ARC Digital Bioprocess Development Hub

TUTORIAL

Towards Autonomous Machine Learning:

Evolution of AutoML, Roles of Humans, and Related Topics

Prof. Bogdan Gabrys and Dr. Thanh Tung Khuat

University of Technology Sydney

The 37th Australasian Joint Conference on Artificial Intelligence (AI 24)

November 25, 2024

Melbourne, VIC, Australia

ACKNOWLEDGEMENTS

ARC Digital Bioprocess Development Hub

Australian Government
Australian Research Council

This tutorial was **partly** supported under the **Australian Research Council's Industrial Transformation Research Program** (ITRP) funding scheme (project number IH210100051). However, *the vast majority of the materials, including four monographs, had been generated and published prior to and independent from the Hub*.

The ARC Digital Bioprocess Development Hub is a collaboration between The University of Melbourne, University of Technology Sydney, RMIT University, CSL Innovation Pty Ltd, Cytiva (Global Life Science Solutions Australia Pty Ltd) and Patheon Biologics Australia Pty Ltd.

Presenters

ARC **Digital Bioprocess Development** Hub

Prof. Bogdan Gabrys

Bogdan Gabrys is currently a Professor of Data Science and a Co-Director of the Complex Adaptive Systems Laboratory at the University of Technology Sydney, Sydney, Australia. His research activities have concentrated on the areas of *data science*, *complex adaptive systems*, *computational intelligence*, *machine learning*, *predictive analytics*, and *their diverse applications*. He has published over 220 research papers, chaired conferences, workshops, and special sessions, and been on program committees of a large number of international conferences with the data science, computational intelligence, machine learning, and data mining themes.

Dr. Thanh Tung Khuat

Thanh Tung Khuat is currently a postdoctoral research fellow at the Complex Adaptive Systems Lab at the University of Technology Sydney, focusing on building modern explainable, adaptive, and robust machine learning algorithms. His research interests include *machine learning*, *interpretable machine learning*, *fuzzy systems*, *knowledge discovery*, *evolutionary computation*, *intelligent optimization techniques*, and applications in biopharmaceuticals, medicine, finance, and agriculture. He has authored and co-authored over 40 peer-reviewed publications in the areas of machine learning and computational intelligence.

Preamble

ARC Digital Bioprocess Development Hub

In which directions and how do AutoML systems evolve?

What are the roles of humans as AutoML systems become smarter?

What is AutoML and AutonoML

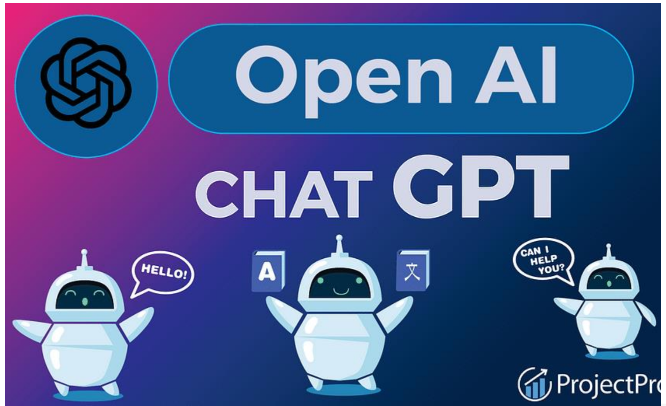
AutoML (Automated Machine Learning) is a set of tools and frameworks that automate various steps in the machine learning pipeline, from data preprocessing to model selection, training, hyperparameter optimization, and deployment. AutoML aims to make machine learning more accessible to non-experts by automating complex and time-consuming processes that typically require deep expertise.

Key Features of AutoML:

1. **Data preprocessing**: Automatically handles missing values, feature scaling, encoding categorical variables, etc.
2. **Model selection**: Tests a wide range of machine learning algorithms (e.g., decision trees, random forests, neural networks) to find the best one.
3. **Hyperparameter tuning**: Optimizes hyperparameters of selected models using techniques like grid search, random search, or Bayesian optimization.
4. **Cross-validation**: Ensures the robustness of models through automated k-fold cross-validation.
5. **Model ensembling**: Combines multiple models to boost performance, like stacking or bagging methods.
6. **Model deployment**: Automates the deployment of models into production environments.

AutoML platforms include tools like Google Cloud AutoML, H2O.ai, DataRobot, and Auto-sklearn. The goal is to democratize machine learning by reducing the need for deep technical expertise, shortening the development time, and improving the performance of models.

Will Autonomous AI systems be able to completely replace humans in the future?



2. `app.js` **(Node.js Server)**

This code sets up the login system and routes using Express.js.

javascript — Copy code

```javascript
const express = require('express');
const bodyParser = require('body-parser');
const app = express();
const port = 3000;

// Middleware for parsing form data
app.use(bodyParser.urlencoded({ extended: true }));

// Set EJS as the templating engine
app.set('view engine', 'ejs');

// Hardcoded credentials for login (in a real app, you'd use a database)
const username = "admin";
const password = "password123";

// Route: Render login page
app.get('/', (req, res) => {
  res.render('login');
});

// Route: Handle login submission
app.post('/login', (req, res) => {
  const { username: enteredUsername, password: enteredPassword } = req.body;

  // Check credentials
  if (enteredUsername === username && enteredPassword === password) {
    res.render('dashboard', { usern[illegible] enteredUsername });
  } else {
```

Tutorial Outline

ARC Digital Bioprocess Development Hub

1. **Introduction to AutoML, AutonoML, and related topics** (30 mins)
2. **Interacting with AutoML Systems**: Current Practices (35 mins)
3. **Interacting with AutoML Systems**: Constrained but Fully Automated (30 mins)
4. **Interacting with AutoML Systems**: Open-ended Environments (25 mins)
5. **Critical Discussion and Future Directions** (30 mins)

Questions and Answering (5 mins) at the end of each section

ARC Digital Bioprocess Development Hub

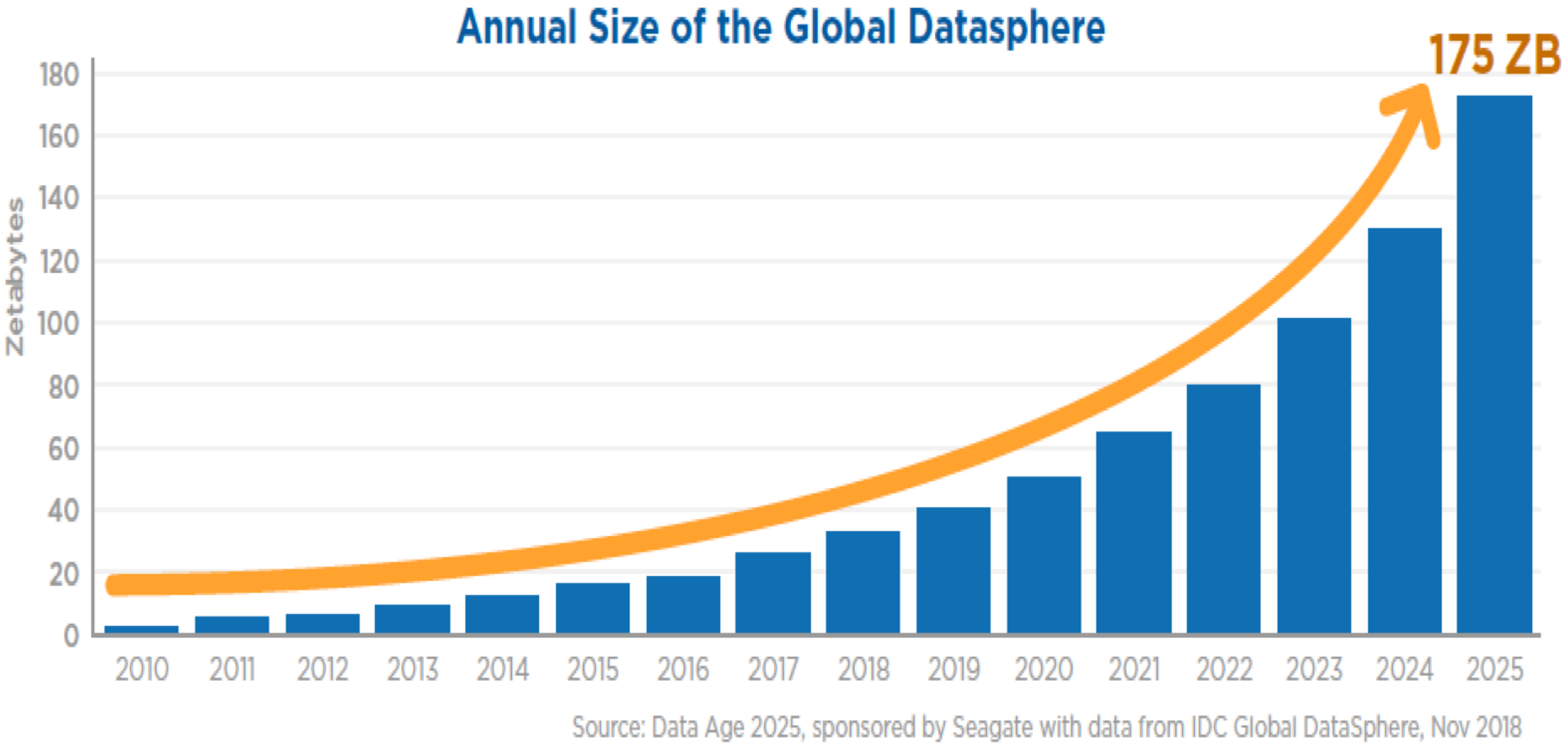
PART 1:

