Engineering Data Science

Christian Cabrera

University of Cambridge

06/11/2024

Previously...

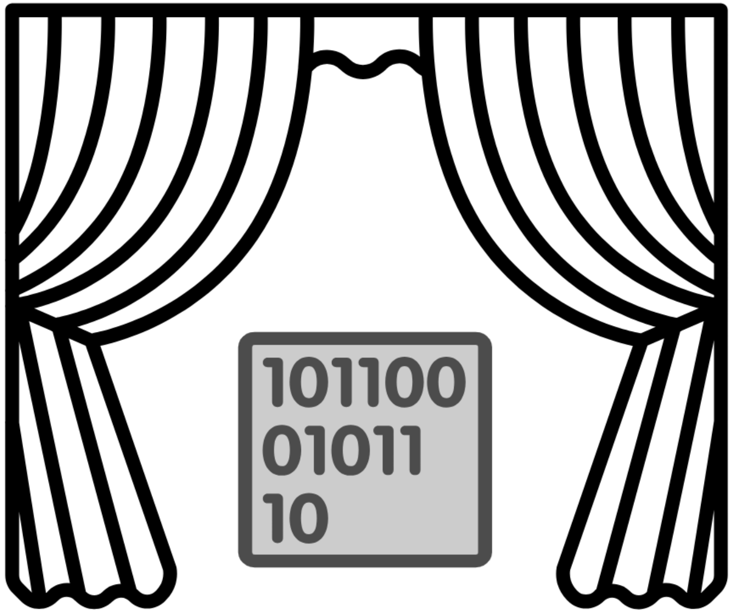
- Data Science Challenges

Previously...

- Data Science Challenges

  - **Bias**

  Systematic tendency in which methods used to gather data and compute statistics generate inaccurate depictions of reality.



Source: https://mlatcl.github.io/advds/lectures/04-02-ai-and-data-science.html

Challenges our ability to deploy safe and effective solutions:

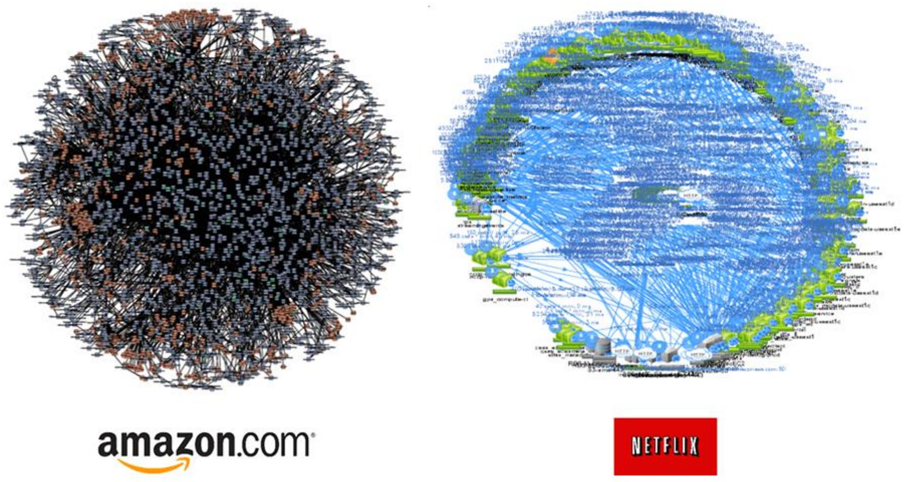
- **Alignment**
- **Fairness**
- **Inclusiveness**

Previously…

- Data Science Challenges

    - **Complexity**

    Systems are highly dynamic and have grown in size. The data processing pipelines involve hundreds or thousands of components.



Source: https://www.divante.com/blog/10-companies-that-implemented-the-microservice-architecture-and-paved-the-way-for-others

Challenges our technical ability to deploy and maintain our solutions:

- **Sustainability**
- **Maintainability**

Previously…

- Data Science Challenges

  - **Intellectual Debt**

  Black-box components make systems hard to understand and threaten human control. We know they components work but do not know how.

Source: (Zittrain-2019) https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c

Challenges our ability to explain our solutions:

- **Interpretability**
- **Accountability**

Why are these important?

- Data Science Challenges
  - **Bias**
  - **Complexity**
  - **Intellectual Debt**
  - **…**

Source: https://www.freepik.com/free-photos-vectors/society

Why are these important?

- Society has challenging problems...

Alignment

Interpretability

Sustainability

Inclusiveness

Fairness

Accountability

Maintainability

Source: https://www.freepik.com/free-photos-vectors/society

How do we build solutions now?

- Focus on technology

How do we build solutions now?

- Focus on technology

For example, we have a learning model that generates predictions from inputs:

<inputs> → **Learning Model** → <prediction>

How do we build solutions now?

- Focus on technology

For example, we have a learning model that generates predictions from inputs:

<inputs> → **Learning Model** → <prediction>

Learning Model	Accuracy
Model v1	88%
Model v2	90%
Model v3	98%

More capable models
Larger models
Powerful infrastructure
Promising applications

How do we build solutions now?

- Focus on technology

For example, we have a learning model that generates predictions from inputs:

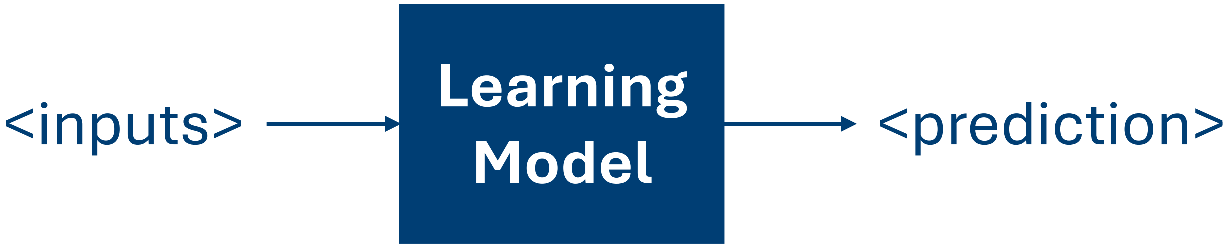
<inputs> → **Learning Model** → <prediction>

Learning Model	Accuracy
Model v1	88%
Model v2	90%
Model v3	98%

Does Model v3 address end users' requirements?

