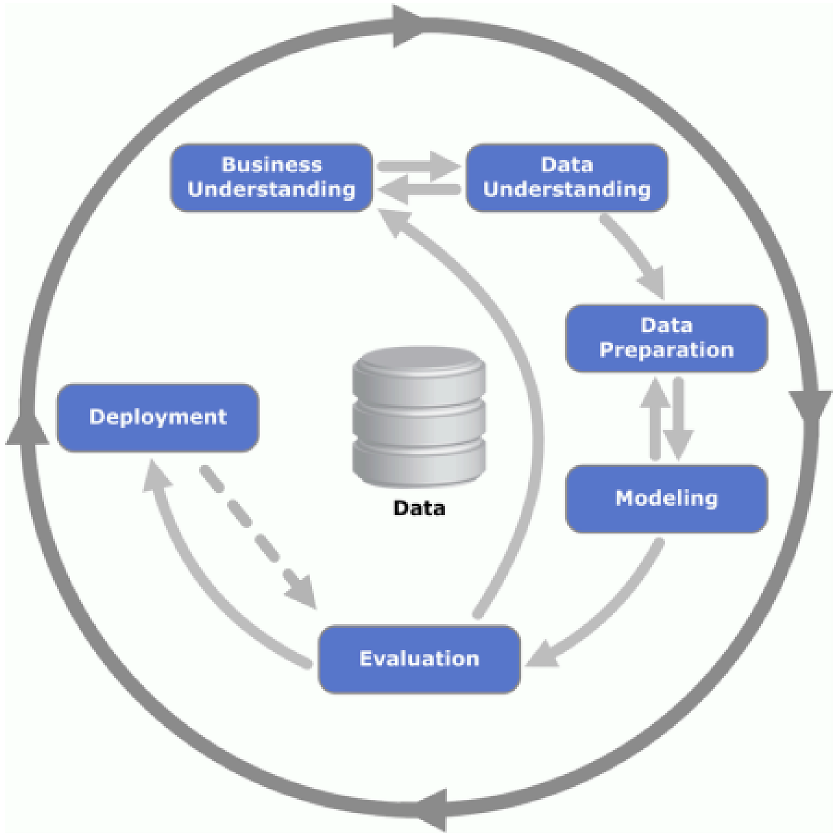
- Applied scientists use Data science and ML to improve business
  outcomes (e.g., revenue, cost, customer experience).
- The systems they build may be internal (e.g., product classification, fraud detection) or customer-facing (e.g., search,
  recommendations).
- Outside of use-case driven applications, 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.

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.

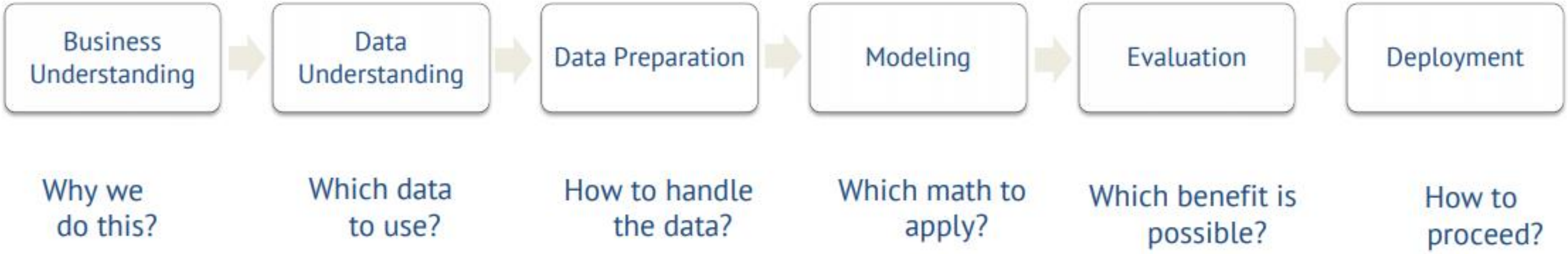
The Cross Industry Standard Process for Data Mining (CRISP-DM)