INTRODUCTION TO AUTOML, AUTONOML, AND RELATED TOPICS

(30 mins – Prof. Gabrys)

Digital Universe – incredible growth of digital content and information

ARC Digital Bioprocess Development Hub



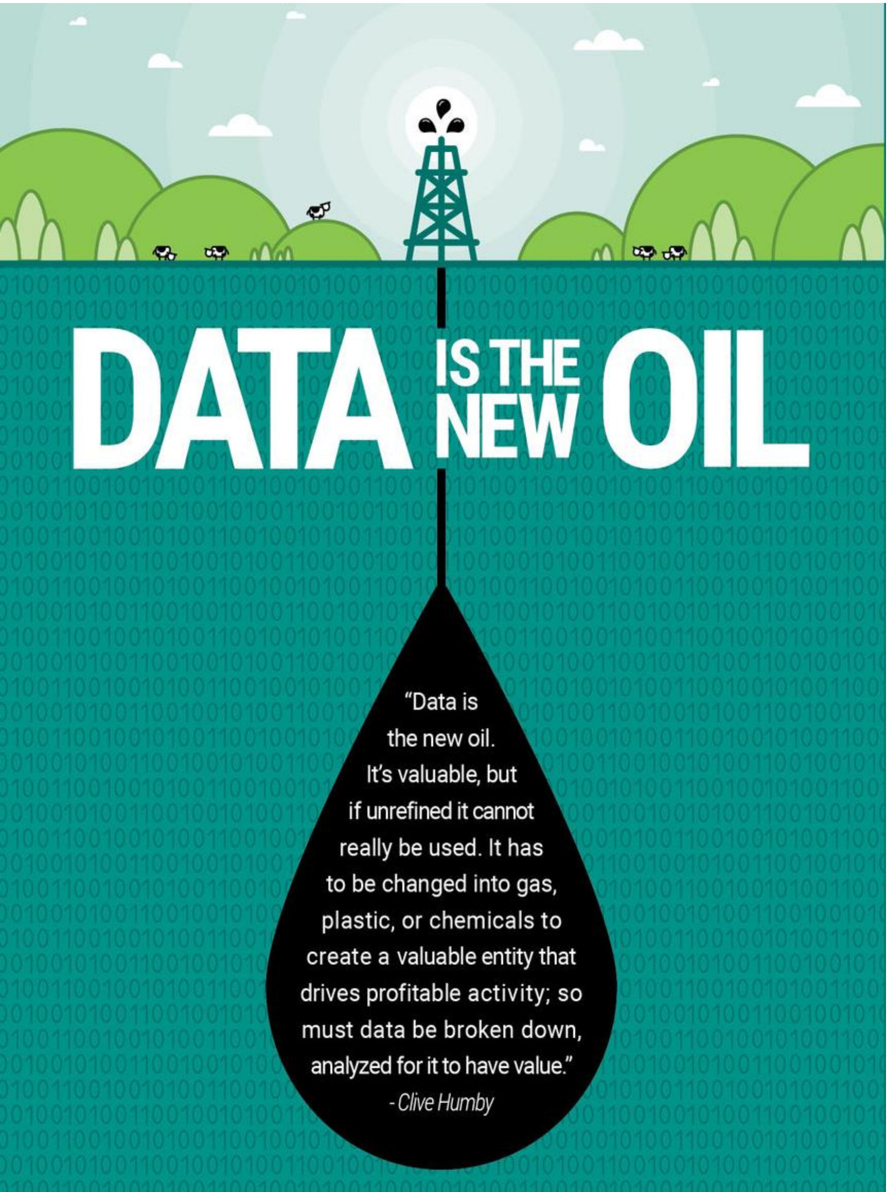
New Contexts in Industry 4.0

ARC Digital Bioprocess Development Hub

BIG DATA

New Contexts in Industry 4.0

ARC Digital Bioprocess Development Hub



Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, or chemicals to create a valuable entity that drives profitable activity; so must data be broken down, analysed for it to have value.

- Clive Humby

Predictive Analytics

&

Data-Driven Machine Learning

Main challenges in Data Science

- **Research in traditionally qualitative disciplines is fundamentally changing** due to the availability of vast amounts of data.
- The **commercial world has been transformed** by a focus on **Big Data** with companies competing on analytics.
- **Public sector organisation and government departments** have been changing their decision-making practices using both data and advanced analytics.
- **Data** has become a commodity and in recent years has been referred to as the '**new oil**'.
- **A new era of predictive analytics and data intensive computing** which has been recognised worldwide.

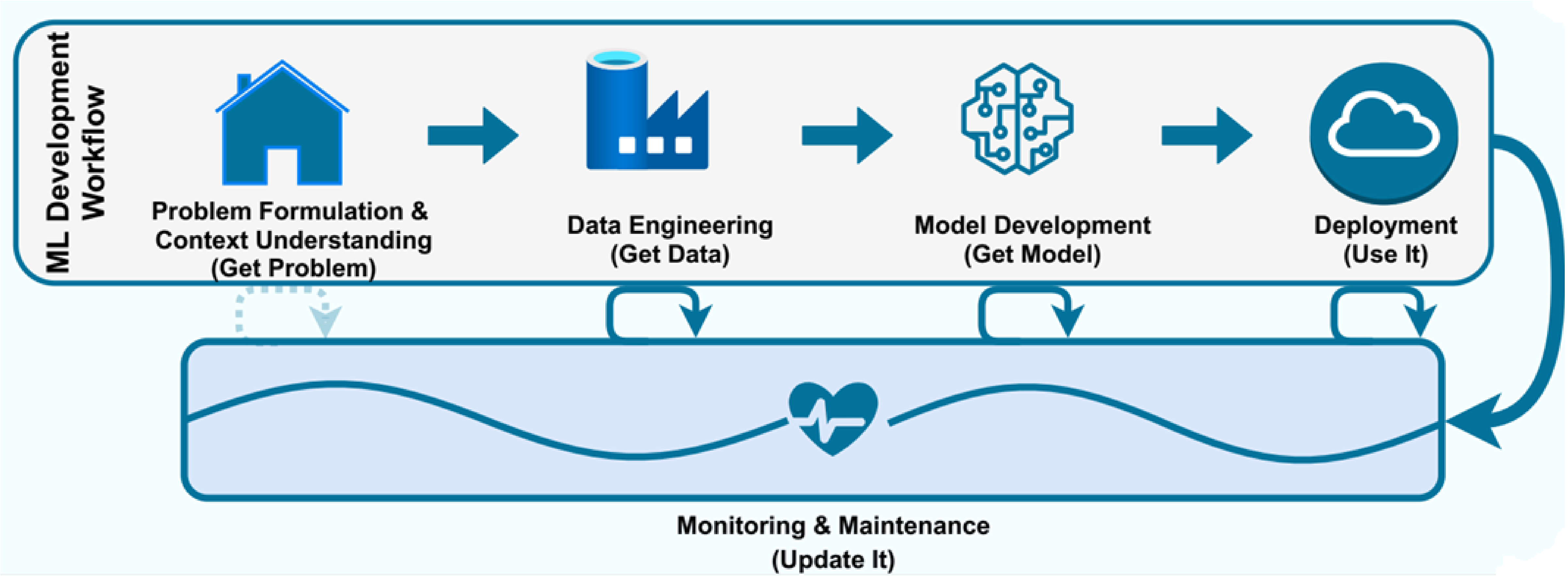
Main challenges in Data Science

ARC **Digital Bioprocess Development** Hub

- **People:** Very strong evidence for **great and immediate need for people with data science and advanced analytics expertise** and skill sets including creativity, excellent communication and business acumen to realise the huge potential that the (big) data revolution brings.
- **Technology:** Great need for **developing scalable, automated and adaptive, predictive and visual data analytics techniques, tools and products** allowing a broader spectrum of non-specialist users to take advantage of the wealth residing in huge and quickly growing digital universe.

