An Introduction

Project Lifecycle

The Alan Turing Institute

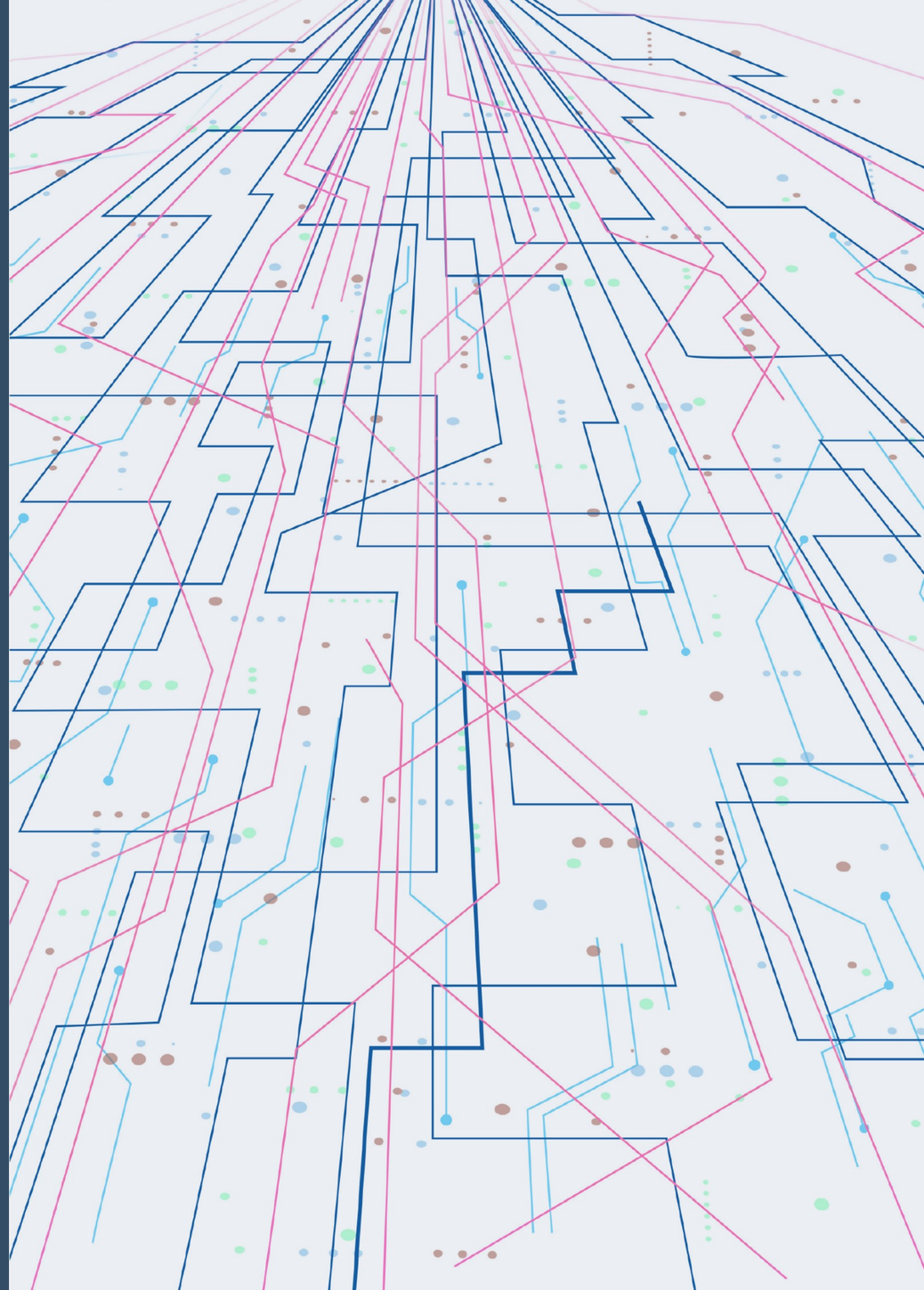
Presentation

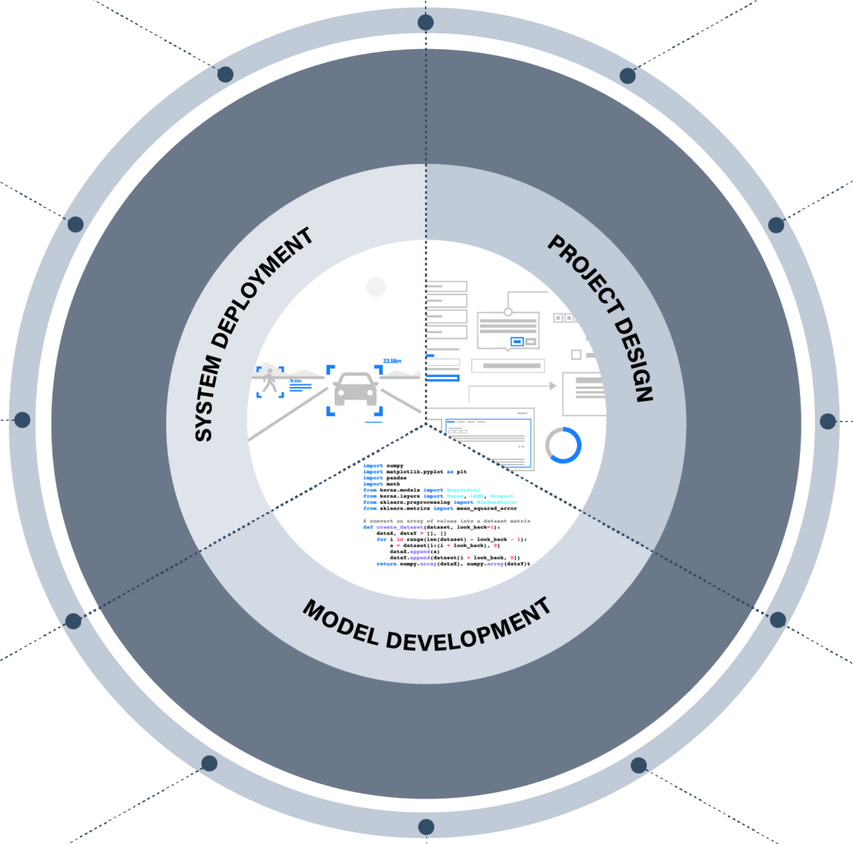
Overview

- Introducing the ML/AI Project
  - (Project) Design
  - (Model) Development
  - (System) Deployment
- Operationalising Ethical Principles
  - SAFE-D Principles