- 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

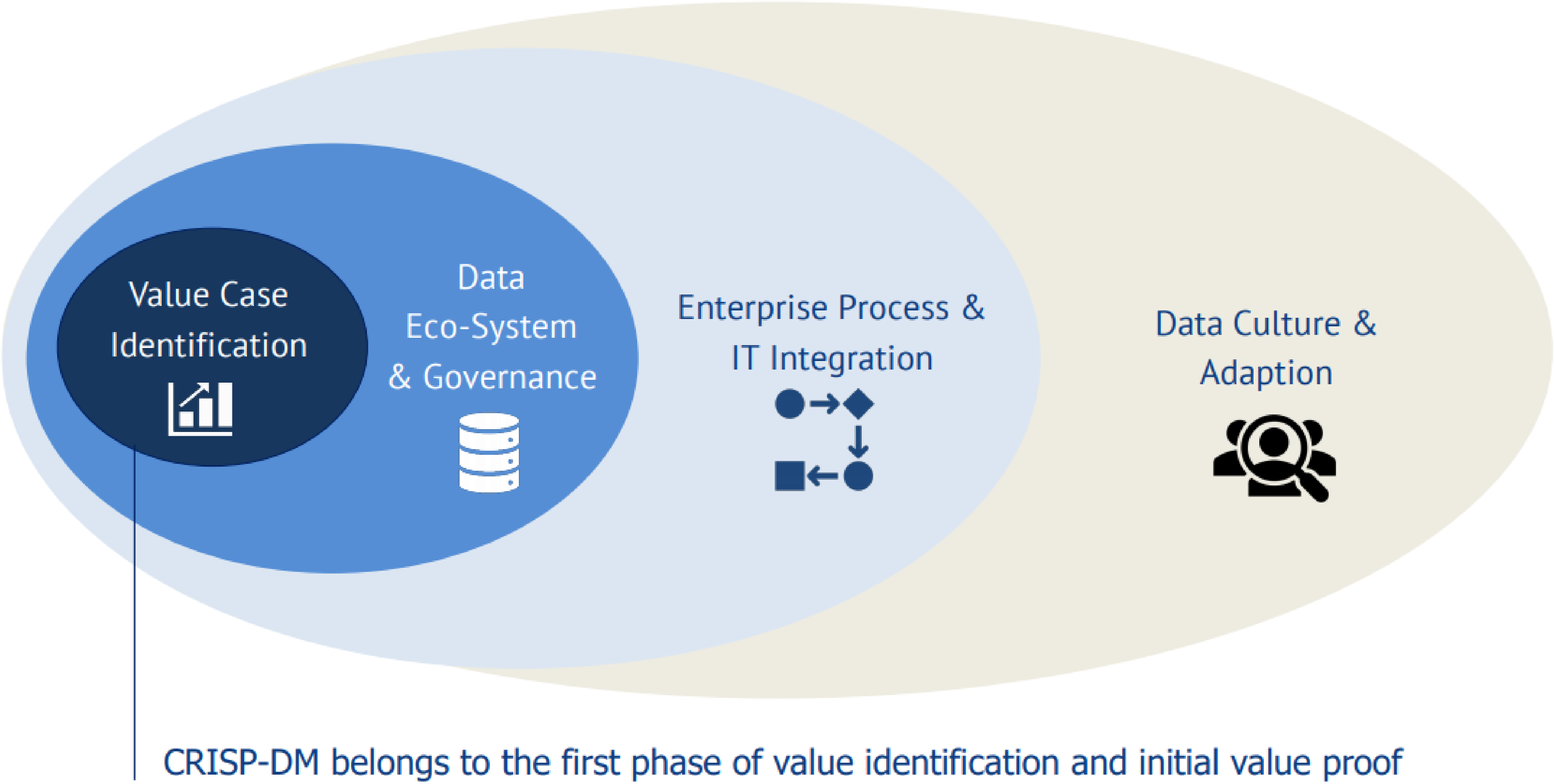


The CRISP-DM describes the high-level tasks of each phase.

- Business Understanding
  - Determine Business Objectives
  - Assess Situation
  - Determine Data Science Goals
  - Produce Project Plan
- Data Understanding
  - Collet Initial Data
  - Describe Data
  - Explore Data
  - Verify Data Quality
- Data Preparation
  - Select Data
  - Clean Data
  - Construct Data
  - Format Data
- Modeling
  - Select Modeling Techniques
  - Generate Test Design
  - Build Models
  - Assess Model (Math)
- Evaluation
  - Evaluate Result (Business)
  - Review Process
  - Determine Next Steps
- Deployment
  - Plan Deployment
  - Plan Monitoring & Maintenance
  - Produce Final Report
  - Review Project