A General Machine Learning Workflow

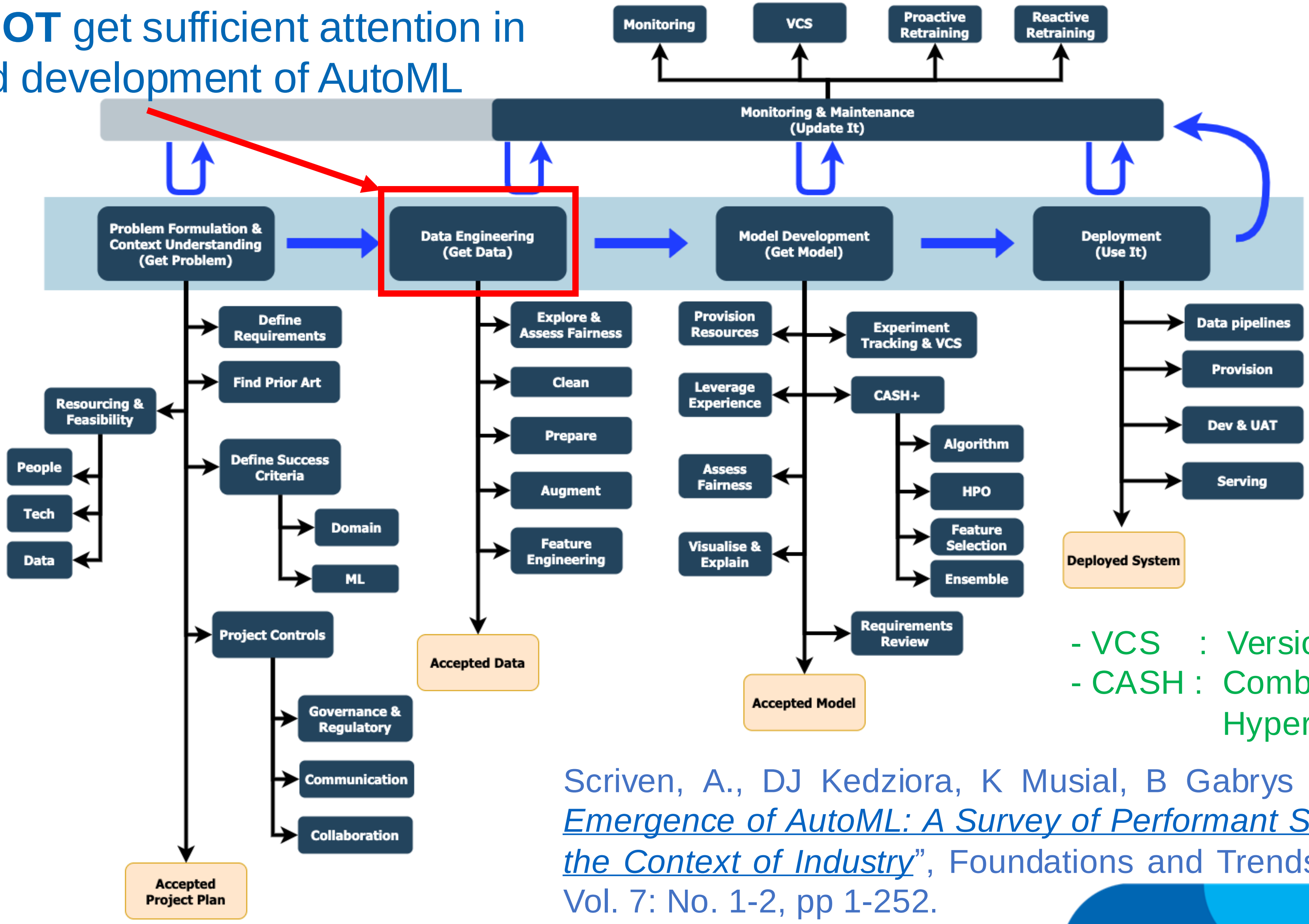
ARC Digital Bioprocess Development Hub



Machine Learning Workflow: Detailed View

ARC **Digital Bioprocess Development** Hub

Usually do **NOT** get sufficient attention in research and development of AutoML



- VCS : Version Control Systems, e.g., Git
- CASH : Combined Algorithm Selection and Hyperparameter Optimisation

Scriven, A., DJ Kedziora, K Musial, B Gabrys (2023) "*The Technological Emergence of AutoML: A Survey of Performant Software and Applications in the Context of Industry*", Foundations and Trends® in Information Systems: Vol. 7: No. 1-2, pp 1-252.

Practical Problems of Predictive Modelling Exploitation

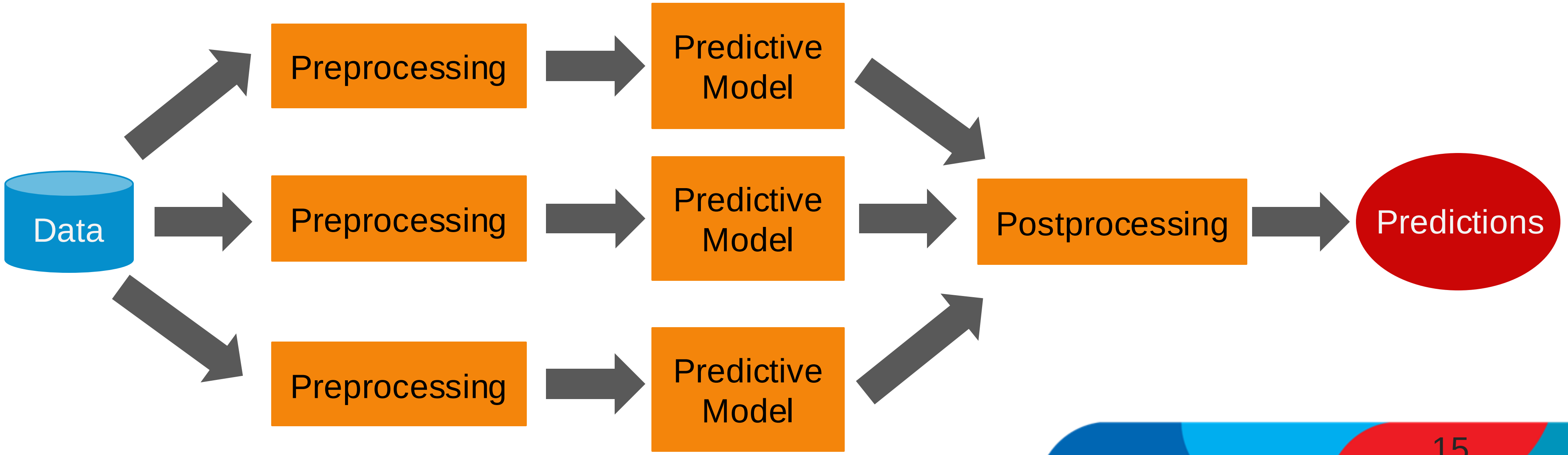
ARC Digital Bioprocess Development Hub

- **Very heavy reliance on predictive modelling expert knowledge and experience** (and bias towards their favourite methods/tools).
- Most of the existing **flexible/adaptable predictive systems rely on** (often a large number) **user specified (hyper-)parameters** on which the system performance usually critically depends.
- **Data acquisition** and **robust pre-processing**, which are frequently omitted in scientific papers, are absolutely **critical for the successful operation** of a predictive system.
- **Labour intensive** and **expensive process of building** and especially **maintaining** predictive models.
- **Need for repetitive analysis** as the predictive models are **not adaptive** / have **no mechanisms for reliable** and **robust learning over time** when the situations/environments change.

Automated Predictive Model Building - in a nutshell

ARC **Digital Bioprocess Development** Hub

Automatically composed, optimised and deployed Multicomponent Predictive System (MCPS)



CASH Problem for Multi-Component Predictive Systems

ARC **Digital Bioprocess Development** Hub

Combined Algorithm Selection and Hyperparameter (CASH) configuration problem

k-fold cross validation

Training dataset

$$\underset{(P,T,F)^{(j)}\in\mathcal{A},\lambda\in\Lambda^{(j)}}{\arg\min}\ \frac{1}{k}\sum_{i=1}^{k}\mathcal{L}\left((P,T_\lambda,F)^{(j)},\mathcal{D}^{(i)}_{train},\mathcal{D}^{(i)}_{valid}\right)$$

Validation dataset

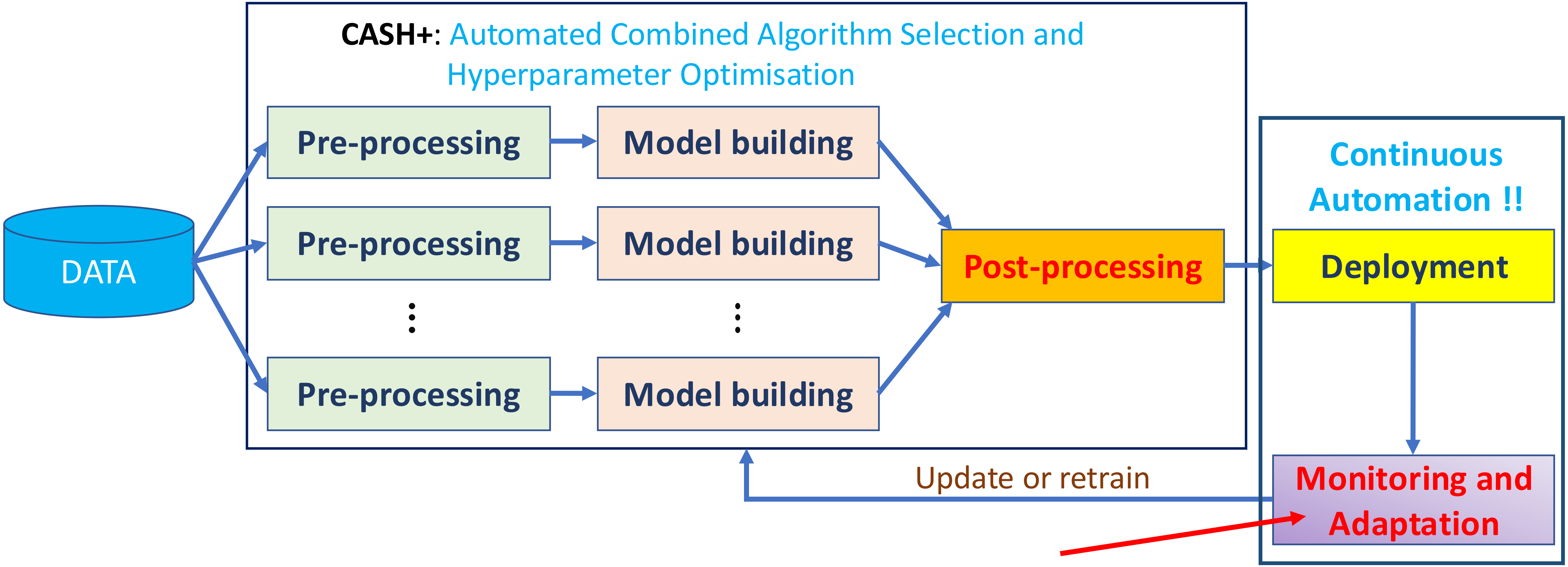
MCPSs

Hyperparameters

Objective function (e.g. classification error)

Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K.: *Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms*. In: Proc. of the 19th ACM SIGKDD. (2013) 847–855

Overview of Automated Predictive Model Building: AutoML to AutonoML

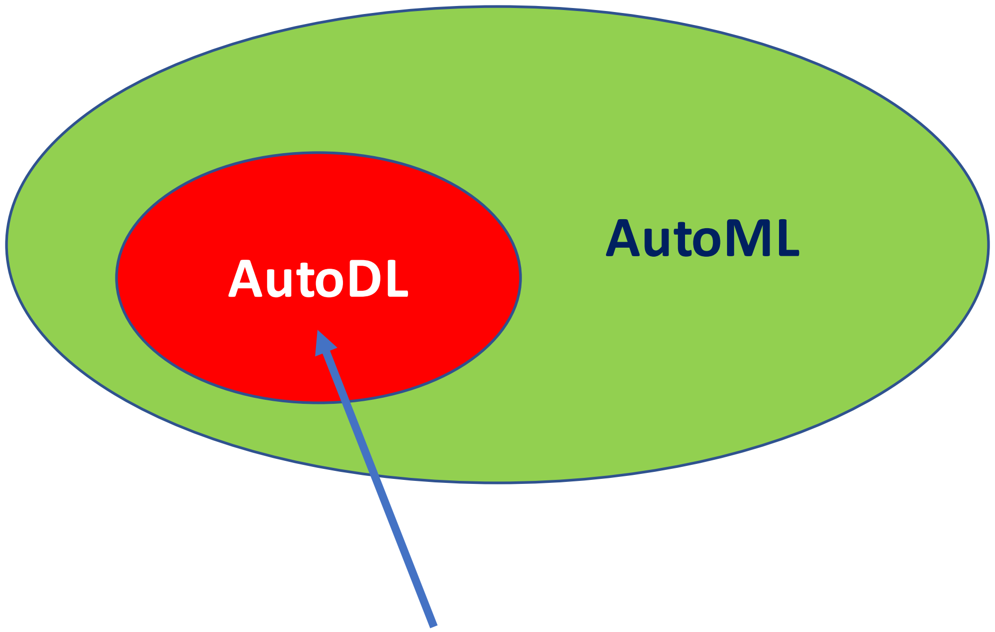


A key aspect to ensure the good performance of learning systems in real world

Automated Predictive Model Building: Related Topics

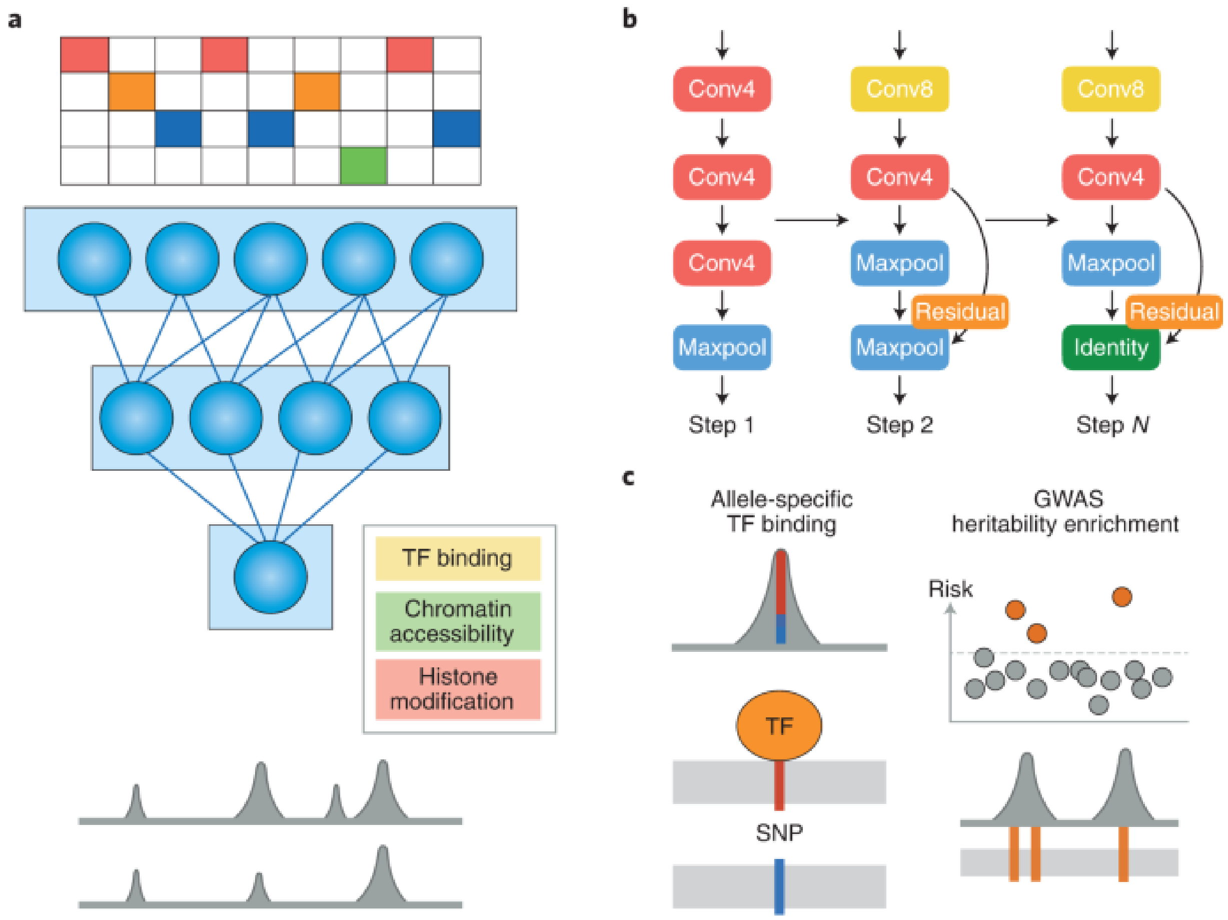
ARC **Digital Bioprocess Development** Hub

- **Automated Deep Learning (AutoDL)**



NEURAL ARCHITECTURE SEARCH

Usually applied for unstructured data, e.g., image, text, sound



Automated Predictive Model Building: Related Topics

ARC **Digital Bioprocess Development** Hub

- **Automated Deep Learning (AutoDL)**
  - Assessment Criteria for AutoDL solutions

✓ Novelty
✓ Solution quality
✓ Efficiency
✓ Stability
✓ Interpretability
✓ Reproducibility
✓ Engineering Quality
✓ Scalability
✓ Generalizability
✓ Eco-friendliness

See the following paper for more information:

X. Dong, D.J. Kedziora, K. Musial, B. Gabrys (2024) “Automated Deep Learning: Neural Architecture Search Is Not the End”, Foundations and Trends® in Machine Learning, vol. 17 (5), 767-920 (https://arxiv.org/pdf/2112.09245.pdf)

AutoML Toolboxes

ARC Digital Bioprocess Development Hub

- **Open-source Toolboxes**
  - [Auto-sklearn]
  - [AutoGluon]
  - [H2O AutoML]
  - [AutoMLPipeline]
  - [TPOT]
  - [Auto-WEKA for MCPS]
  - [Pycaret]
  - [AutoKeras]
  - [Auto-Pytorch]
  - …

Usually be used via **CODE:**
- Suitable for Developers, Data Scientists, and ML Experts

More tools can be found in:

Scriven, A., DJ Kedziora, K Musial, B Gabrys (2023) "*The Technological Emergence of AutoML: A Survey of Performant Software and Applications in the Context of Industry*", Foundations and Trends® in Information Systems: Vol. 7: No. 1-2, pp 1-252 (https://arxiv.org/pdf/2211.04148.pdf).