Section 1

Introducing the ML/AI Project Lifecycle





Project Planning

Problem Formulation

Data Extraction & Procurement

Data Analysis

Preprocessing & Feature Engineering

Model Selection & Training

Model Testing & Validation

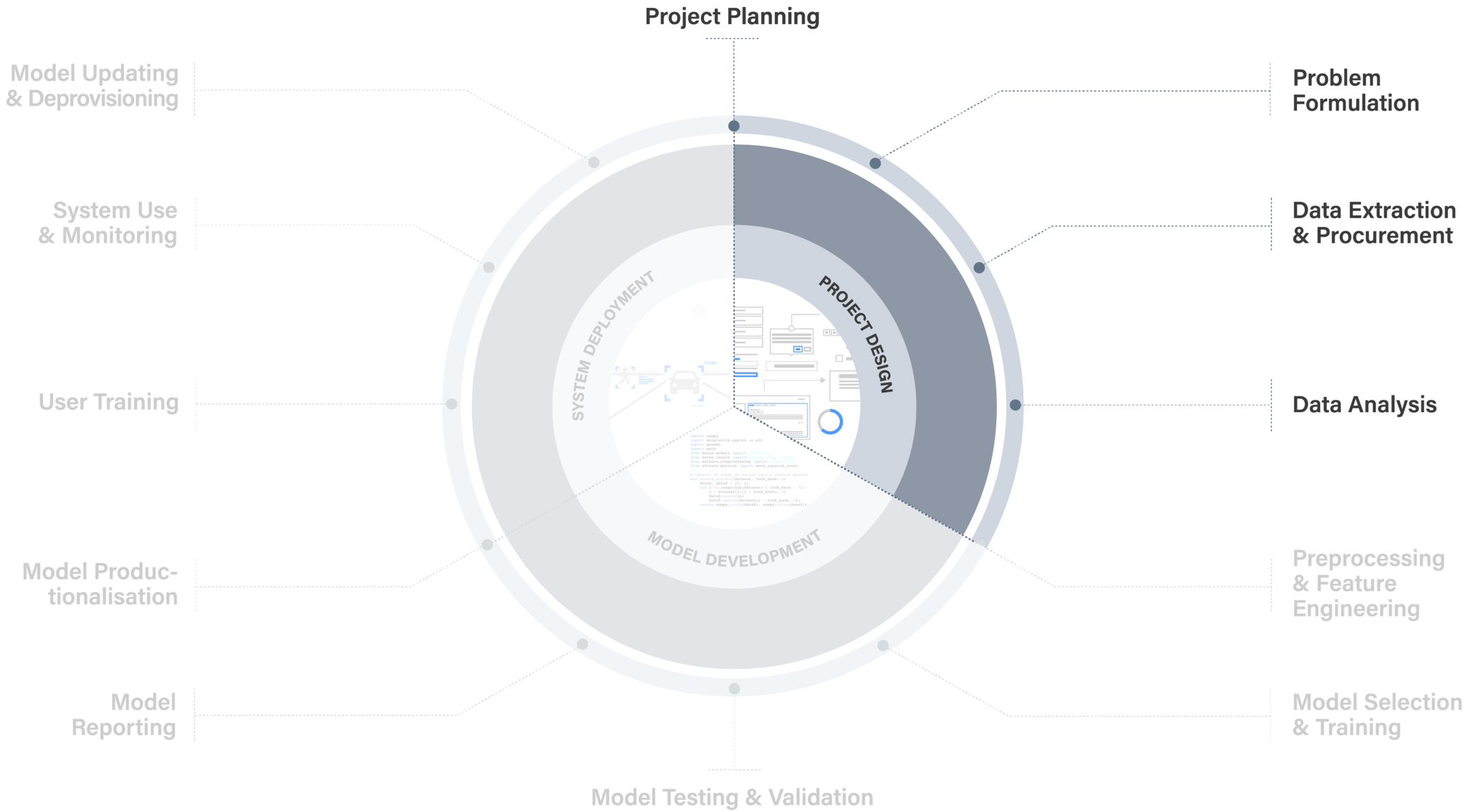
Model Reporting

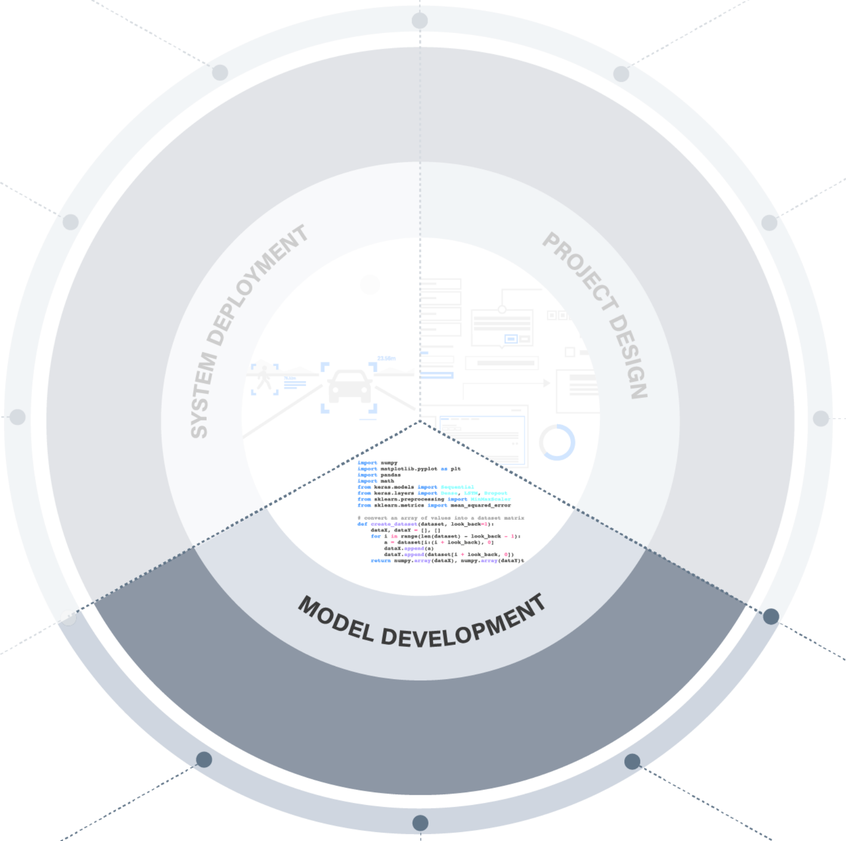
Model Productionalisation

User Training

System Use & Monitoring

Model Updating & Deprovisioning


Project Planning
Problem Formulation
Data Extraction & Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
Model Testing & Validation
Model Reporting
Model Productionalisation
User Training
System Use & Monitoring
Model Updating & Deprovisioning
PROJECT DESIGN
MODEL DEVELOPMENT
SYSTEM DEPLOYMENT