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

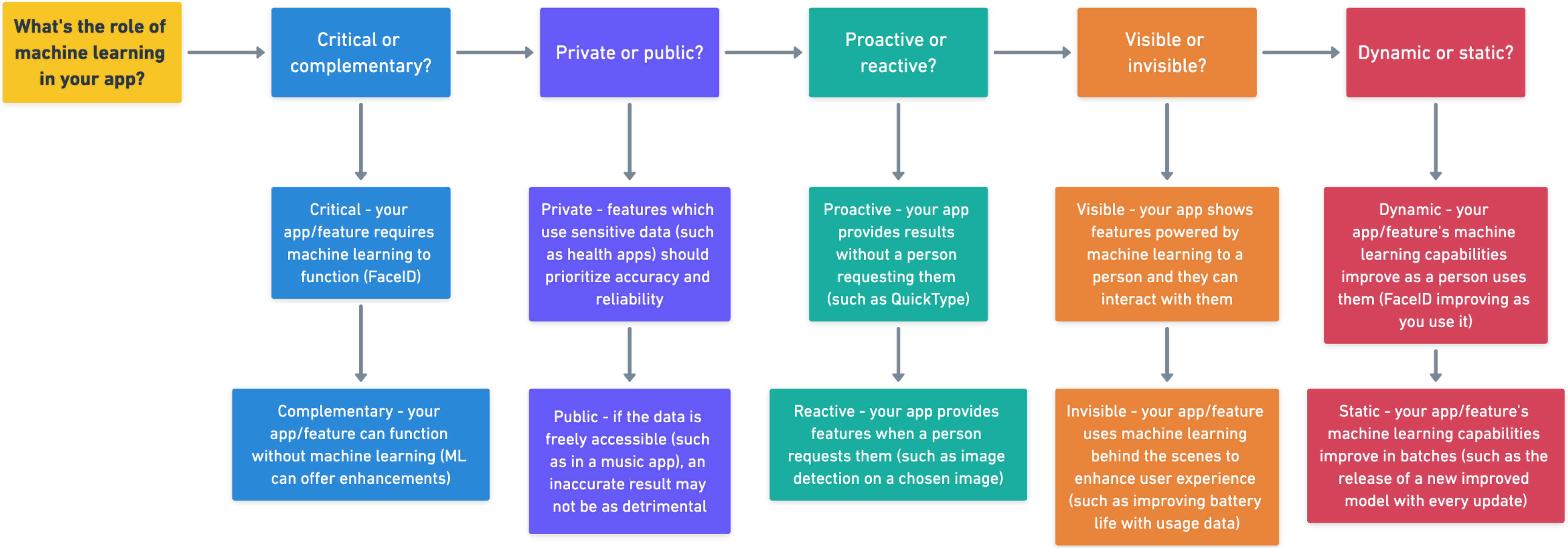


CRISP-DM belongs to the first phase of value identification and initial value proof