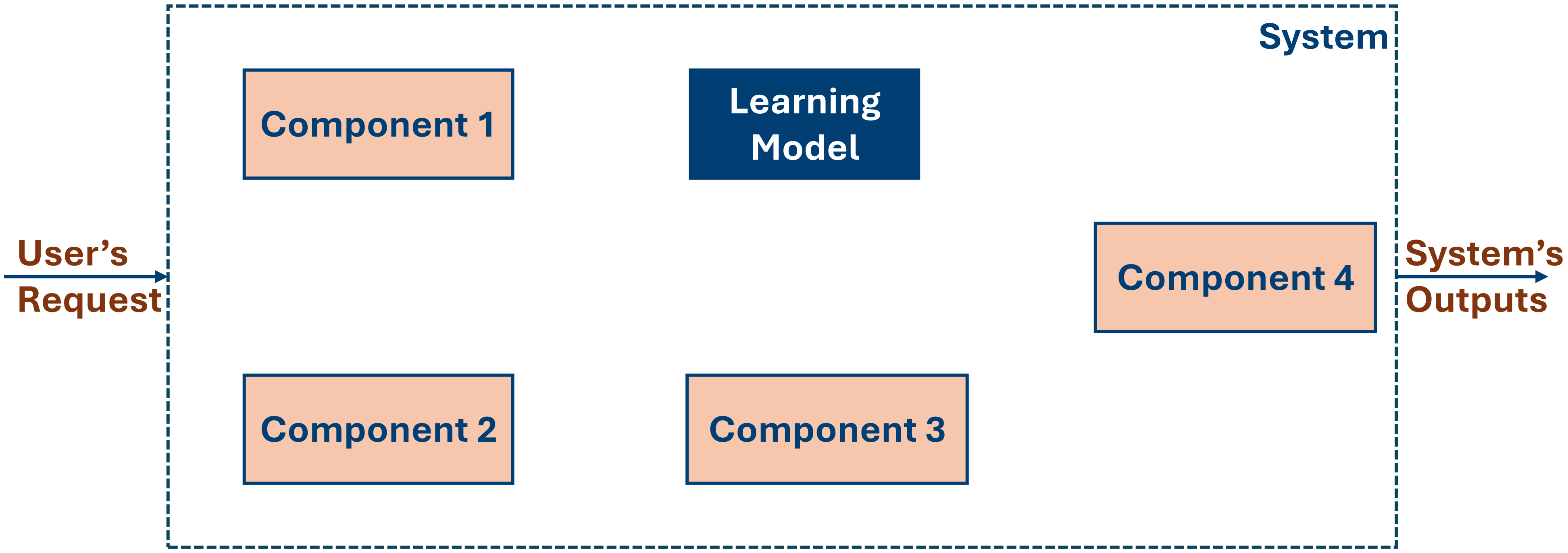
Real-world Deployments

- Context matters



Real-world Deployments

- Context matters



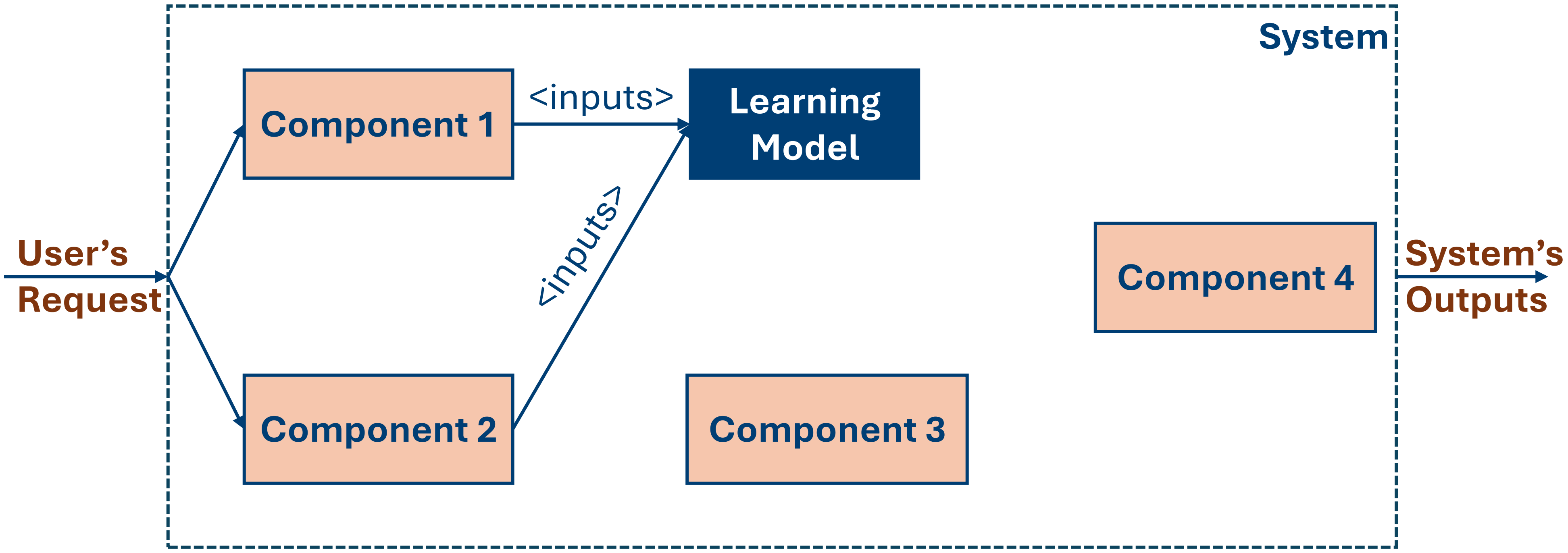
Learning models are deployed as part of larger systems.

Systems address end users' requirements.

Users' requirements must drive the design of the systems.