Project Planning

Model Updating & Deprovisioning

Problem Formulation

System Use & Monitoring

Data Extraction & Procurement

User Training

Data Analysis

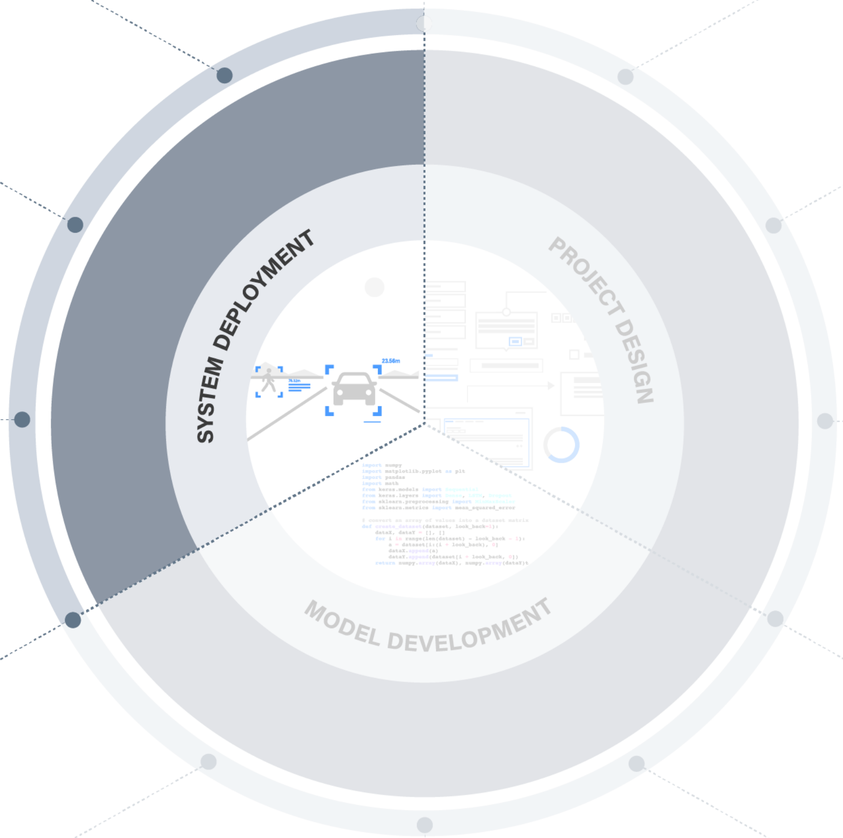
Model Produc-tionalisation

Preprocessing & Feature Engineering

Model Reporting

Model Selection & Training

Model Testing & Validation



Project Planning

Model Updating & Deprovisioning

System Use & Monitoring

User Training

Model Productionalisation

Model Reporting

Problem Formulation

Data Extraction & Procurement

Data Analysis

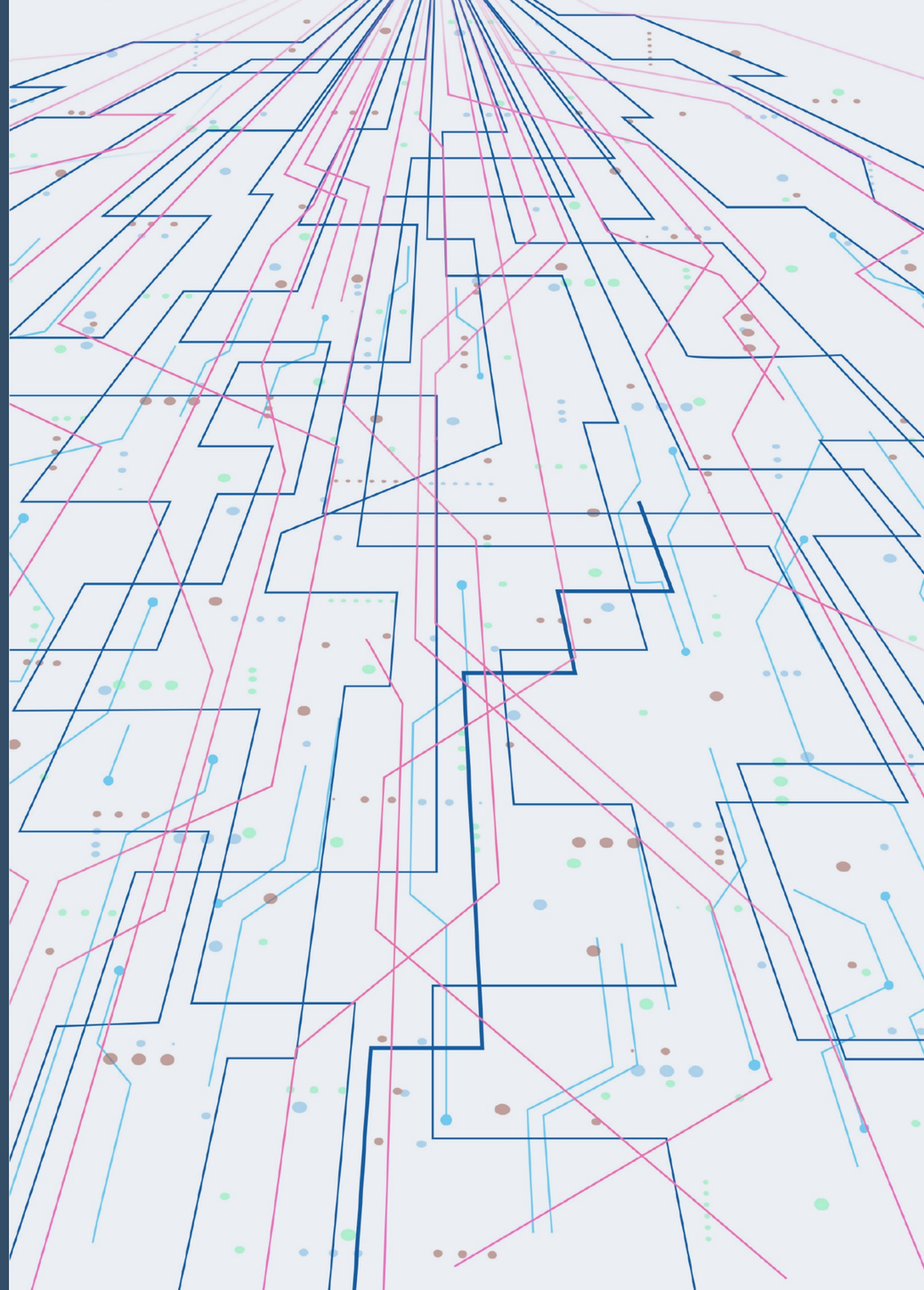
Preprocessing & Feature Engineering

Model Selection & Training

Model Testing & Validation

Section 2

Operationalising Ethical Principles



Operationalising

Ethical Principles

What is the role of ethical principles?