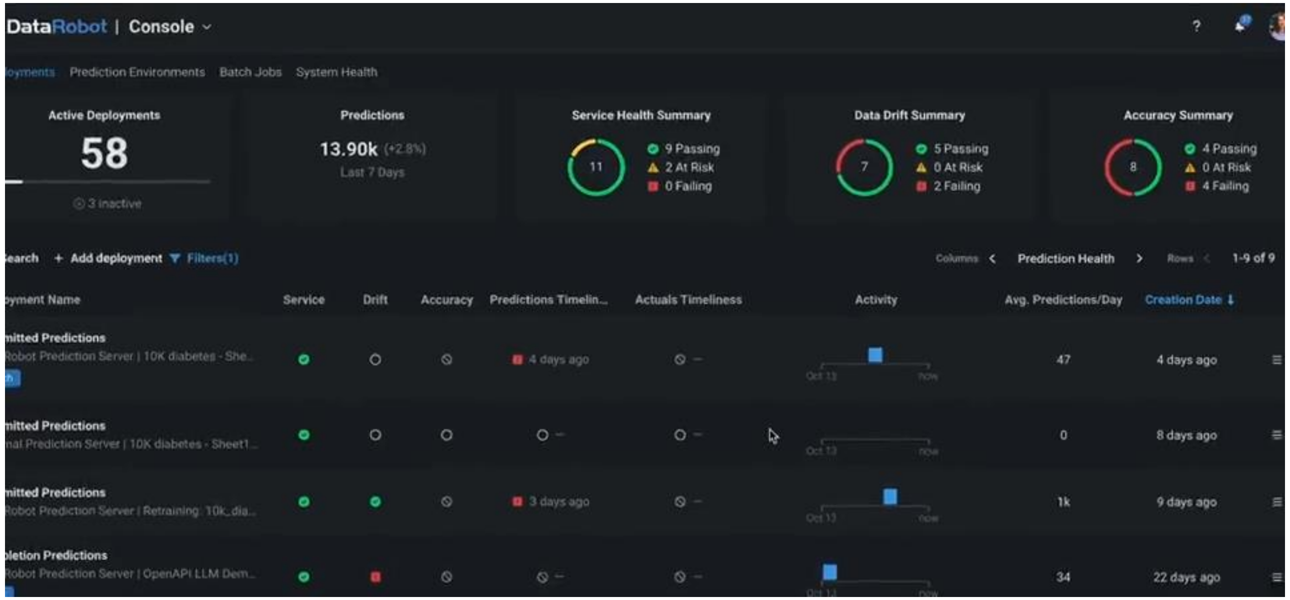
AutoML Toolboxes

ARC Digital Bioprocess Development Hub

- **Commercial Products**
  - Data Robot
  - H2O Driverless AI
  - Dataiku
  - RapidMiner
  - KNIME
  - AWS SageMaker
  - Google Cloud AutoML
  - Microsoft Azure AutoML
  - …

Good **Graphical User Interfaces**

Easy for less experienced users in ML to use them



More tools can be found in:

Scriven, A., DJ Kedziora, K Musial, B Gabrys (2023) "*The Technological Emergence of AutoML: A Survey of Performant Software and Applications in the Context of Industry*", Foundations and Trends® in Information Systems: Vol. 7: No. 1-2, pp 1-252 (https://arxiv.org/pdf/2211.04148.pdf).

ARC Digital Bioprocess Development Hub

PART 1:

QUESTIONS & ANSWERING SESSION

(5 mins)

ARC **Digital Bioprocess Development** Hub

PART 2:

INTERACTING WITH AUTOML SYSTEMS: CURRENT PRACTICES

(35 mins – Dr. Khuat)

Groups of Stakeholders Interacting with AutoML Systems (Who?)

Develop, deploy, monitor, and maintain ML solutions and AutoML systems	Translate business requirements and raw data into application inputs, verify the outputs of ML solutions	Ensure compliance around specific social requirements: user safety, model reliability, outcome fairness, ethical standards, etc.	Interact with an **AutoML-produced ML solution**
Technical group	**Business group**	**Regulatory group**	**End-user group**
Data scientist	Business analyst	Governance staff	End-users
Data analyst	Domain expert	Third-party auditors	
ML Engineer	Project manager	Government agencies	
Software developer	…	…	
System engineer			
…			

Roles of humans within the current AutoML systems (What?)

Problem Formulation & Context understanding

- Define requirements
- Define success criteria and management plans
- Define working contexts
- Resource and feasibility assessment
- Data collection
- Raw data exploration
- Data quality verification

Data engineering

- Interactive feature exploration
- Data selection
- Data imputation
- Outlier removal
- Feature transformation and visualisation
- Bias checking

Model development

- Select model types and search spaces
- Select evaluation metrics
- Visualise hyper-parameters
- Visualise ML pipelines
- Compare ML pipelines
- Evaluate performance
- Verify explainability
- Bias auditing
- Verify Business criteria

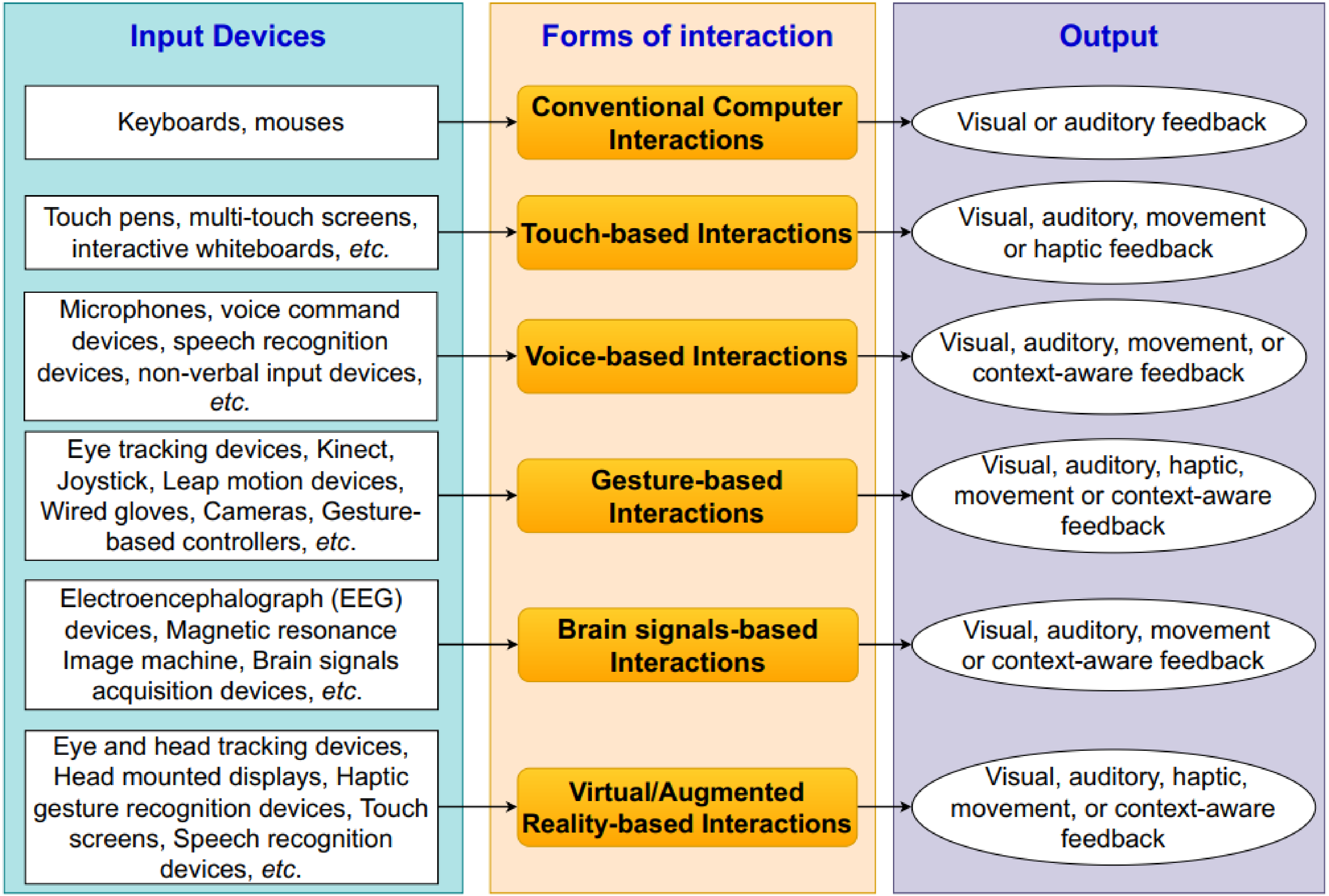
Deployment

- Evaluate the deployed models under production conditions
- Evaluate user acceptance
- Build deployment strategies

Continuous monitoring & Maintenance

- Incoming data monitoring and visualisation
- Performance tracking
- Bias auditing
- Model updating and retraining

The User Interface: Many Modalities (How?)



Communication modalities for Human-Machine Interface with ML solutions and AutoML systems

Improving the Outcomes of Interactions (Why?)

ARC Digital Bioprocess Development Hub

- **Towards Trustworthy AutoML**
  - Building trust between humans and intelligent systems
    - Allow stakeholders to foresee the behaviours of learning systems
  - Increase trust of humans about decisions generated automatically by ML algorithms
    - Users need to understand the characteristics and operations of algorithms
    - Transparency, accountability, fairness, and explainability are crucial factors
    - Accommodation of explanatory means within learning systems
  - Based on interpersonal trust in sociology
    - Intrinsic trust
      - Appears only when users can successfully understand the actual reasoning process of learning systems
    - Extrinsic trust
      - Achieve through the persuasiveness of model outcomes and other behaviours
      - Data used to test an ML model should represent the real world

Improving the Outcomes of Interactions (Why?)

ARC Digital Bioprocess Development Hub

- **The importance of explainability**
  - Most modern AutoML tools are black-box systems
    - Hard to trust the resulting ML solutions that they produce
  - Humans do not usually use systems that are not explainable, accountable, tractable and trustworthy
  - Interpretability is a prerequisite
    - To improving the fairness of learning models,
    - Assisting the anticipation of system behaviours under challenging circumstances
  - Easier for stakeholders to diagnose errors and refine the operation of the systems

Improving the Outcomes of Interactions (Why?)