Real-world Deployments

- Context matters

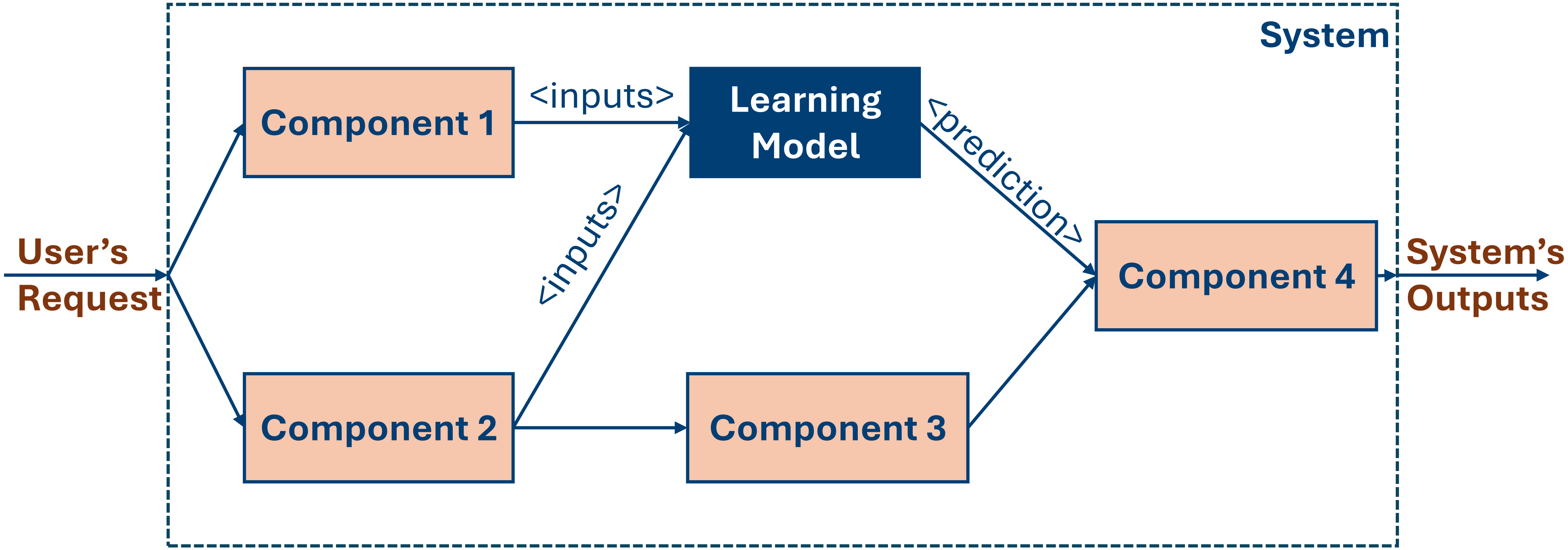


Systems are composed of multiple components

These components can impact the behavior of the learning models

Real-world Deployments

- Context matters

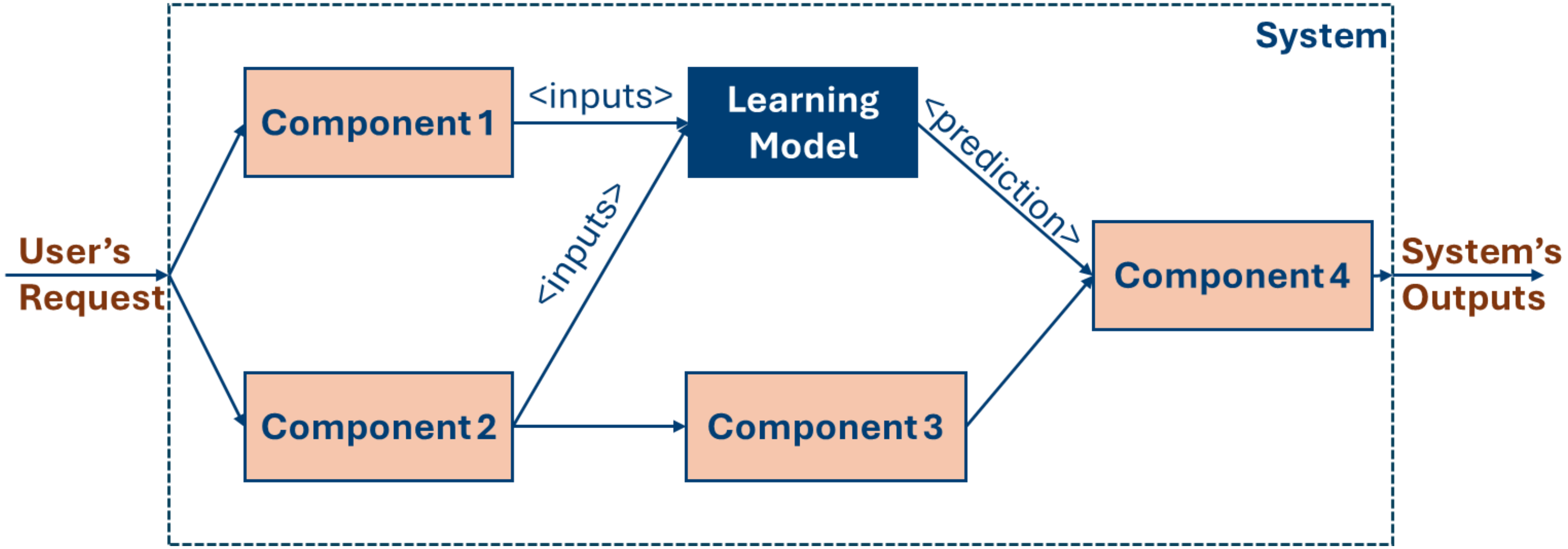


Learning models outputs can be inputs for other components

Systems' outputs are aggregations of the data processed by different components.

Real-world Deployments

- Context matters



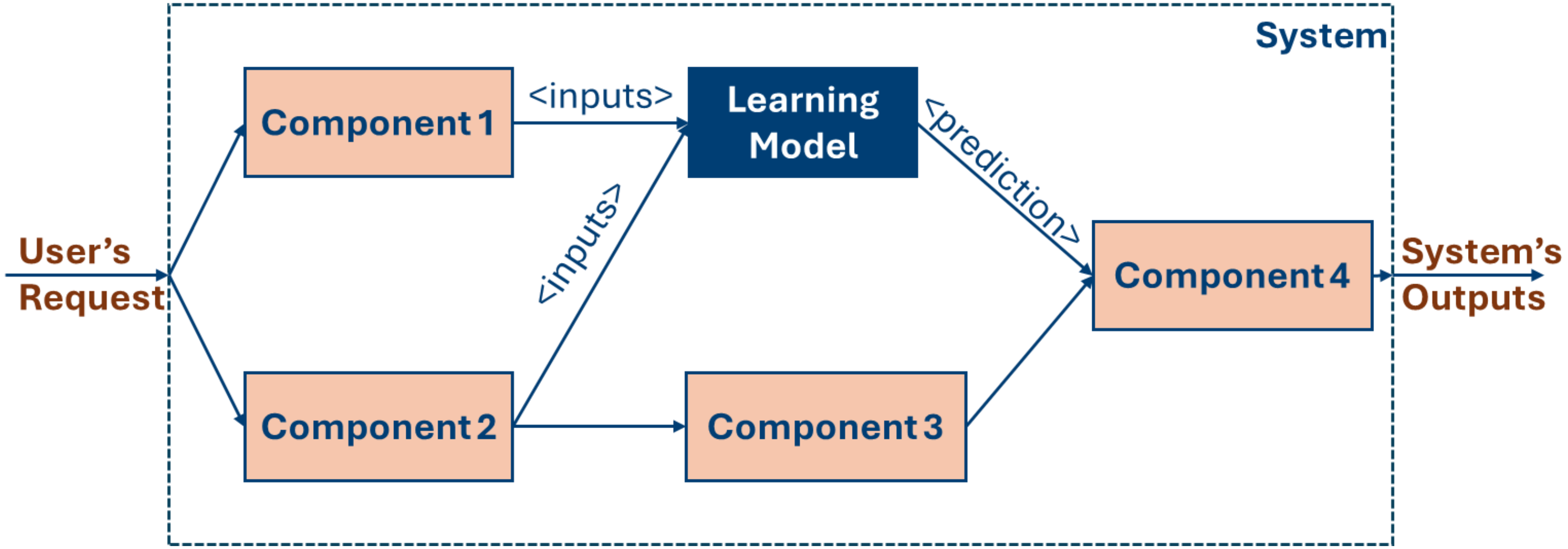
*** Low latency requirement!**

(< 5secs)

Learning Model	Accuracy	Latency
Modelv1	88%	3 secs
Modelv2	90%	4 secs
Modelv3	98%	10 secs

Real-world Deployments

- Context matters



*** Low latency requirement!**

(< 5secs)

***Constraint resources.**

Learning Model	Accuracy	Latency	Resources Demand
Modelv1	88%	3 secs	Low
Modelv2	90%	4 secs	Medium
Modelv3	98%	10 secs	High

Context

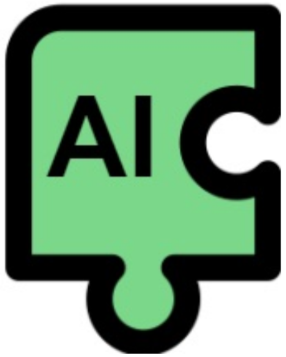
What do people need?
What are the social problems data science can help with?