> **They serve to distil values and norms into action-guiding constraints for reflection and deliberation.**

However, they are often insufficient to determine actions without further specification.

Specification

SAFE-D Principles

The following principles, known as the SAFE-D principles have been designed to support the responsible governance of data science and AI:

- Sustainable
- Accountable
- Fair
- Explainable
- Data (Quality, Integrity and Protection)

But specifying and operationalising them can only be done in conjunction with stakeholders and domain experts.

SAFE-D Principles

Sustainability

Sustainability requires the outputs of a project to be:

- safe, secure, robust, and reliable
- informed by ongoing consideration of the risk of exposing individuals to harms even after the system has been deployed and the project completed—a long-term (or sustainable) form of safety.

Core Attributes

Sustainability

- Safety
- Security
- Robustness
- Reliability
- Accuracy and Performance

SAFE-D Principles

Accountability

Accountability can refer to transparency of processes and associated outcomes that enable people to understand how a project was conducted (e.g., project documentation), or why a specific decision was reached. But it can also refer to broader processes of responsible project governance that seek to establish clear roles of responsibility where full transparency may be inappropriate (e.g., confidential projects).

Core Attributes

Accountability

- Traceability
- Answerability
- Auditability
- Clear Data Provenance and Lineage
- Accessibility
- Reproducibility

SAFE-D Principles

Fairness

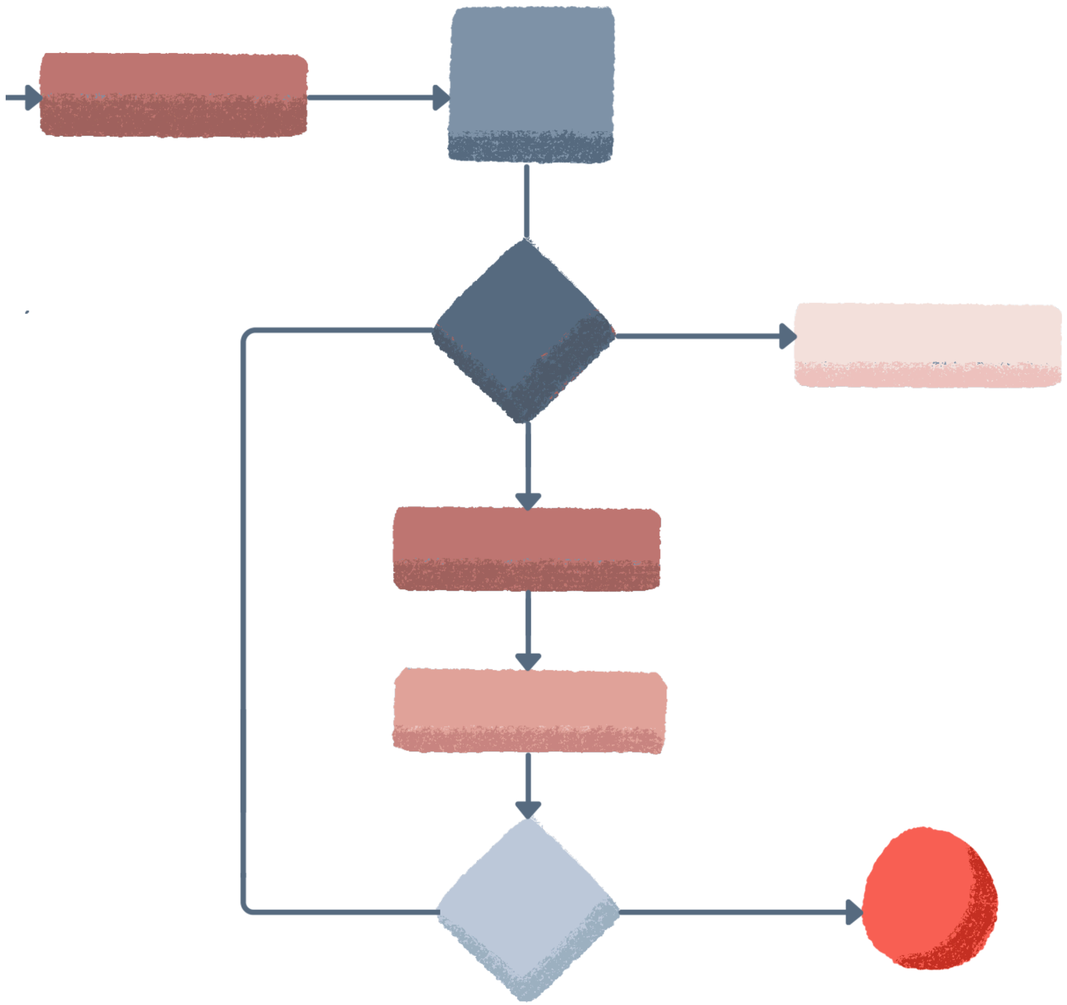
Fairness is inseparably connected with legal conceptions of equality and justice, which may emphasise a variety of features such as non-discrimination, equitable outcomes, or procedural fairness through bias mitigation.

However, these notions serve as a subset of broader normative considerations pertaining to social justice, socioeconomic capabilities, diversity and inclusivity.