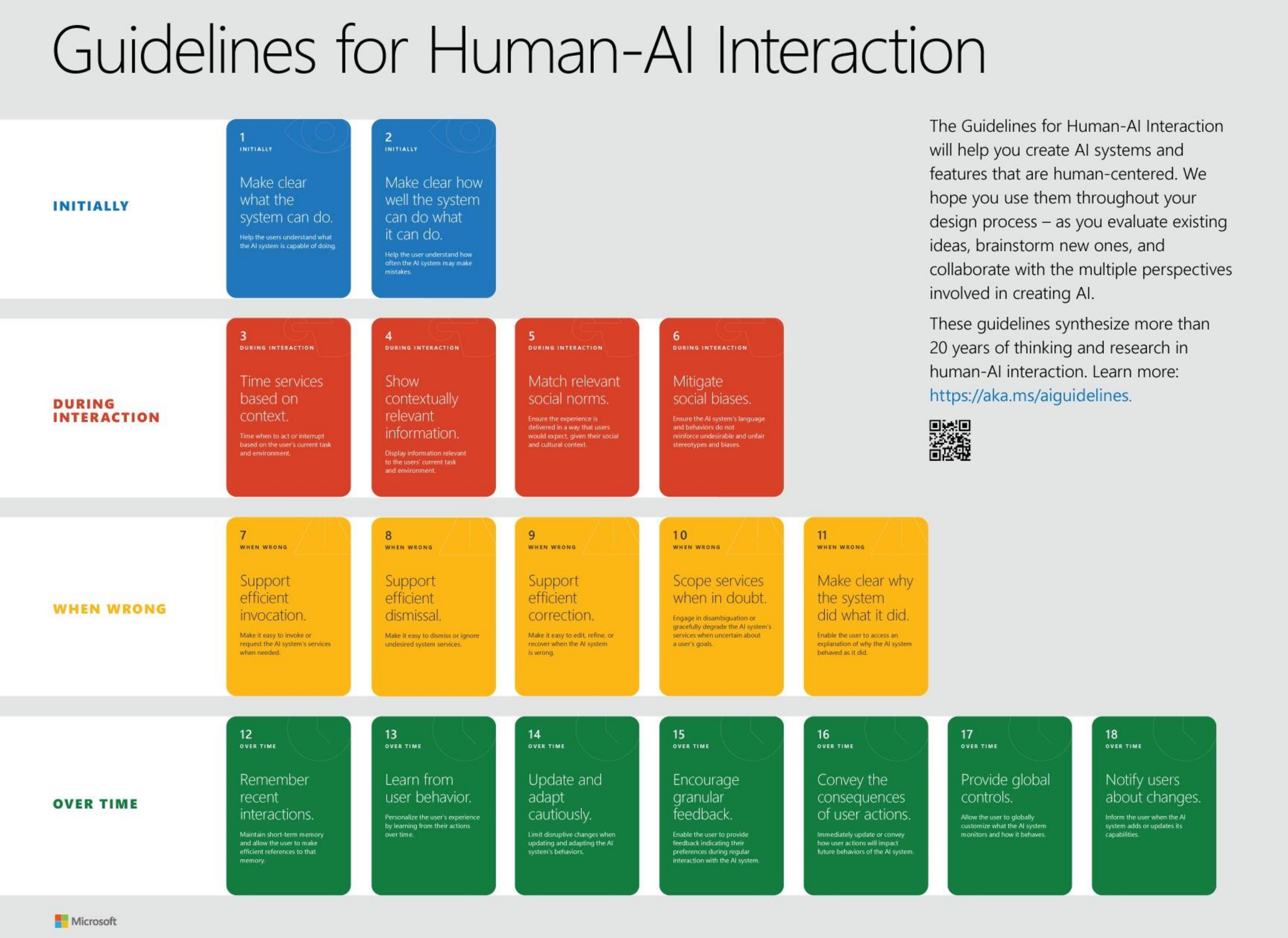
APPLE

Apple's Human Interface Guidelines for Machine Learning

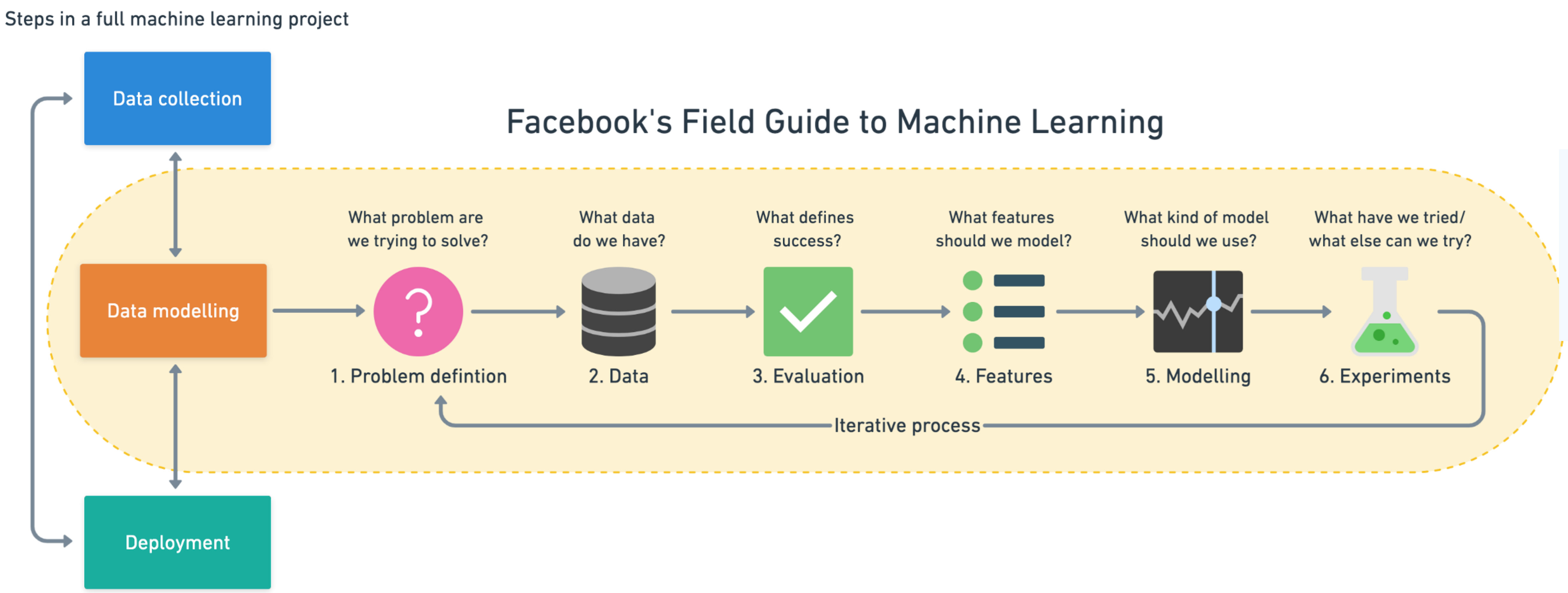
Introduction - Machine Learning - Human Interface Guidelines - Apple Developer