Source:
https://commons.wikimedia.org/wiki/File:NP_coffee_cooperative_%285867722870%29.jpg

Source:
https://www.flickr.com/photos/scottishgovernment/23657582298/in/photostream/

Technology Adoption

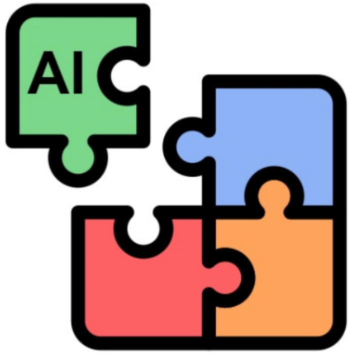
Socio-Technical Systems



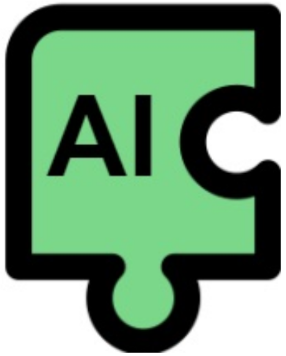
Artificial Intelligence

Technology Adoption

Socio-Technical Systems

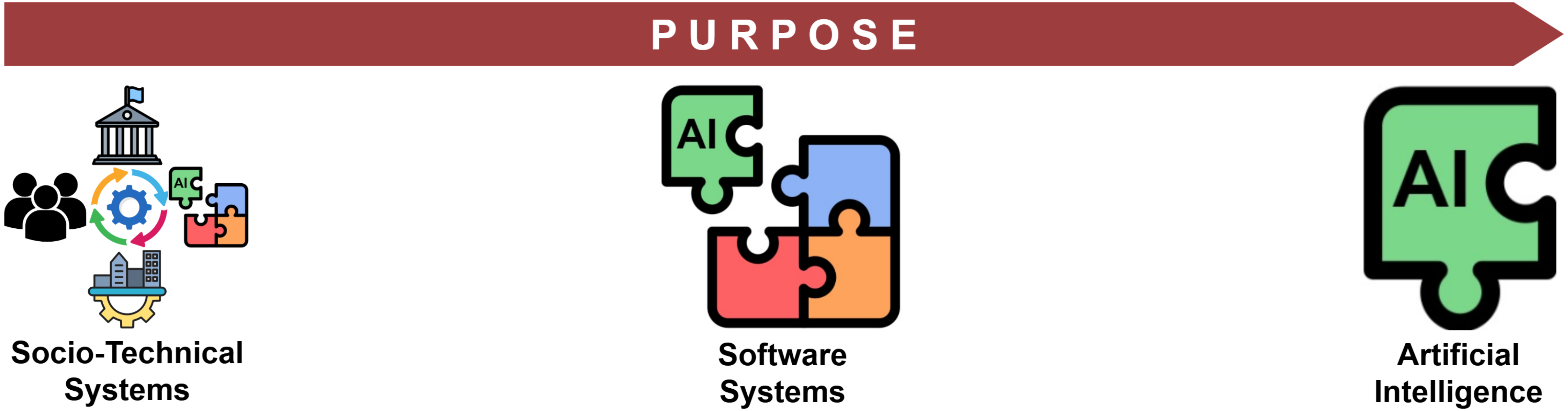


Software Systems

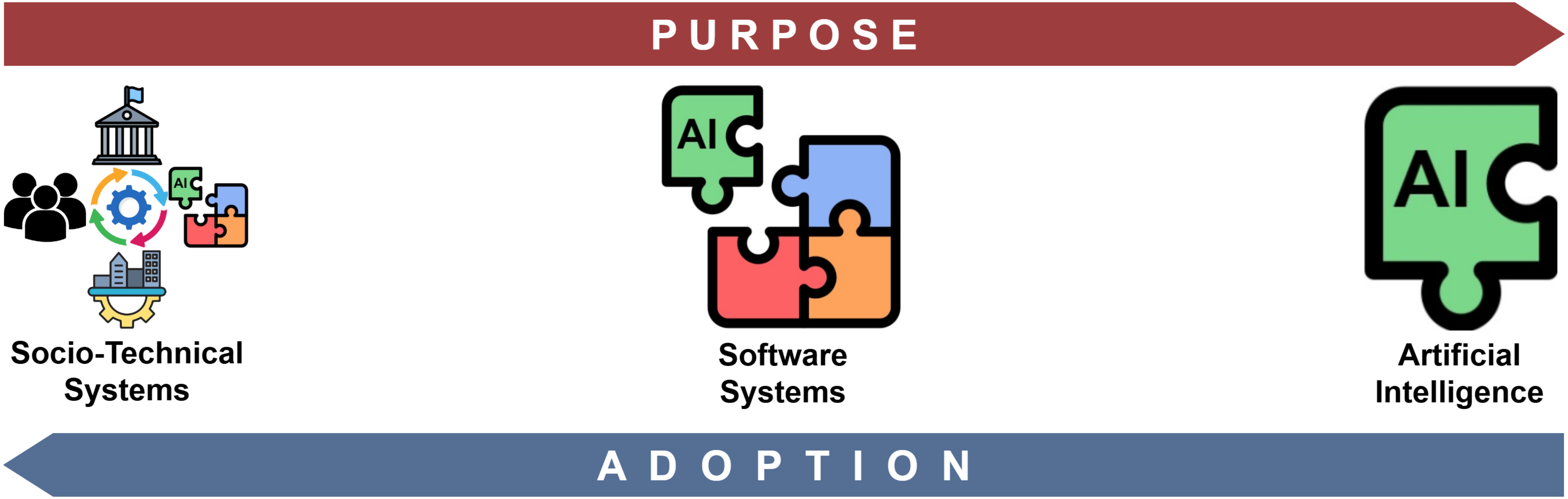


Artificial Intelligence

Technology Adoption



Technology Adoption



How do we engineer software systems?

The Systems Engineering Approach

- Problem first
  - Why is this problem important?
  - What are the people needs?
  - Which are the problem constraints?
  - What are the important variables to consider?
  - …

How do we engineer software systems?

The Systems Engineering Approach

- Engineering Principles
  - Systems Thinking
  - Process Model

How do we engineer software systems?

Systems Thinking

- System Views: Defining problems from different perspectives.
- Agility System: Defining flexible systems architectures.
- System Dynamics: Modelling the changing nature of systems.

How do we engineer software systems?