Core Attributes

Fairness

- Bias Mitigation
- Diversity and Inclusiveness
- Non-Discrimination
- Equality

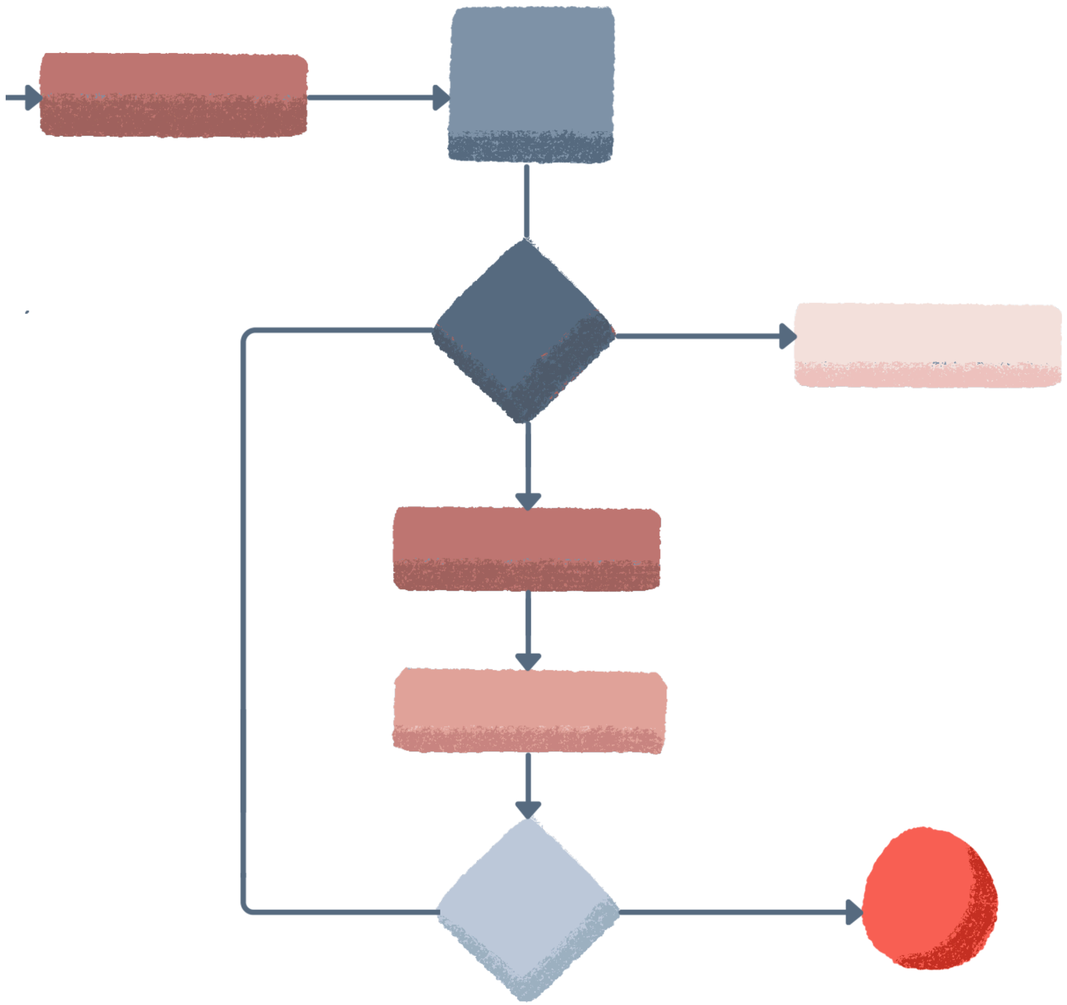


SAFE-D Principles

Explainability

Explainability is a key condition for autonomous and informed decision-making in situations where data-driven systems interact with or influence human judgement and choice behaviour.

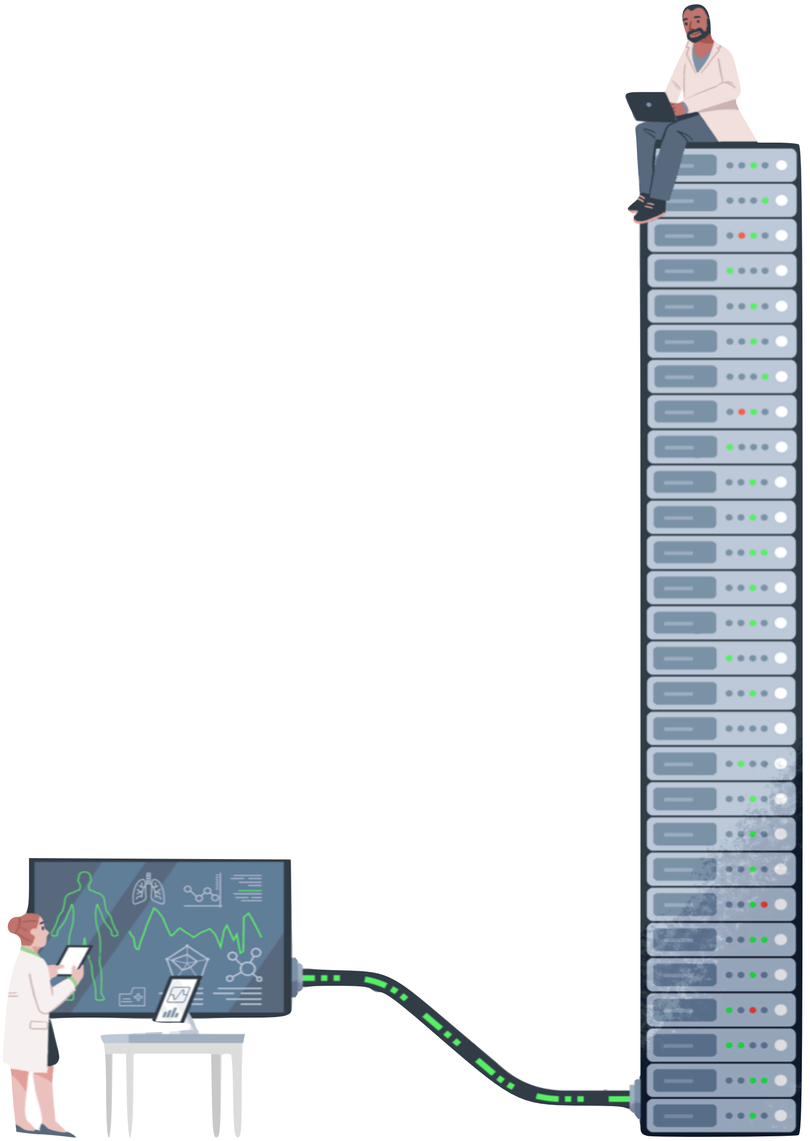
Explainability goes beyond the ability to merely interpret specific aspects of a project (e.g., interpreting the parameters of a model); it also depends on the ability to provide an accessible and relevant information base about the processes behind the outcome.