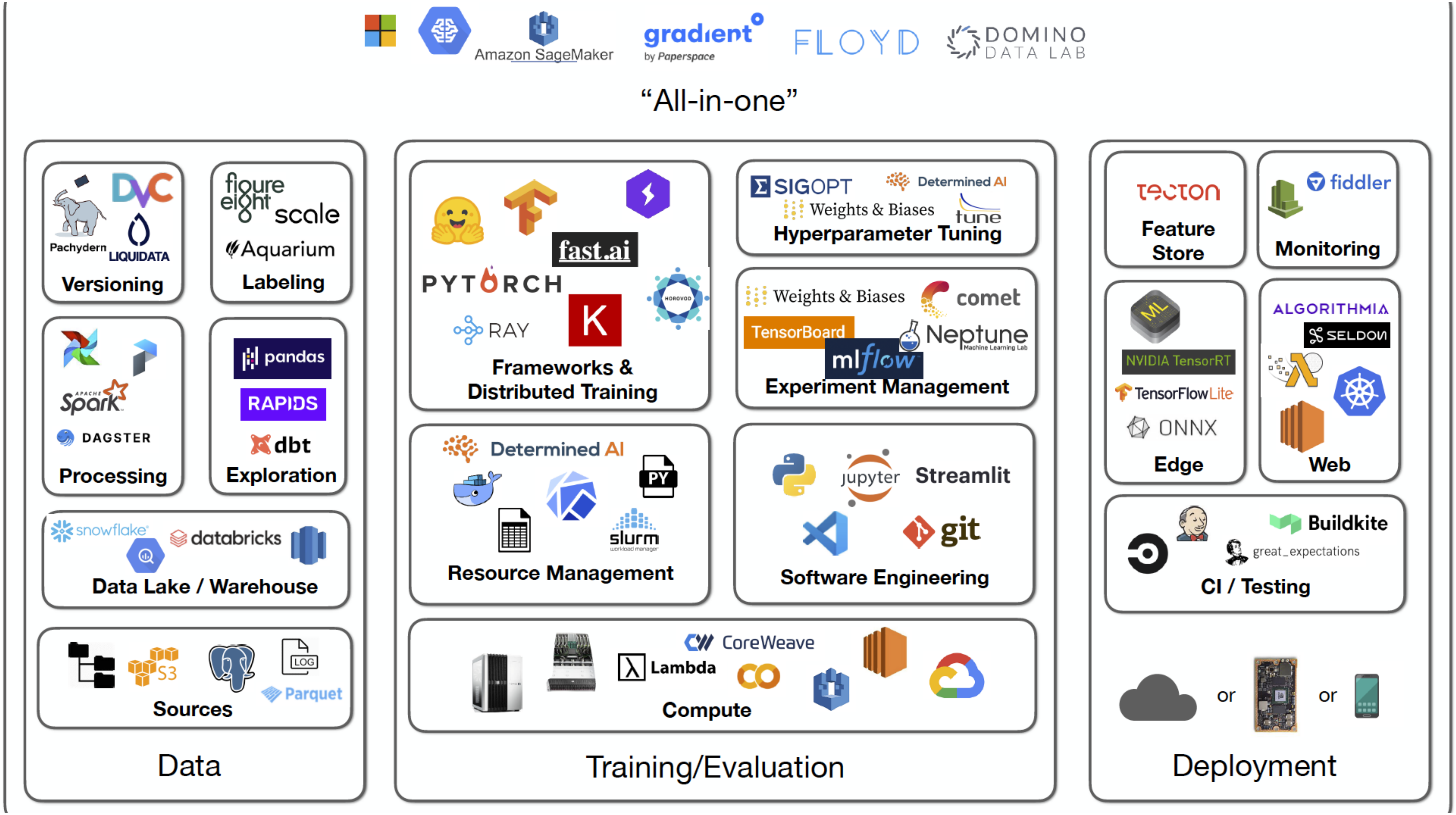
MICROSOF


Guidelines for Human-AI Interaction
INITIALLY
DURING INTERACTION
WHEN WRONG
OVER TIME
1 INITIALLY Make clear what the system can do. Help the users understand what the AI system is capable of doing.
2 INITIALLY Make clear how well the system can do what it can do. Help the user understand how often the AI system may make mistakes.
3 DURING INTERACTION Time services based on context. Time when to act or interrupt based on the user's current task and environment.
4 DURING INTERACTION Show contextually relevant information. Display information relevant to the users' current task and environment.
5 DURING INTERACTION Match relevant social norms. Ensure the experience is delivered in a way that users would expect, given their social and cultural context.
6 DURING INTERACTION Mitigate social biases. Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.
7 WHEN WRONG Support efficient invocation. Make it easy to invoke or request the AI system's services when needed.
8 WHEN WRONG Support efficient dismissal. Make it easy to dismiss or ignore undesired system services.
9 WHEN WRONG Support efficient correction. Make it easy to edit, refine, or recover when the AI system is wrong.
10 WHEN WRONG Scope services when in doubt. Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.
11 WHEN WRONG Make clear why the system did what it did. Enable the user to access an explanation of why the AI system behaved as it did.
12 OVER TIME Remember recent interactions. Maintain short-term memory and allow the user to make efficient references to that memory.
13 OVER TIME Learn from user behavior. Personalize the user's experience by learning from their actions over time.
14 OVER TIME Update and adapt cautiously. Limit disruptive changes when updating and adapting the AI system's behaviors.
15 OVER TIME Encourage granular feedback. Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.
16 OVER TIME Convey the consequences of user actions. Immediately update or convey how user actions will impact future behaviors of the AI system.
17 OVER TIME Provide global controls. Allow the user to globally customize what the AI system monitors and how it behaves.
18 OVER TIME Notify users about changes. Inform the user when the AI system adds or updates its capabilities.
The Guidelines for Human-AI Interaction will help you create AI systems and features that are human-centered. We hope you use them throughout your design process – as you evaluate existing ideas, brainstorm new ones, and collaborate with the multiple perspectives involved in creating AI.
These guidelines synthesize more than 20 years of thinking and research in human-AI interaction. Learn more: https://aka.ms/aiguidelines.
Microsoft

Facebook


Steps in a full machine learning project
Data collection
Data modelling
Deployment
Facebook's Field Guide to Machine Learning
What problem are we trying to solve?
What data do we have?
What defines success?
What features should we model?
What kind of model should we use?
What have we tried/ what else can we try?
1. Problem defintion
2. Data
3. Evaluation
4. Features
5. Modelling
6. Experiments
Iterative process

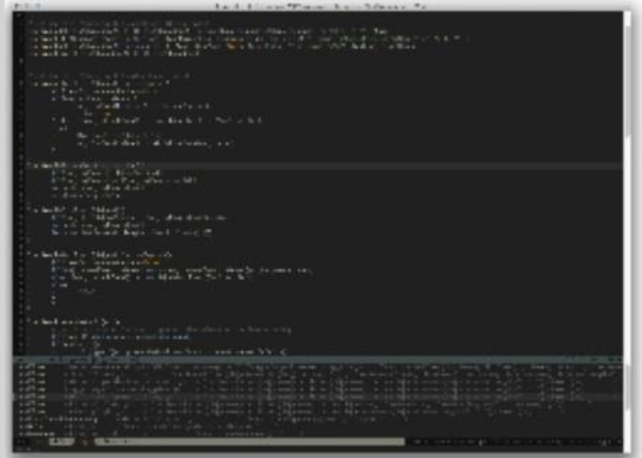
TOOLS


Amazon SageMaker
gradient
by Paperspace
FLOYD
DOMINO DATA LAB
“All-in-one”
Pachyderm
DVC
LIQUIDATA
Versioning
figure eight
scale
Aquarium
Labeling
Spark
DAGSTER
Processing
pandas
RAPIDS
dbt
Exploration
snowflake
databricks
Data Lake / Warehouse
S3
LOG
Parquet
Sources
Data
fast.ai
PYTORCH
RAY
K
HOROVOD
Frameworks &
Distributed Training
SIGOPT
Determined AI
Weights & Biases
tune
Hyperparameter Tuning
Weights & Biases
comet
TensorBoard
Neptune
Machine Learning Lab
mlflow
Experiment Management
Determined AI
PY
slurm
workload manager
Resource Management
jupyter
Streamlit
git
Software Engineering
CoreWeave
Lambda
Compute
Training/Evaluation
tecton
Feature Store
fiddler
Monitoring
ML
NVIDIA TensorRT
TensorFlowLite
ONNX
Edge
ALGORITHMIA
SELDON
Web
Buildkite
great_expectations
CI / Testing
or
or
Deployment