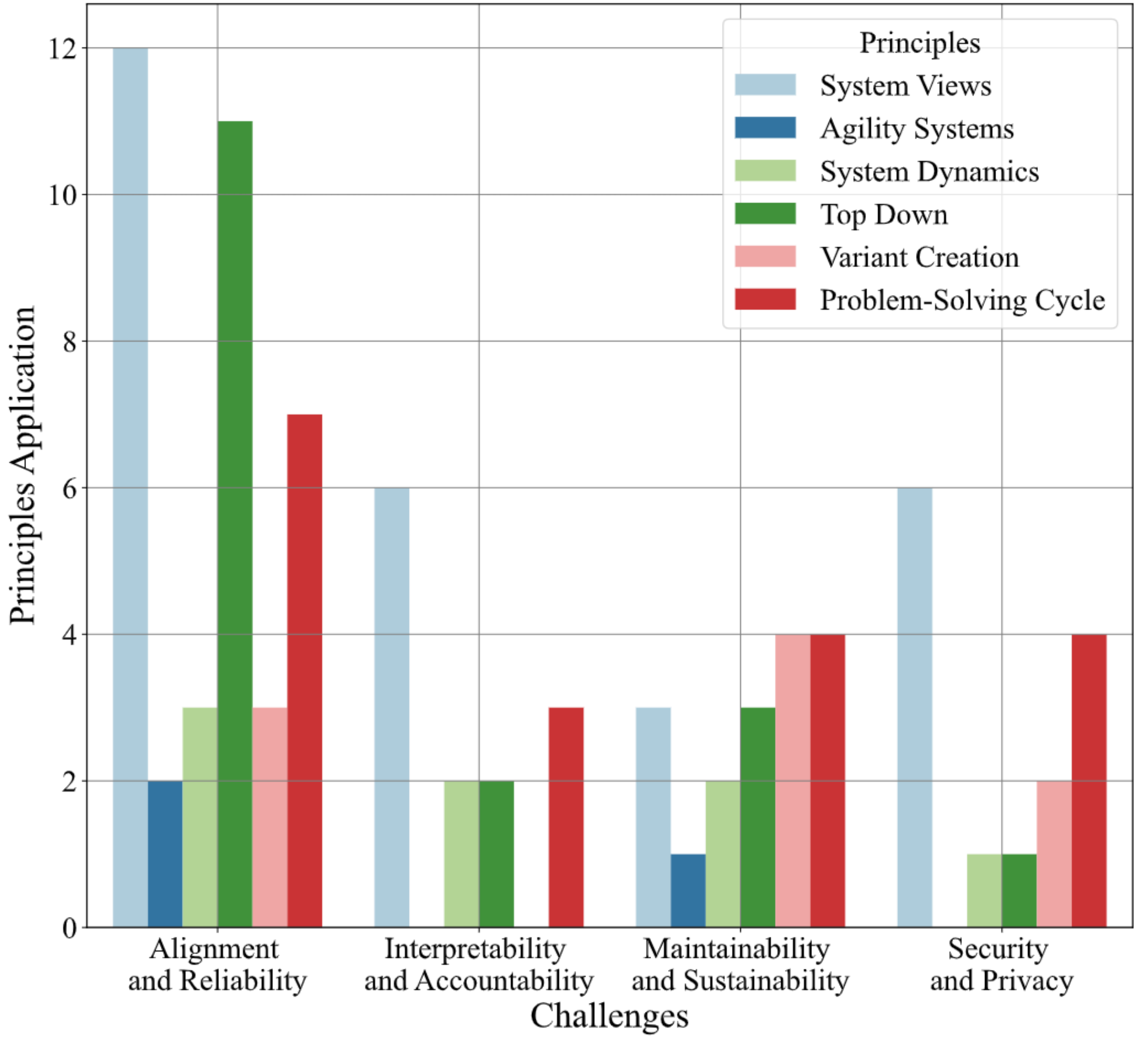
Process Model

- Top-Down Analysis: Divide and conquer. Problem decomposition.
- Variant Creation: Assessing different alternatives to solve a problem.
- Problem-Solving Cycle: Defining and following a methodology.

Systems Engineering Principles in Practice

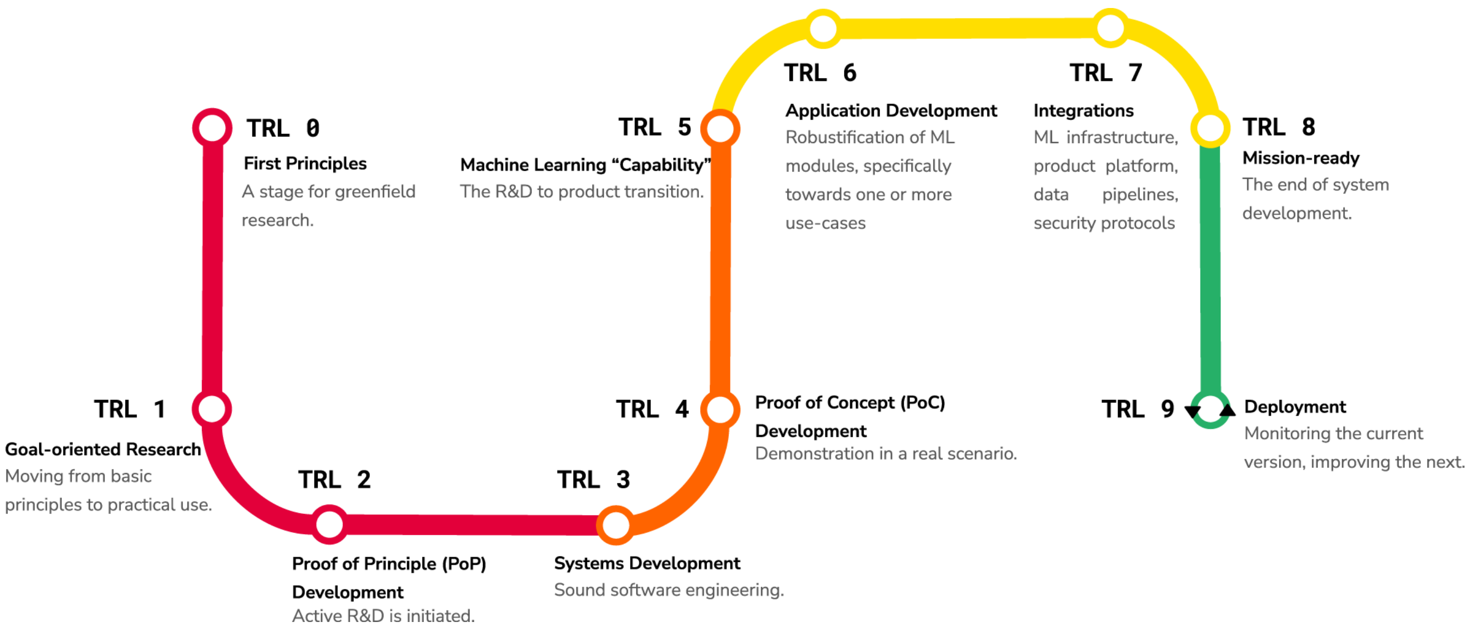
A survey of 24 works that apply the principles to address when deploying AI-based Systems:

- Alignment and Reliability
- Interpretability and Accountability
- Maintainability and Sustainability
- Security and Privacy



Cabrera, Christian, Bastidas Viviana, Schooling Jennifer, and Neil David Lawrence. “The Systems Engineering Approach in Times of Large Language Models." *Proceedings of the 58th Hawaii International Conference on System Sciences* 2025. (To appear)