Core Attributes

Explainability

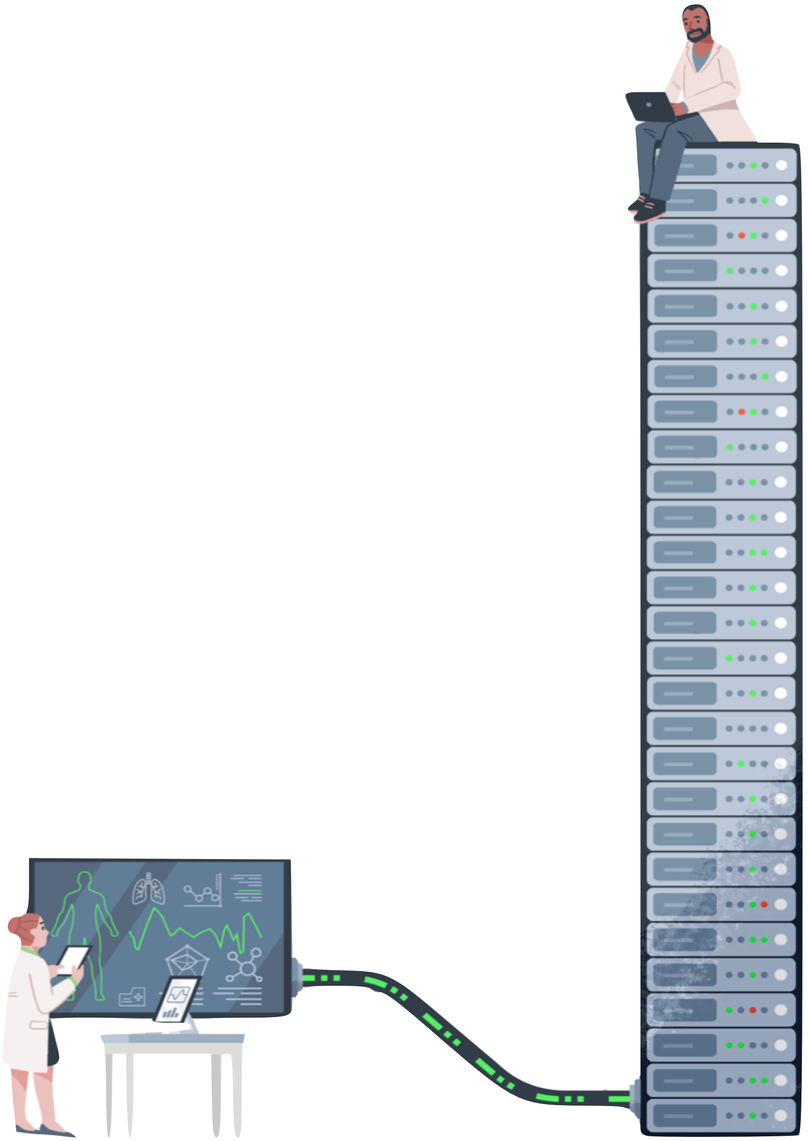
- Interpretability
- Responsible Model Selection
- Accessible Rationale Explanation
- Implementation and User Training



SAFE-D Principles

Data Quality

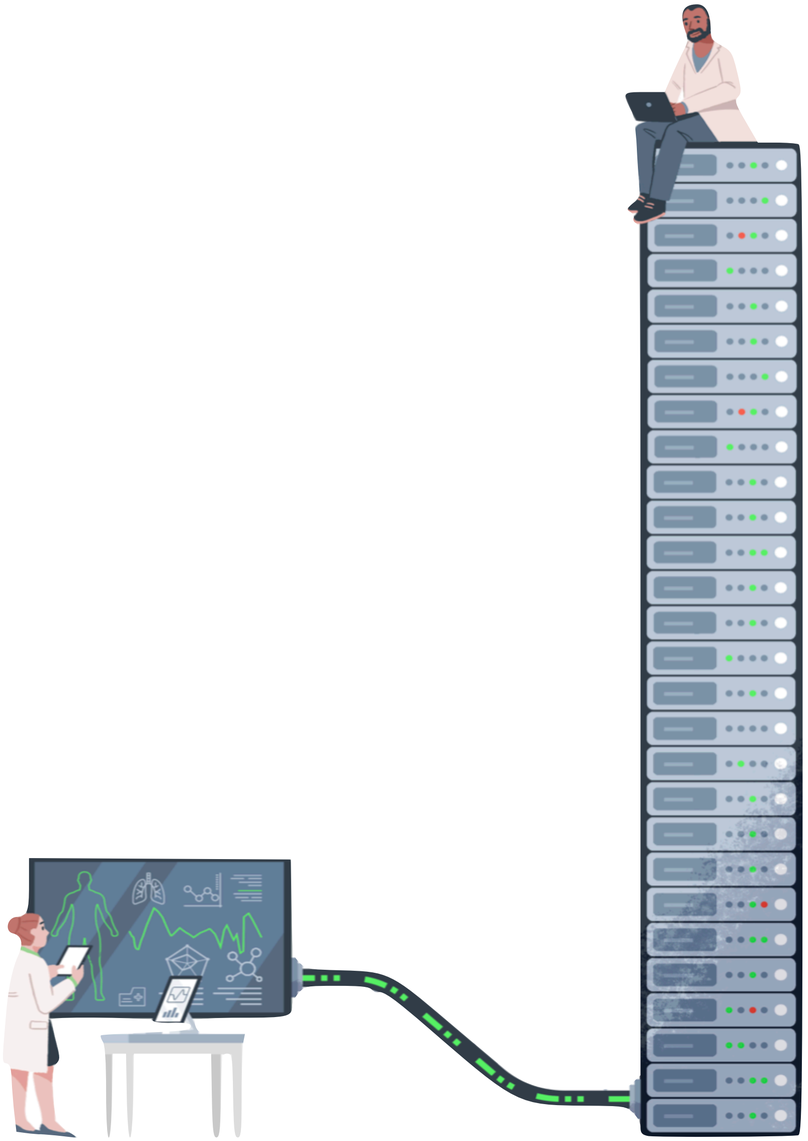
‘Data Quality’ captures the static properties of data, such as whether they are (a) relevant to and representative of the domain and use context, (b) balanced and complete in terms of how well the dataset represents the underlying data generating process, and (c) up-to-date and accurate as required by the project.