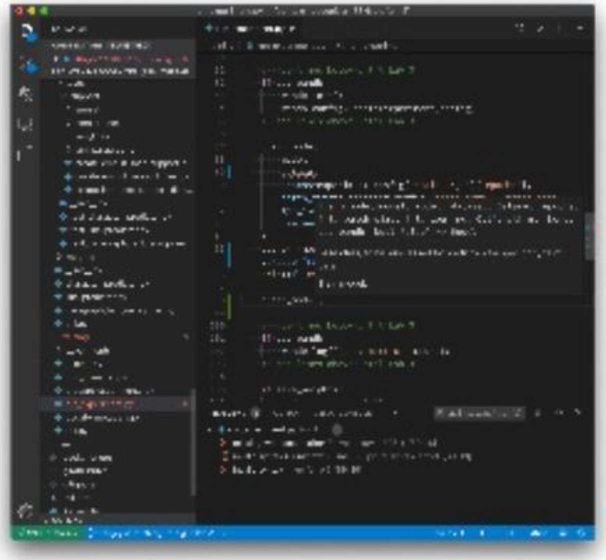
Programming Language

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

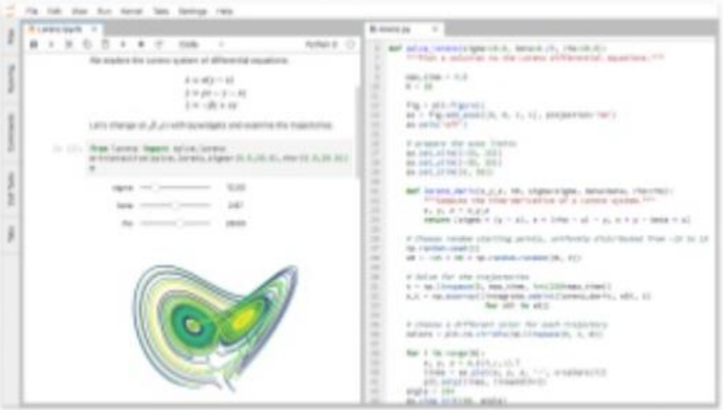
Editors



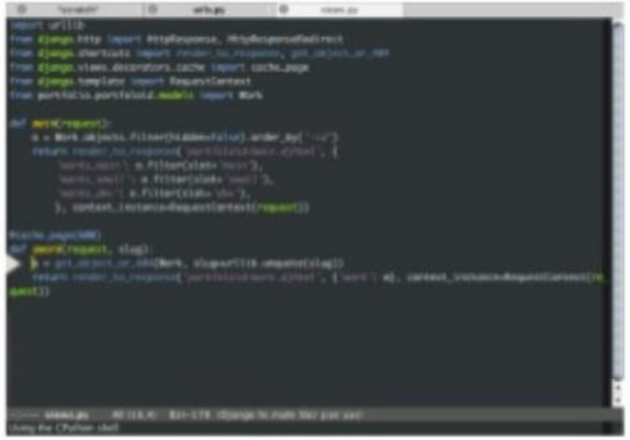
Vim



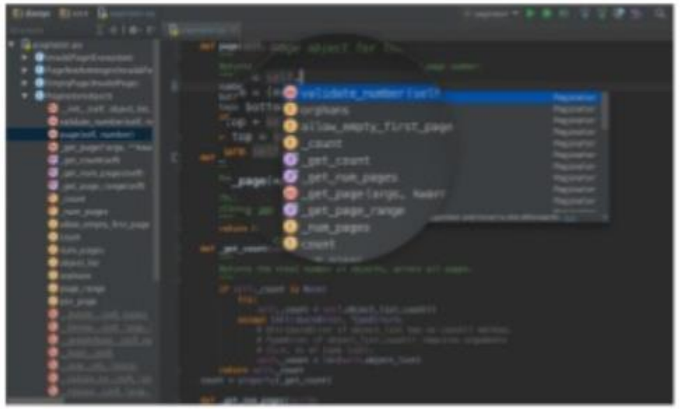