Systems Engineering Principles in Practice



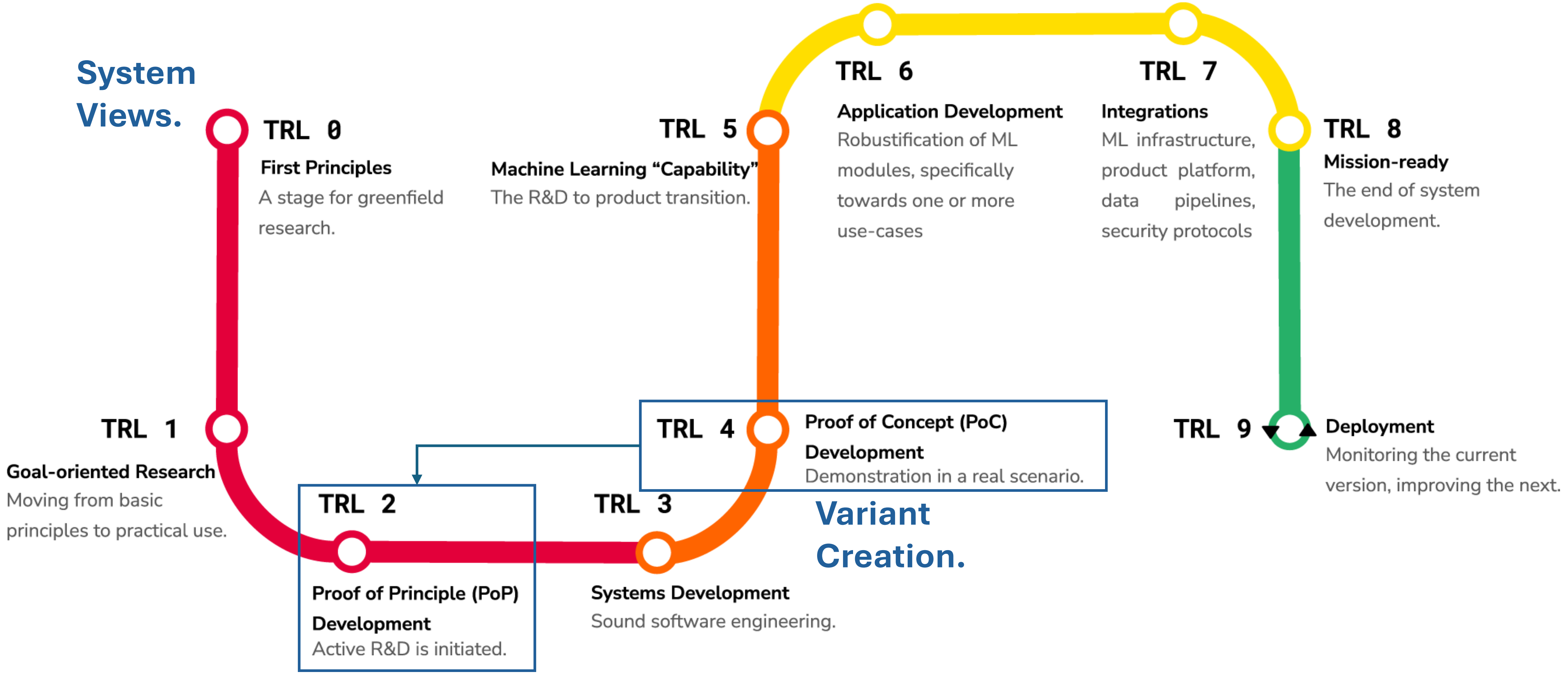
MLTRL – Technology readiness levels for machine learning systems

Domain: Critical Systems

Alignment and reliability, interpretability and accountability, maintainability and sustainability, and security and privacy.

Lavin, A., Gilligan-Lee, C.M., Visnjic, A. et al. Technology readiness levels for machine learning systems. Nat Commun 13, 6039 (2022). https://doi.org/10.1038/s41467-022-33128-9

Systems Engineering Principles in Practice



Lavin, A., Gilligan-Lee, C.M., Visnjic, A. et al. Technology readiness levels for machine learning systems. Nat Commun 13, 6039 (2022). https://doi.org/10.1038/s41467-022-33128-9

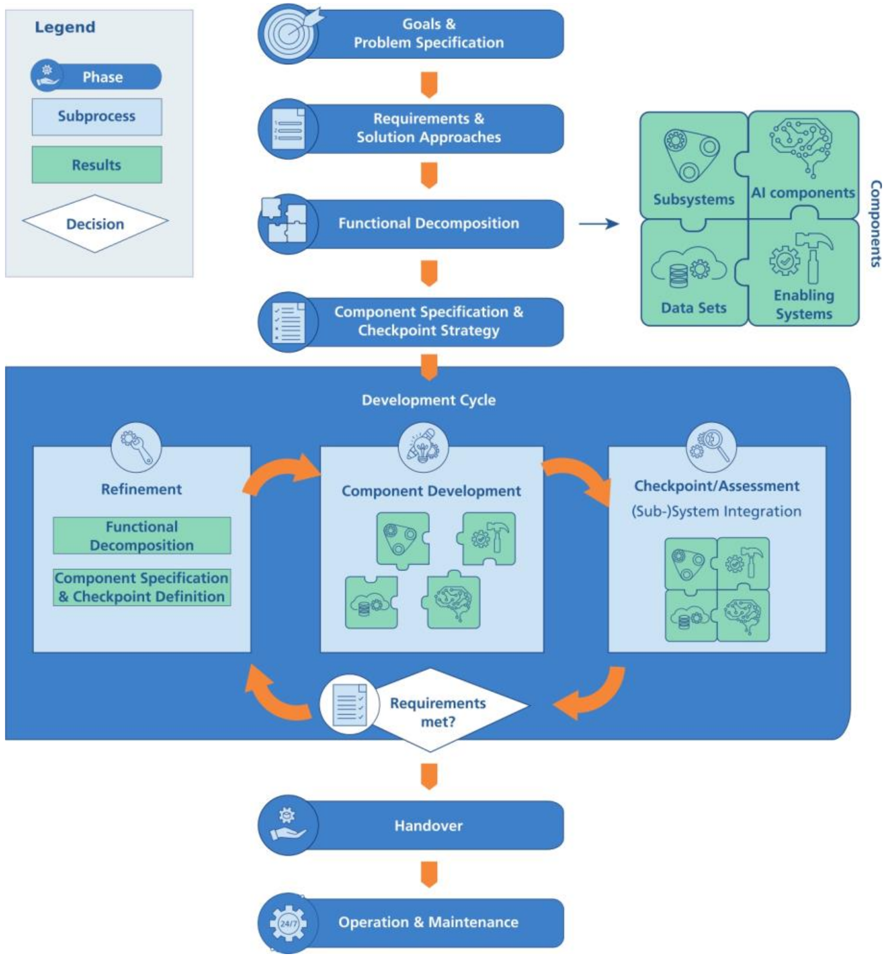
Systems Engineering Principles in Practice

PAISE® – Process Model for AI Systems Engineering

Domain: ML-based Systems

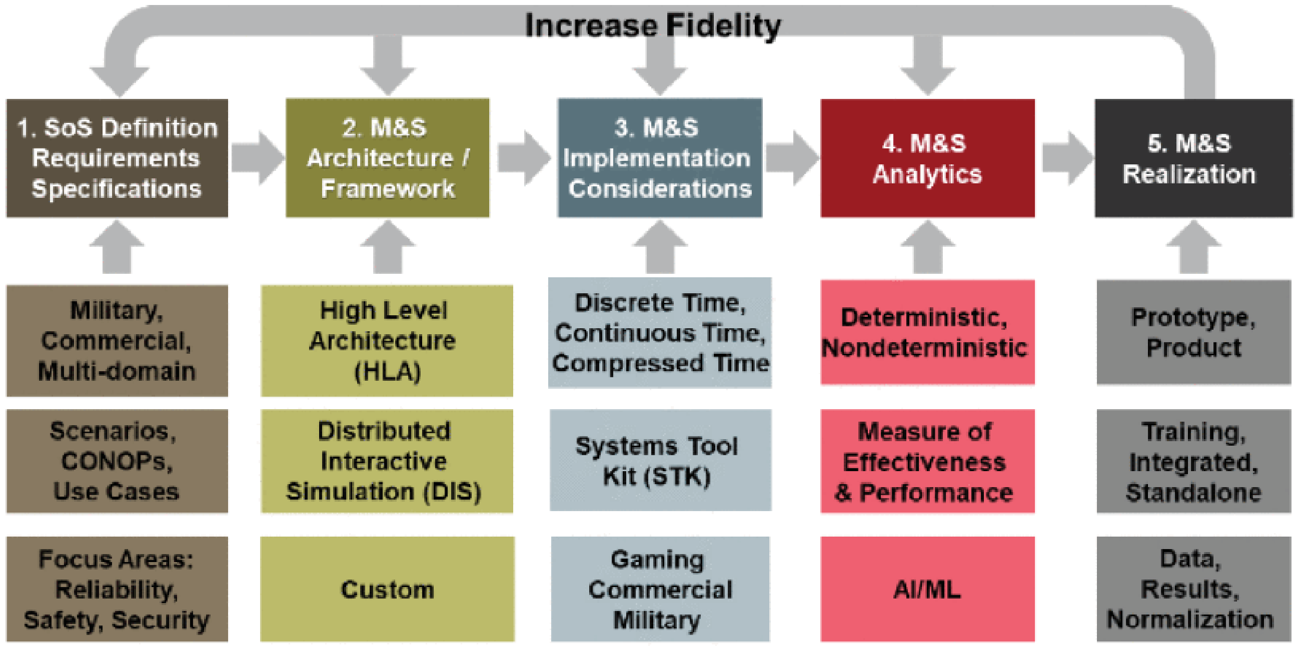
Alignment and reliability:

- Problem Solving Cycle.
- Top-Down Analysis.
- System Views.
- Agility Systems.