Core Attributes

Data Quality

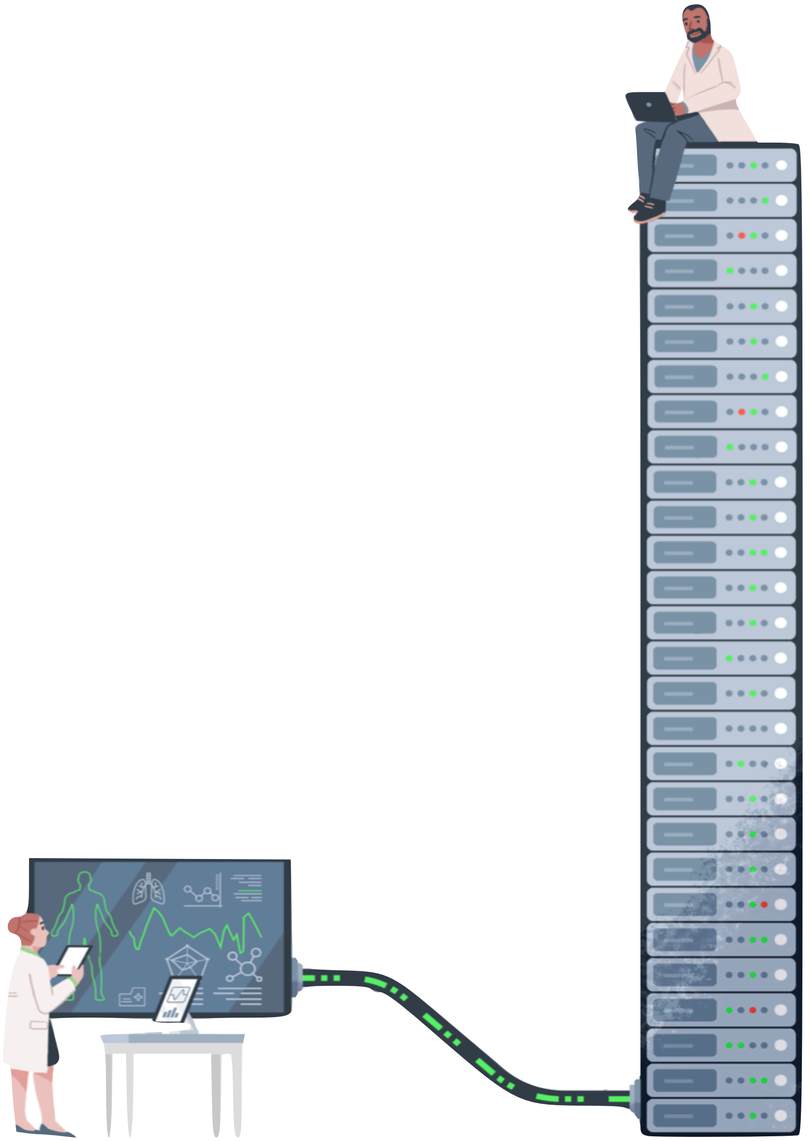
- Source Integrity and Measurement Accuracy
- Timeliness and Recency
- Relevance, Appropriateness, and Domain Knowledge
- Adequacy of Quantity and Quality
- Balance and Representativeness



SAFE-D Principles

Data Integrity

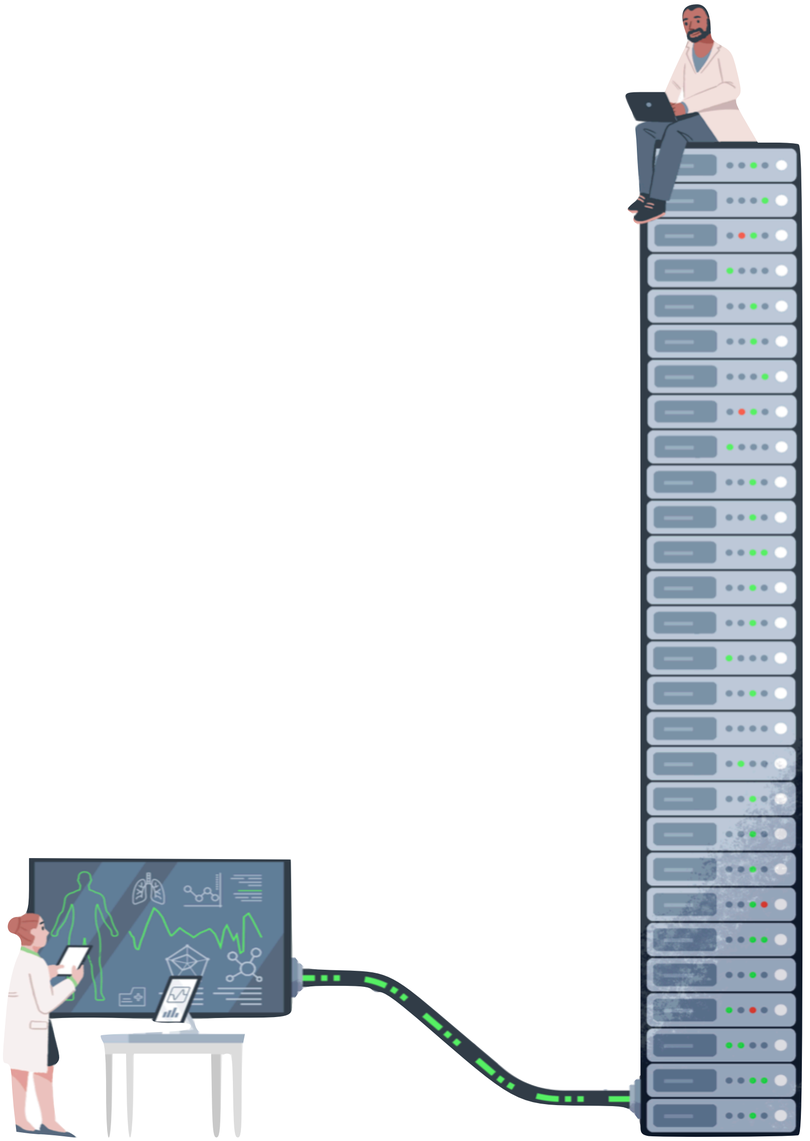
‘Data Integrity’ refers to more dynamic properties of data stewardship, such as how a dataset evolves over the course of a project lifecycle. In this manner, data integrity requires (a) contemporaneous and attributable records from the start of a project (e.g., process logs; research statements), (b) ensuring consistent and verifiable means of data analysis or processing during development, and (c) taking steps to establish findable, accessible, interoperable, and reusable records towards the end of a project’s lifecycle.