VS Code



Jupyter



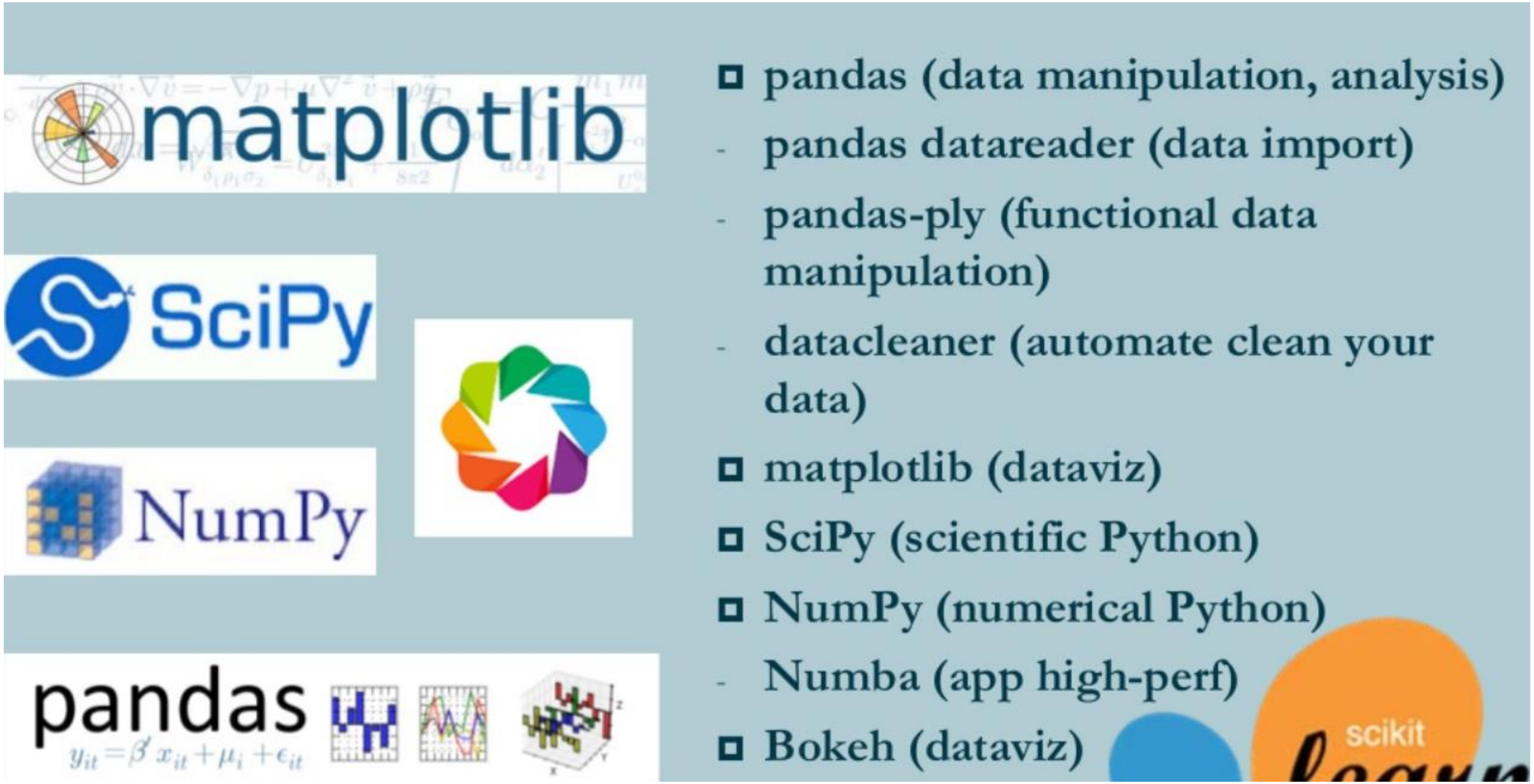
Emacs



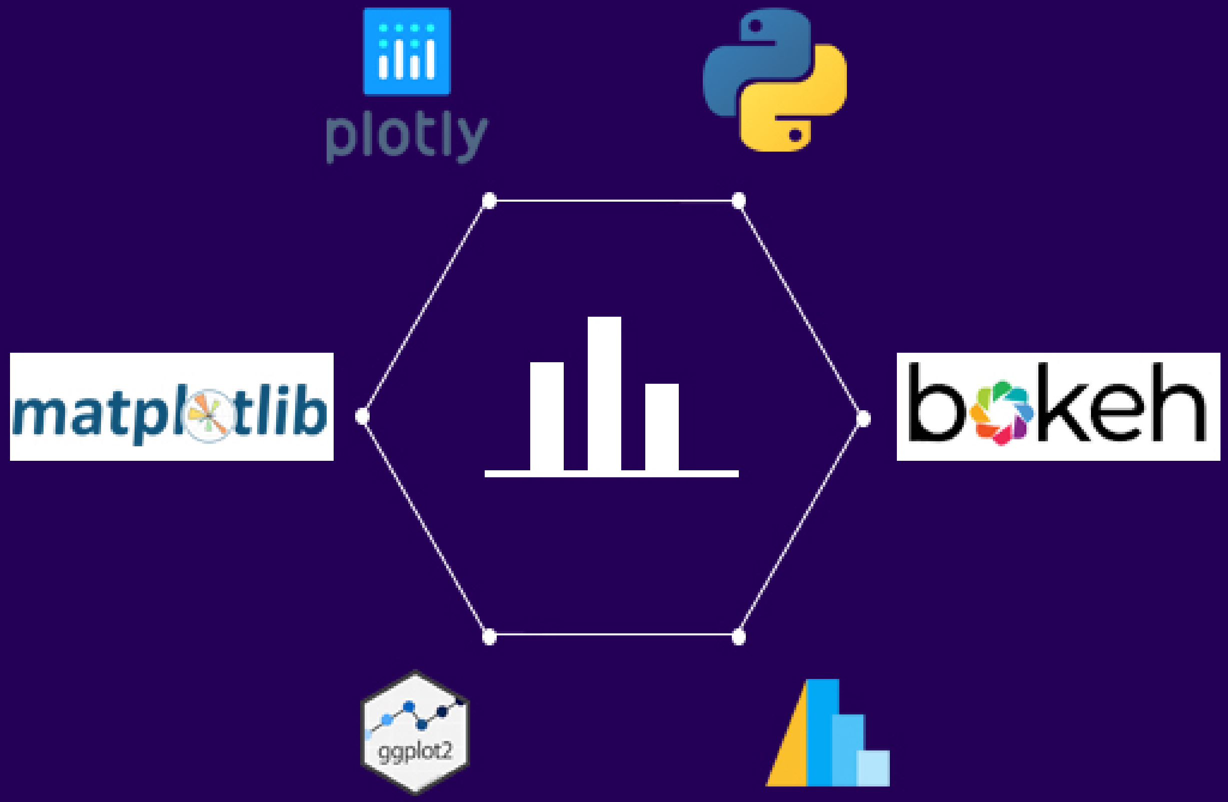
PyCharm

Version control (GIT & GITHUB)

DATA ANALYSIS (Pandas & NumPy)

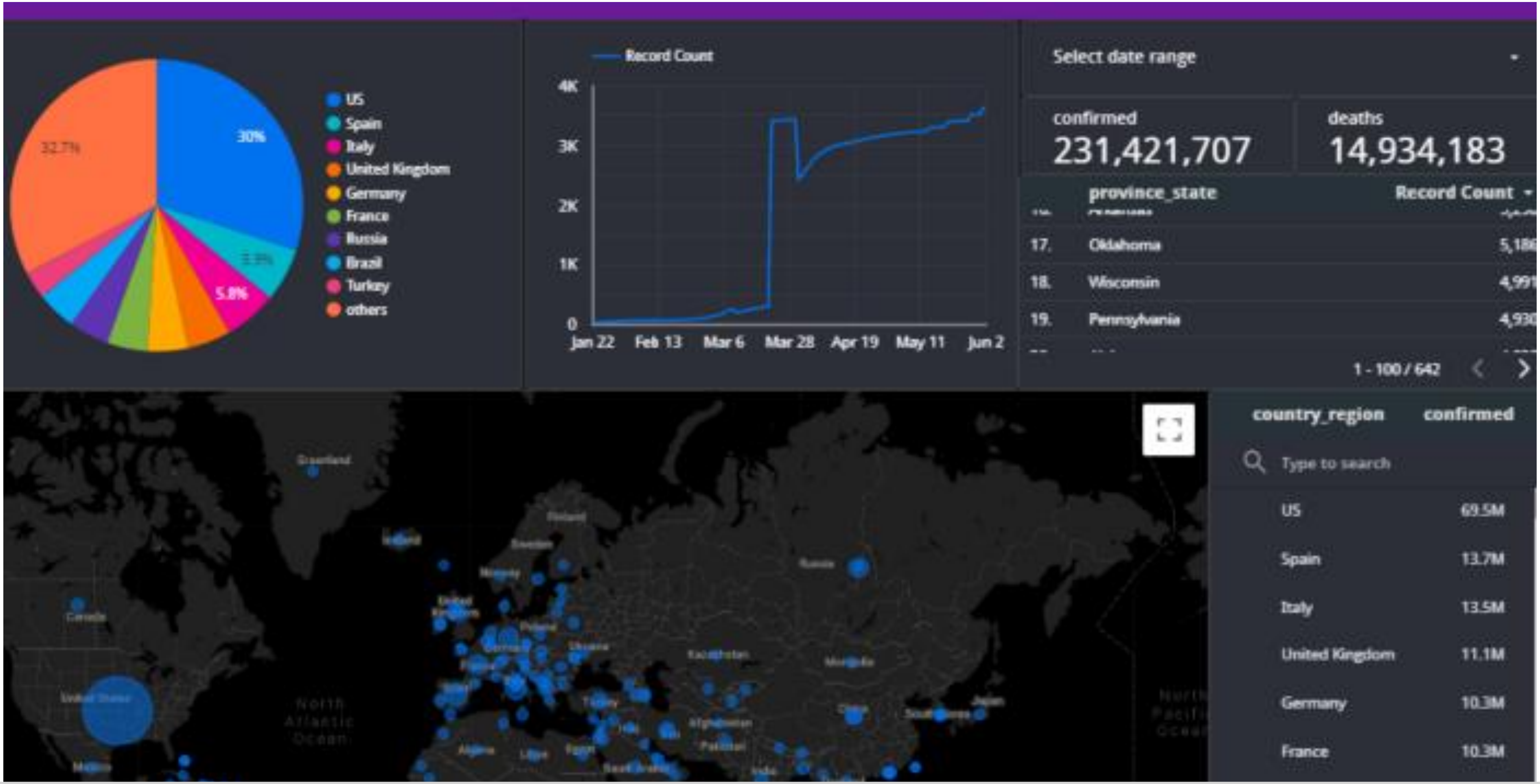


DATA VISUALIZATION (Matplotlib, Seaborn& Plotly)



Maps & location (Mapbox)

DASHBOARD (DASH & STREAMLIT)



MACHINE LEARNING (Sklearn)

DEEP LEARNING

TensorFlow

Keras

Natural language processing

spaCy

Set up

1 | Google Colabs

2 | Anaconda

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