Hasterok, Constanze and Stompe, Janina. "PAISE® – process model for AI systems engineering" at - Automatisierungstechnik 70, no. 9 (2022): 777-786. https://doi.org/10.1515/auto-2022-0020

Systems Engineering Principles in Practice



Alignment and reliability, maintainability and sustainability, and security and privacy.

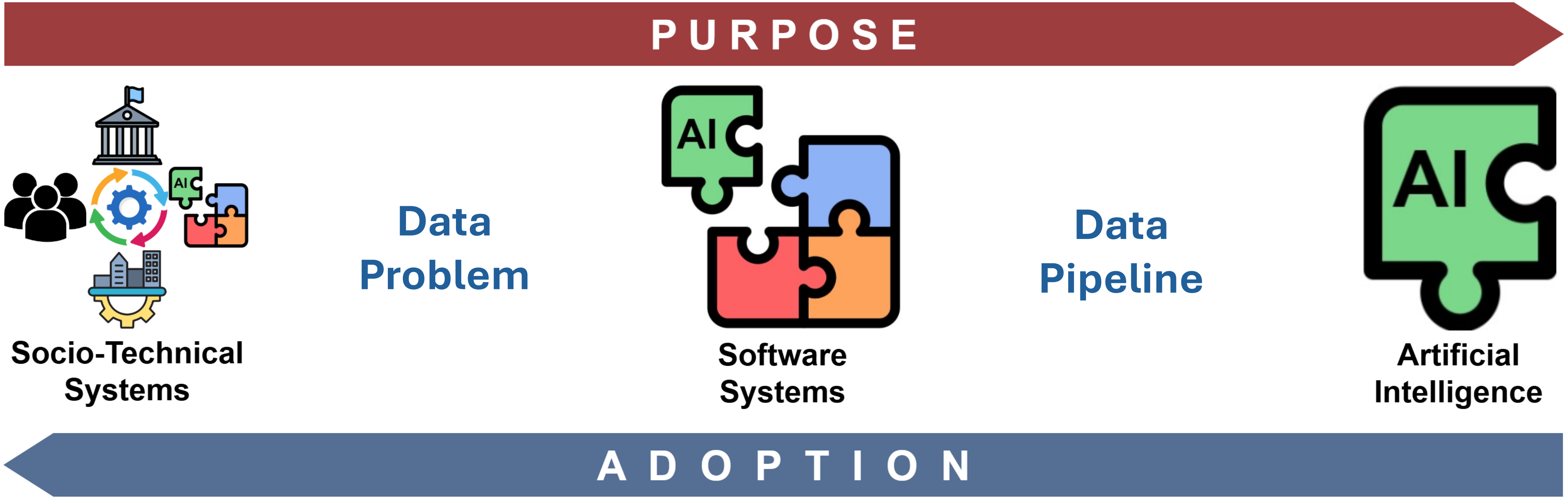
ACDANS – System of Systems Engineering Approach for Complex Deterministic and Nondeterministic Systems

Domain: Military Systems

- **Problem Solving Cycle.**
- **System Dynamics**
- **Variant Creation.**

P. Hershey, "System of Systems Engineering Approach for Complex Deterministic and Nondeterministic Systems (ACDANS)," 2021 16th International Conference of System of Systems Engineering (SoSE), Västerås, Sweden, 2021, pp. 185-190, doi: 10.1109/SOSE52739.2021.9497496.

Engineering Data Science



Data Problem

- Understanding stakeholders' needs
- Understanding resources and constraints
- Understanding data sources
- Understanding the nature of the data

Data Problem

- Understanding stakeholders' needs
- Understanding resources and constraints
- Understanding data sources
- Understanding the nature of the data

Systems Thinking

- Different points of view: **city council, landlord, tenants.**
- Agility Systems: **Flexible resources (e.g., cloud database), and flexible architecture (i.e., library).**
- Dynamic Systems: **Heterogeneous data, temporal features, spatial data, etc.**

Data Pipeline

The Fynesse framework:

- Access
- Assess
- Address

Data Pipeline

The Fynesse framework:

- Access
- Assess
- Address

Process Model

- Top-down: **downloading and uploading data chunks.**
- Variant creation: **exploring different data sources and management strategies.**
- Problem-solving cycle: **access, assess, and address loop.**