ARC **Digital Bioprocess Development** Hub

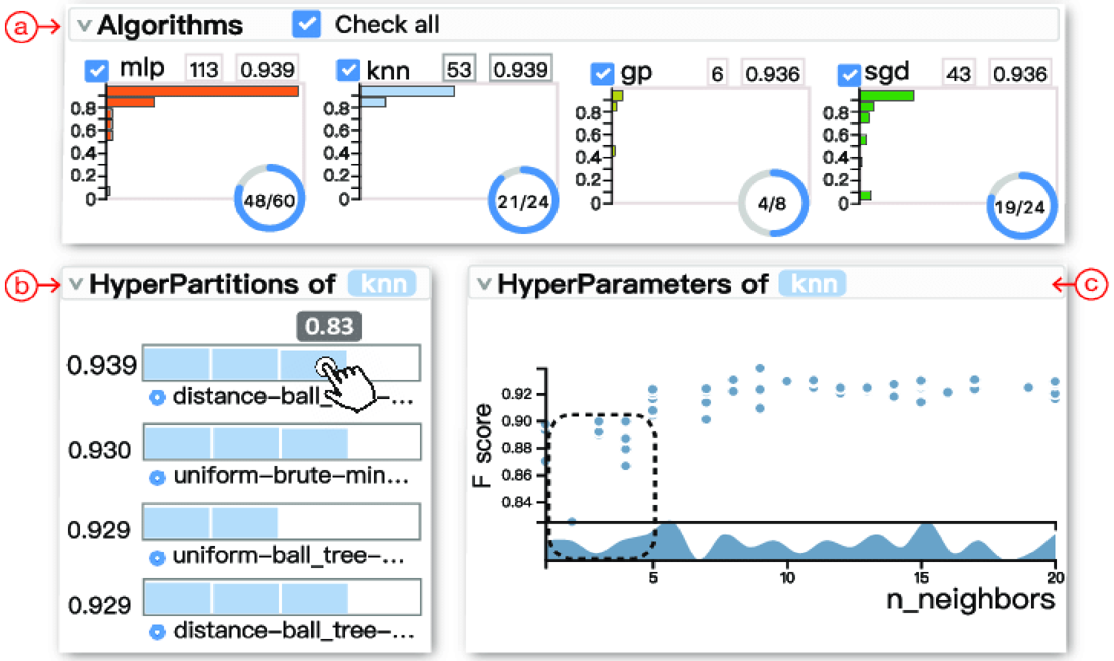
- **Methodology of Explainability**
  - In terms of **scope**, explanation methods can be divided into
    - Global type
      - Understand the entire inferential process of a learning system when producing predictive results
    - Local type
      - Explain why the learning system generates a specific outcome for a ***single*** sample
  - In terms of **stage**, explanation methods can be divided into
    - Ante-hoc type
      - Consider and enforce explainability from the beginning of model development (e.g., decision trees, rule-based systems, etc.)
    - Post-hoc type
      - Incorporate external explainers to understand the behaviours of learning systems

Improving the Outcomes of Interactions (Why?)

ARC **Digital Bioprocess Development** Hub

- **Building explanations within a purely human context**
  - Answer ‘why-questions’ as a contrastive explanation
    - ***Why was decision X made rather than decision Y?***
    - Example: **IF** *customer age* < 30 **AND** *monthly income* < $5000, **THEN** home loan is rejected

    ⇒*Contrastive* explanation: **IF** *monthly income* ≥ $5000, **THEN** home loan is accepted
  - Visualisation
    - E.g., Deploy graphs and scatter plots to illustrate searchable configuration spaces



Improving the Outcomes of Interactions (Why?)

ARC **Digital Bioprocess**
Development Hub

- **Building explanations within a purely human context**
  - Textual descriptions
    - Written or spoken format
    - Using examples in textual descriptions to make them more convinced

    "*Movie **A** directed by XYZ is likely to score a good rating because it is an action film. Three out of four action films directed by XYZ were previously rated good, i.e., C, D, E but not F. Movie **B** directed by XYZ is likely to score a bad rating because it is a drama film. One out of three drama films directed by XYZ were previously rated good, i.e., G but not H or I.*"
  - Rule-based explanations
    - Integration of fuzzy rules into learning algorithms

Improving the Outcomes of Interactions (Why?)

- **Bias Mitigation Through Human-Machine Collaboration**
  - Human judgement and expert knowledge will always be required
    - To determine standards of bias/fairness in the design, implementation, operation and deployment of ML models.
  - Human expertise may be sourced from many disciplines
    - Including the humanities, social sciences, law, and ethics
  - Human intervention is essential for assessing the accuracy of the outcomes generated by ML solutions
    - Example: An employee churn prediction at the Xerox corporation identified that a key feature inversely correlating with employment duration was ***commute time***.
    - Managers identified that ***commute time*** is ***indirectly*** a ***protected*** variable
      - The company was located in an affluent area and employees with a lower socioeconomic status were based further out.
      - Therefore, this criterion was eliminated from model inputs.

Improving the Outcomes of Interactions (Why?)

ARC Digital Bioprocess Development Hub

- **Bias Mitigation Through Human-Machine Collaboration**
  - Biases can easily slip through an ML application without close human monitoring and other collaborative interactions.
  - It is crucial to design effective human-machine collaboration via user interfaces.

The User Interface: Key Requirements

ARC Digital Bioprocess Development Hub

- **Given all the underlying concepts that we have covered, are there any general principles for designing a good AutoML UI?**
- Interactive visualization shows a critical role
  - Comprehension, diagnosis and iterative improvement of numerous learning models
  - Fill the gap between human knowledge and the insights generated by AutoML systems
  - Support technical stakeholders, enabling the inspection/control of system operations
    - Visualise and compare multiple ML pipelines
    - Know how and why AutoML algorithms construct ML pipelines for specific problems

The User Interface: Key Requirements

General Guidelines for Human-AI Interaction

Source: [Microsoft]

The User Interface: Key Requirements

FEATURE SELECTION
Enable users to determine the priority of features to be included in the model

FEATURE SELECTION
Enable users to choose features to be combined and to generate a specific rule or function to create the new values

FEATURE SELECTION
Enable users to genenerate additional data points or remove data points from the dataset based on time and space qualifiers to modify the framing of existing variables

FEATURE SELECTION
Enable users to define a rule or function to generate the missing values for a given feature

FEATURE SELECTION
Enable users to merge new features with existing data

FEATURE SELECTION
Enable users to select features to be included or excluded from the model

INSTANCE SELECTION
Enable users to choose instances that need to be included in the model

INSTANCE SELECTION
Enable users to determine parameters indicative of a subpopulation to divide the data

INSTANCE SELECTION
Enable users to define a rule or function to generate a subpopulation of the instances

INSTANCE SELECTION
Enable users to eliminate instances from the dataset

INSTANCE SELECTION
Enable users to group instances into different combinations of training and testing sets

DATA ENGINEERING

MODEL SELECTION
Enable users to determine the class of desired models

MODEL SELECTION
Enable users to select a previous solution and indicate what a new model is desired

PARAMETER SETTINGS
Enable users to determine a model and the parameter values desired

SOLUTION DESIGN
Enable users to determine types of data preparation primitives to be accommodated in a solution (e.g, data imputation)

MODEL DEVELOPMENT

UI design principles for human interactions with AutoML systems

Gil et al., “**Towards human-guided machine learning**”. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 614–624, 2019

The User Interface: Key Requirements

MODEL ASSESSMENT
Enable users to select a particular statistic test and parameters

MODEL ASSESSMENT
Enable users to request results after any step in a solution

MODEL COMPARISON
Enable users to determine multiple models to be included in solutions that have otherwise the same steps

MODEL COMPARISON
Enable users to compare two or multiple solutions for a given model but each trained on different subsets of the instances

MODEL COMPARISON
Generate comparative explanations for two given solutions

MODEL COMPARISON
Contrast two solutions in terms of the steps involved

MODEL COMPARISON
Generate comparative explanations for two given models

PARAMETER COMPARISON
Enable users to select multiple solutions with the same models or model primitive but a range of different values for one or more parameters

UI design principles for human interactions with AutoML systems

MODEL INTERPRETATION

Gil et al., "**Towards human-guided machine learning**". In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 614–624, 2019

PART 2:

QUESTIONS & ANSWERING SESSION

(5 mins)

PART 3:

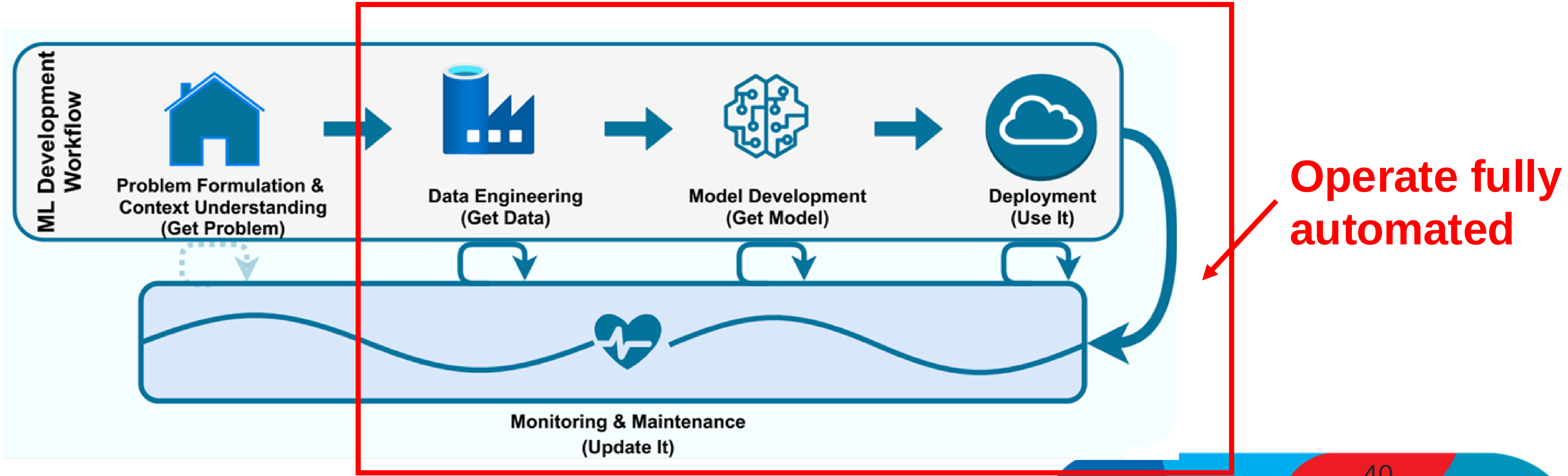
INTERACTING WITH AUTOML SYSTEMS: CONSTRAINED BUT FULLY AUTOMATED

(30 mins – Dr. Khuat)

Constrained Working Contexts for ML Systems

ARC **Digital Bioprocess Development** Hub

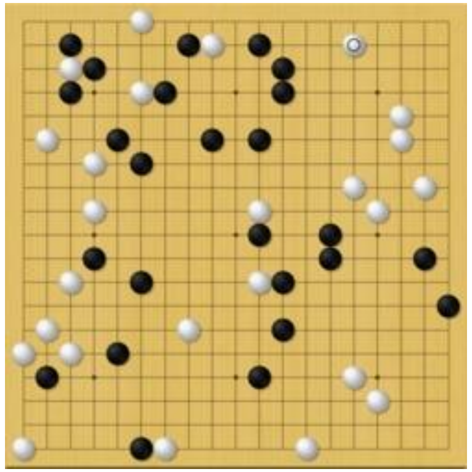
- **What is the constrained working contexts for ML models**
  - Only **problem formulation** and **context understanding phase** in a ML workflow *needs the human intervention*
  - The remaining phases in the ML workflow can operate fully automated

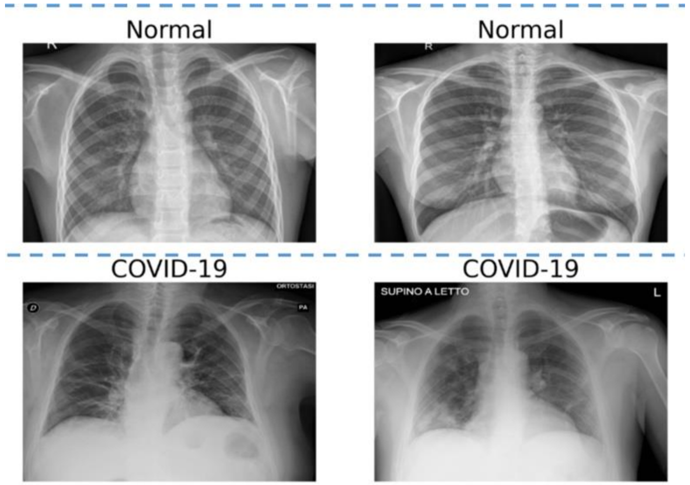


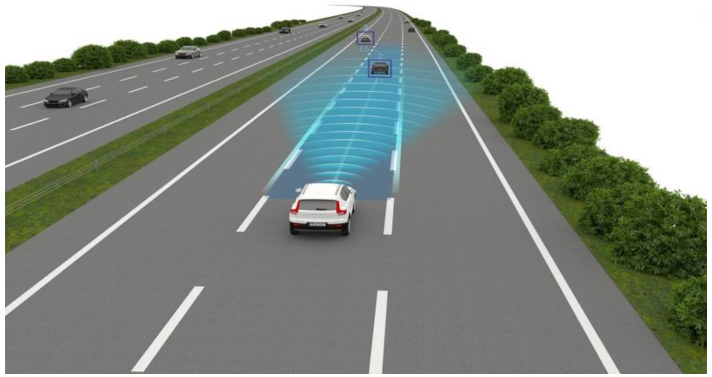
Constrained Working Contexts for ML Systems

ARC **Digital Bioprocess Development** Hub

- How humans constrain the working environments for ML systems?
  - Define the situations and working rules of the problem
  - Define how to assess the solutions
  - *Examples*: Go and chess games, driving cars on the highways, chest X-ray imaging







Constrained and Fully Automated AutoML Systems

ARC Digital Bioprocess Development Hub

- **Goals**
  - Building end-to-end ML solutions
  - AutoML systems are capable of processing almost the entire ML workflow independently
  - Truly remove the demand for ML technicians and support domain workers without ML knowledge to build ML solutions for their problems

Constrained and Fully Automated AutoML Systems

ARC Digital Bioprocess Development Hub

- **How it works**
  - Using superior search strategies
    - Reinforcement learning
    - Random search and self-playing within the constrained environments
    - *AlphaGo, AlphaZero, AlphaFold*
  - Using meta-learning
    - Meta-knowledge and experience from previous ML problems and solutions
    - The promising ML pipelines generated from previous experiences
    - Self-adaptive meta-model and meta-knowledge
  - Self-learning and adaptation to new situations where constraints still work

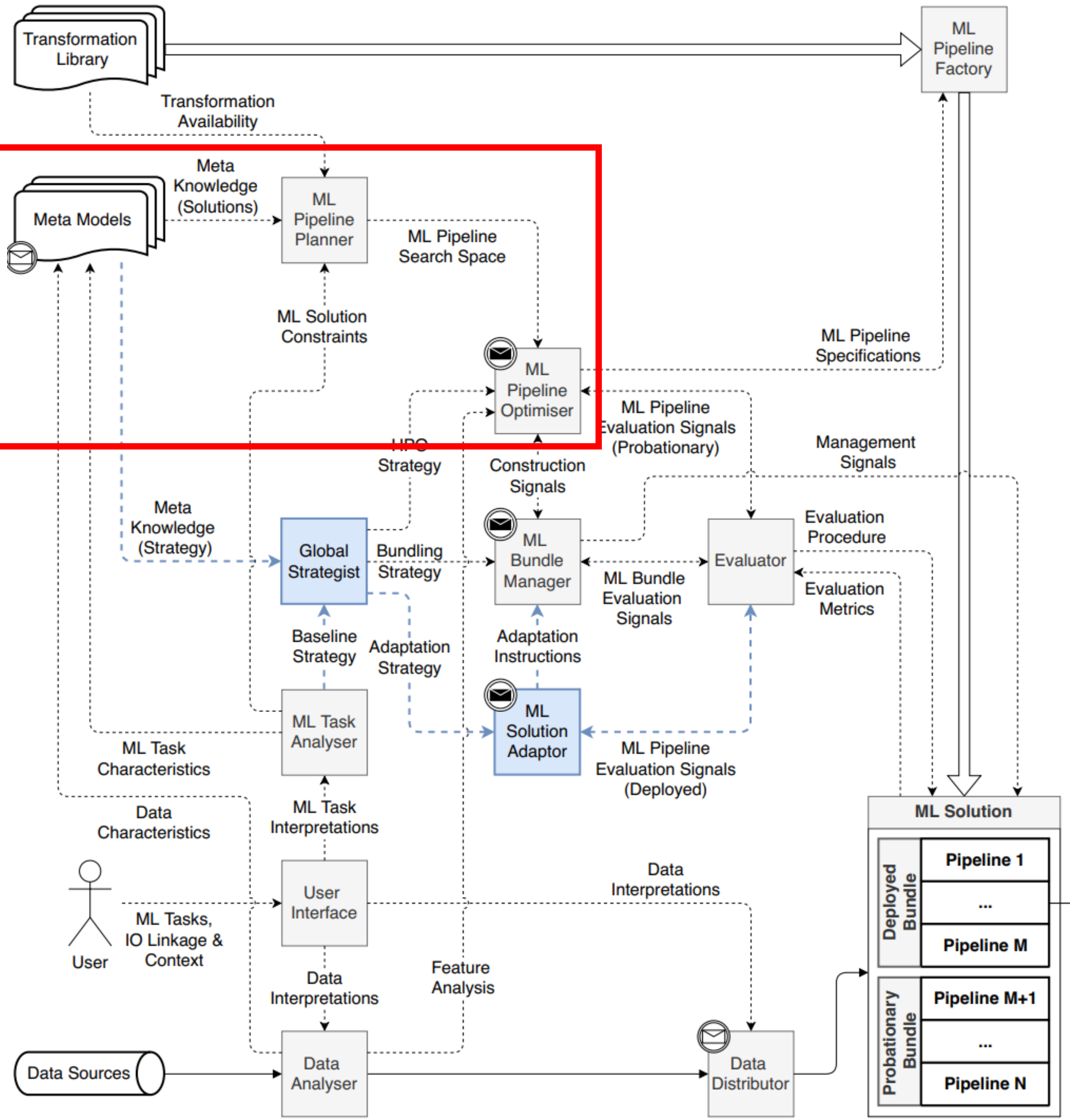
Constrained and Fully Automated AutoML Systems