Core Attributes

Data Integrity

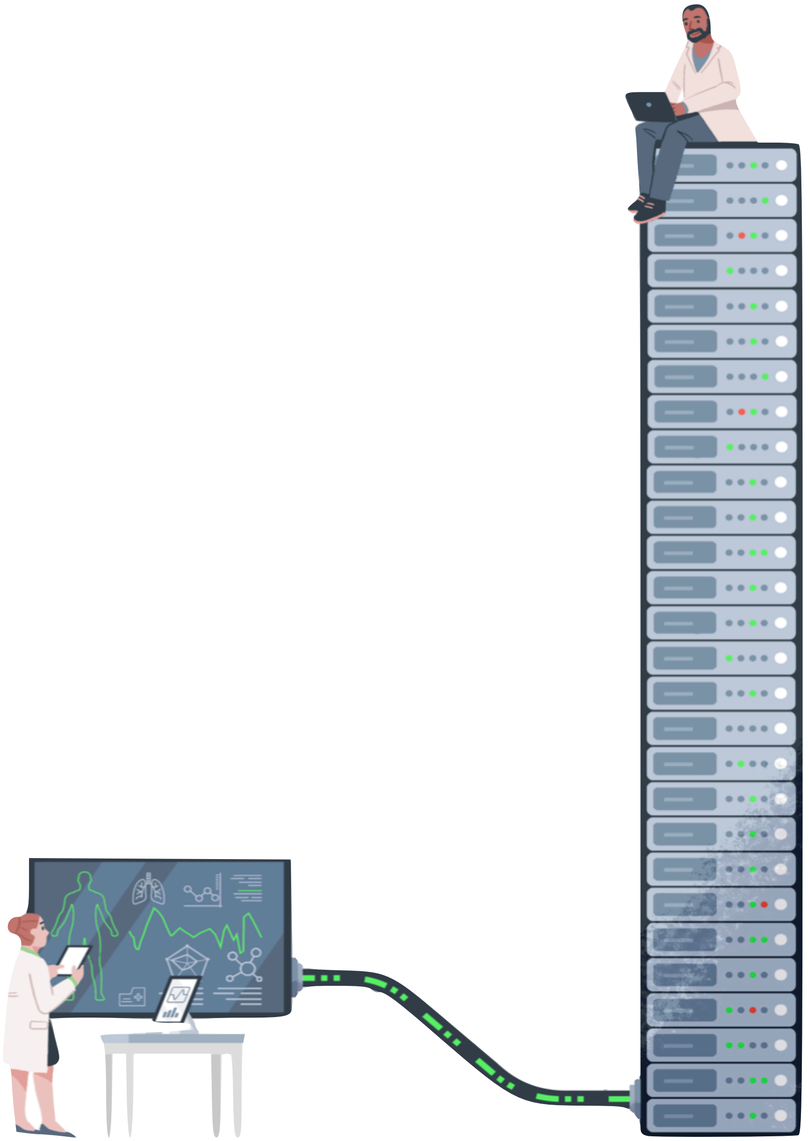
- Attributable
- Consistent, Legible and Accurate
- Complete
- Contemporaneous
- Responsible Data Management
- Data Traceability and Auditability



SAFE-D Principles

Data Protection and Privacy

‘Data protection and privacy’ reflect ongoing developments and priorities as set out in relevant legislation and regulation of data practices as they pertain to fundamental rights and freedoms, democracy, and the rule of law. For example, the right for data subjects to have inaccurate personal data rectified or erased.



Core Attributes

Data Protection and Privacy

- Consent (or legitimate basis) for processing
- Data Security
- Data Minimisation
- Transparency
- Proportionality
- Purpose Limitation

Examples

Property Claims

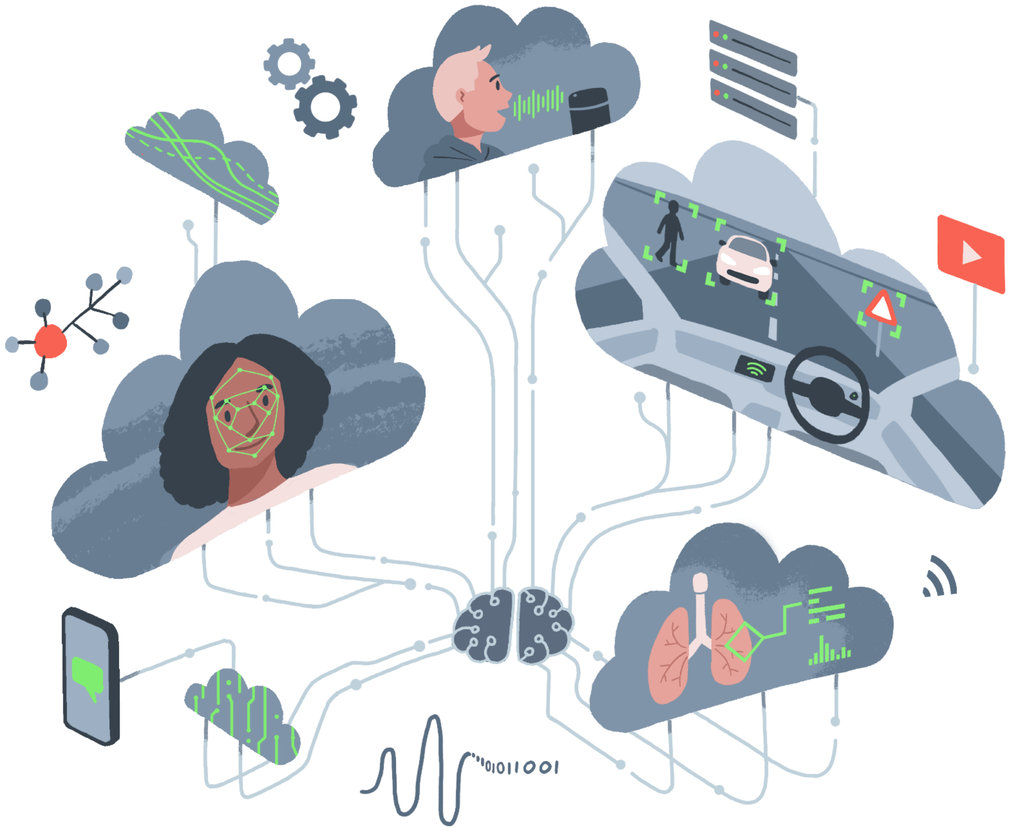
The following claims offer examples for each of the SAFE-D principles (or, goals) along with an attribute and corresponding project lifecycle stage

Goal & Attribute	Example Property Claim	Project Lifecycle Stage
Sustainability (Robustness)	*The model used in our system has been internally and externally validated. The external validation has been carried out across several varied environments to ensure robustness of the system.*	Model Training, Testing and Validation
Accountability (Accessibility)	*All identified stakeholders were consulted prior to the development of our system to help critically evaluate our project plans and ensure they were intelligible.*	Project Planning and Problem Formulation
Fairness (Equality)	*Persons affected by use of the system have avenues of recourse, ability to contest system outputs and demand human intervention.*	System Use & Monitoring
Explainability (Responsible Model Selection)	*Features were hand-selected in conjunction with domain experts to optimise for both interpretability and predictive power.*	Preprocessing & Feature Engineering and Model Selection
Data Quality (Timeliness & Recency)	*Only data that were collected within the previous 3 months were used to ensure the training data were up-to-date.*	Data Extraction or Procurement

Next Steps

Further Resources

- Turing Commons
  - Responsible Research and Innovation
  - Public Engagement of Data Science and AI
  - AI Ethics and Governance
- Understanding artificial intelligence ethics and safety
  - Public Sector Guidance
- Ethical Assurance
  - Article
  - Website



Thank you!

The Alan Turing Institute

Dr Christopher Burr
cburr@turing.ac.uk
@chrisdburr