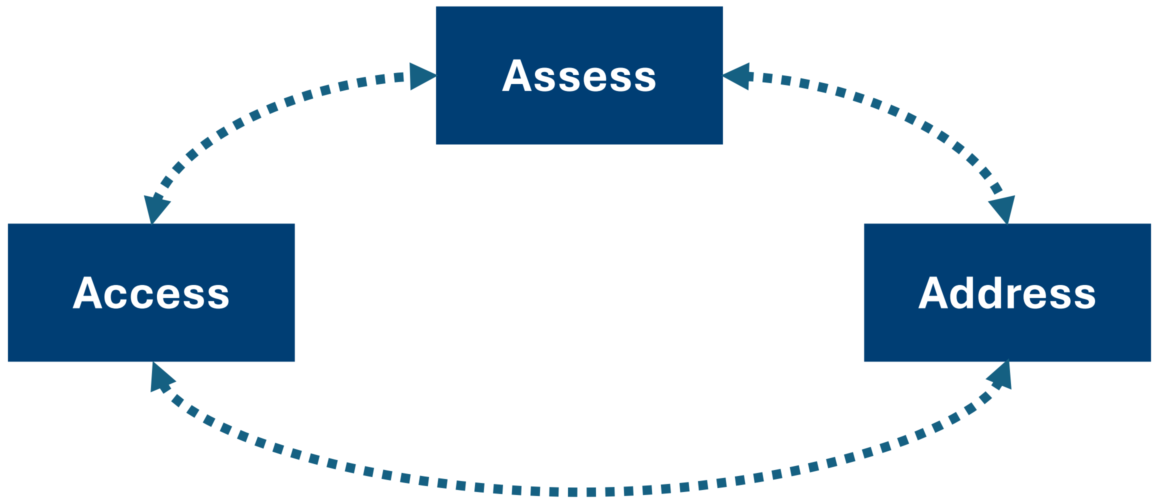
Engineering Data Science

Assess

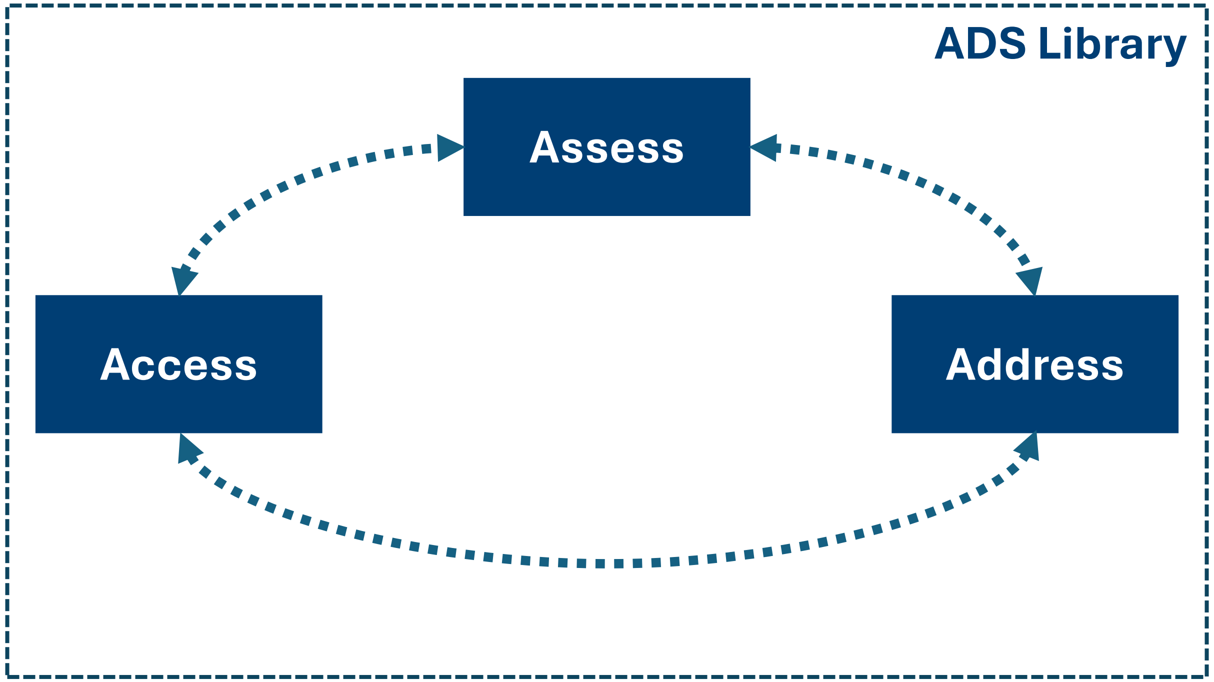
Access

Address

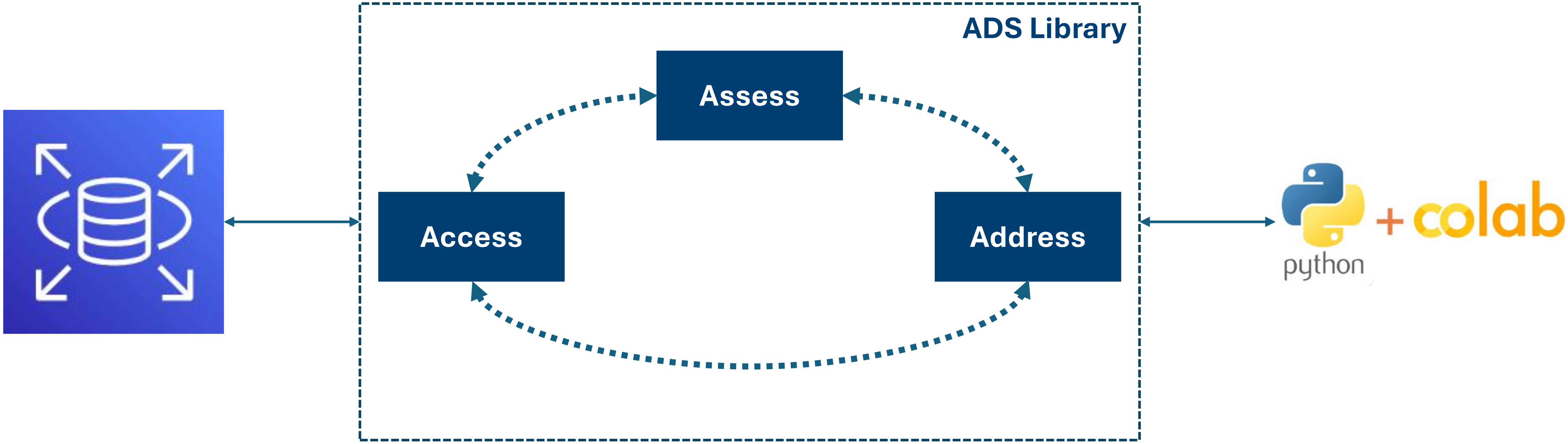
Engineering Data Science



Engineering Data Science

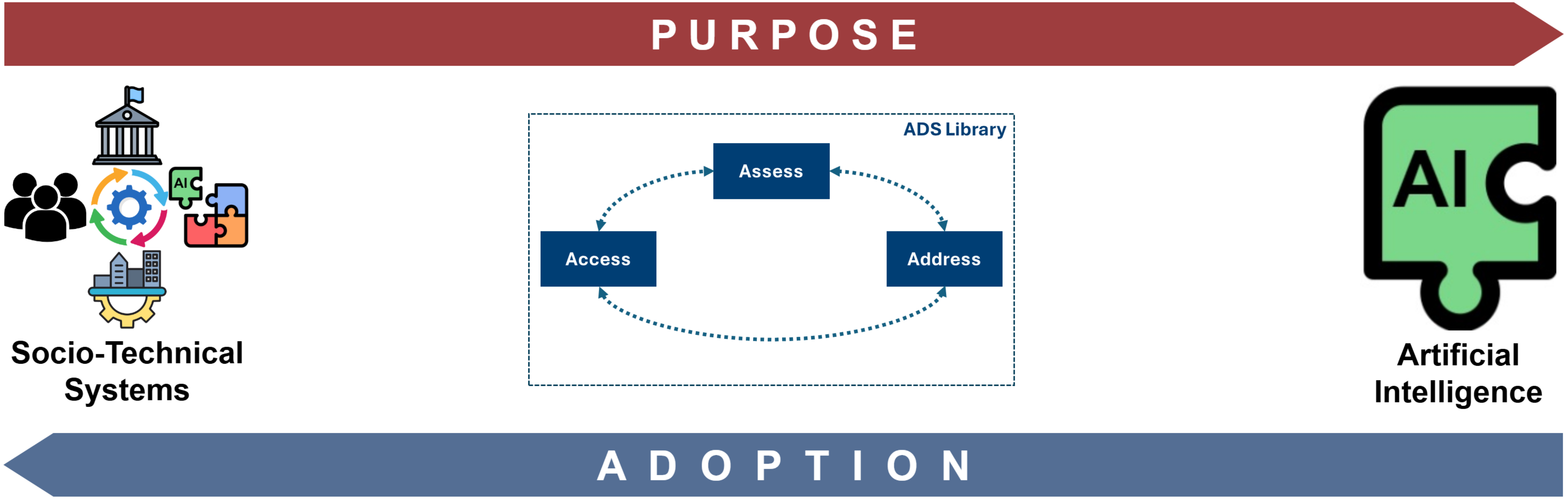


Engineering Data Science



Engineering Data Science

- Data from UK citizens and infrastructure.
- gov.UK and institutions managing data.
- Open Street Maps (OSM).



Learning models

Summary

- Context is important when addressing Data Science challenges.
- We must address these challenges because Data science projects impact people.
- The Systems Engineering approach provides principles to guide our work.
- We should include these principles in our Data Science projects.

Many thanks!