ARC Digital Bioprocess Development Hub

- **Meta-learning**
  - Exploiting prior experience
  - A library stores the accumulation of previous experience
    - Solutions that worked well previously
    - Solution-finding procedures that worked well previously
  - How recognisable the current context is
  - Good suggestions are sent to the ML-pipeline planner or the ML pipeline optimizer
  - Meta-knowledge
    - Meta-features
      - Seek similarity between datasets
      - Good learning models are associated with characteristics of datasets
    - Meta-models
      - A function from meta-features dataset to a recommendation

AutonoML Framework

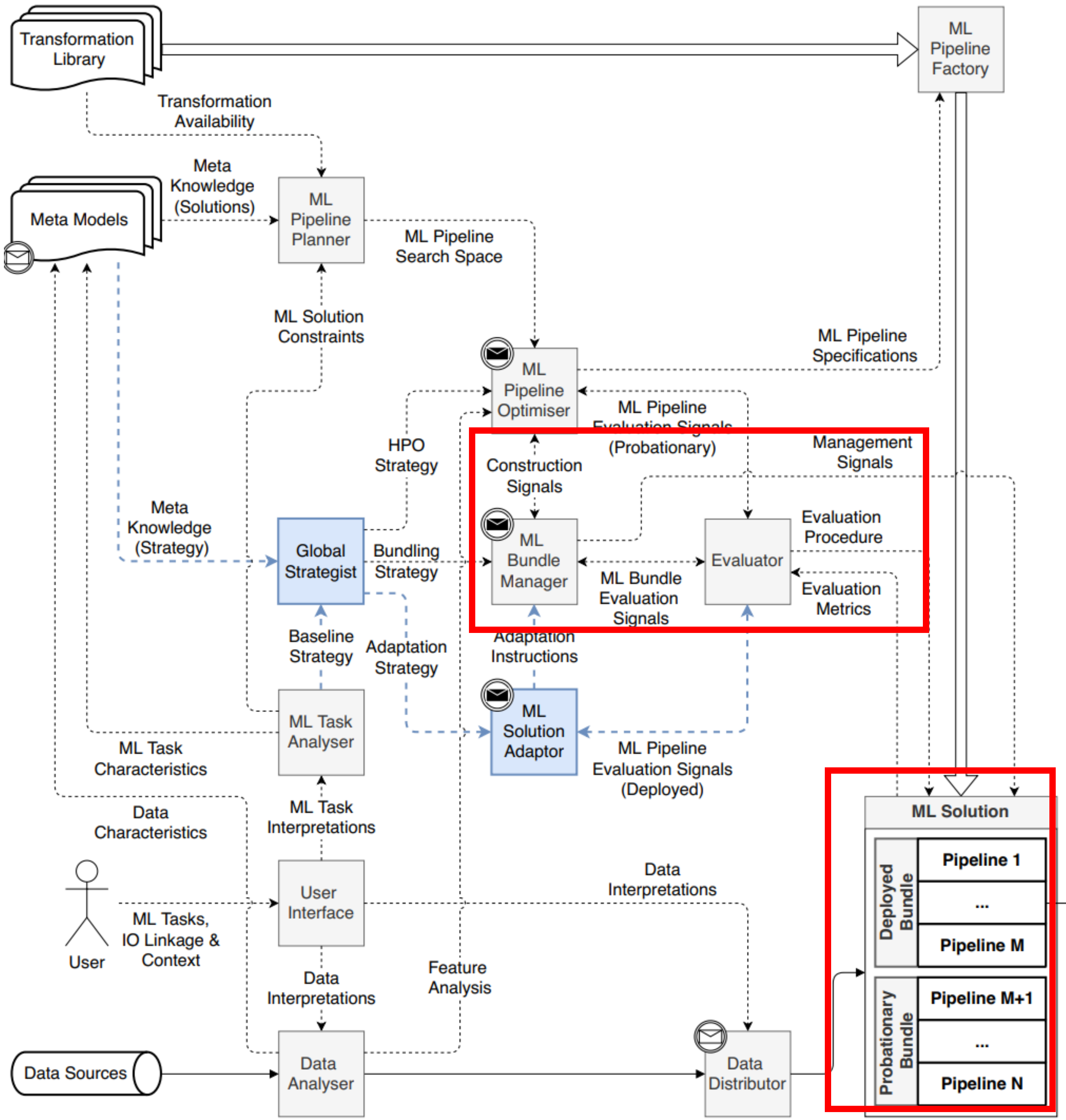
- Meta-learning is commonly used to suggest a good initial point to start searching from.
- Pipeline recommendations have also been supported by a meta-model.
- Meta-learning can also assist in selecting the active predictor within a heterogenous ensemble.



D. J. Kedziora, K. Musial, and B. Gabrys, "Autonoml: Towards an integrated framework for autonomous machine learning," *Foundations and Trends® in Machine Learning*, vol. 17, no. 4, pp. 590-766, 2024.

AutonoML Framework

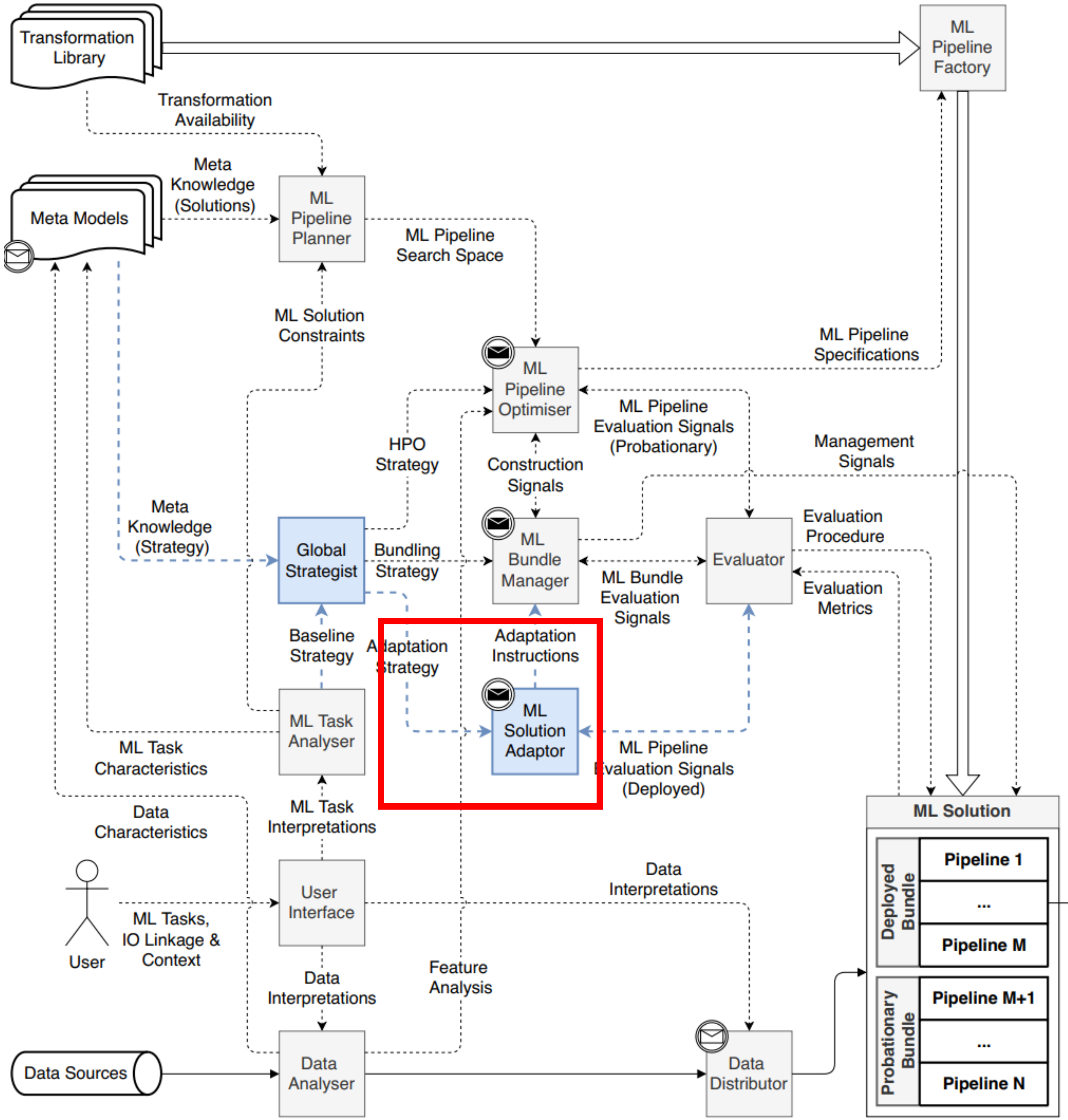
- ML bundle manager decides when to add, remove, develop or deploy ML pipelines within the arrangement.
- The bundle manager controls the aggregator, e.g., weighting the sum of outputs, and also has access to all evaluations that may assist its prescribed ensembling strategy.



D. J. Kedziora, K. Musial, and B. Gabrys, "Autonoml: Towards an integrated framework for autonomous machine learning," *Foundations and Trends® in Machine Learning*, vol. 17, no. 4, pp. 590-766, 2024.

AutonoML Framework

- An AutonoML system should be viewed as an artificial brain, continuously developing while active.
- A fundamental characteristic to define an AutonoML system is the capacity to persist and adapt.
- Adaptive strategies can be integrated into an AutonoML framework.



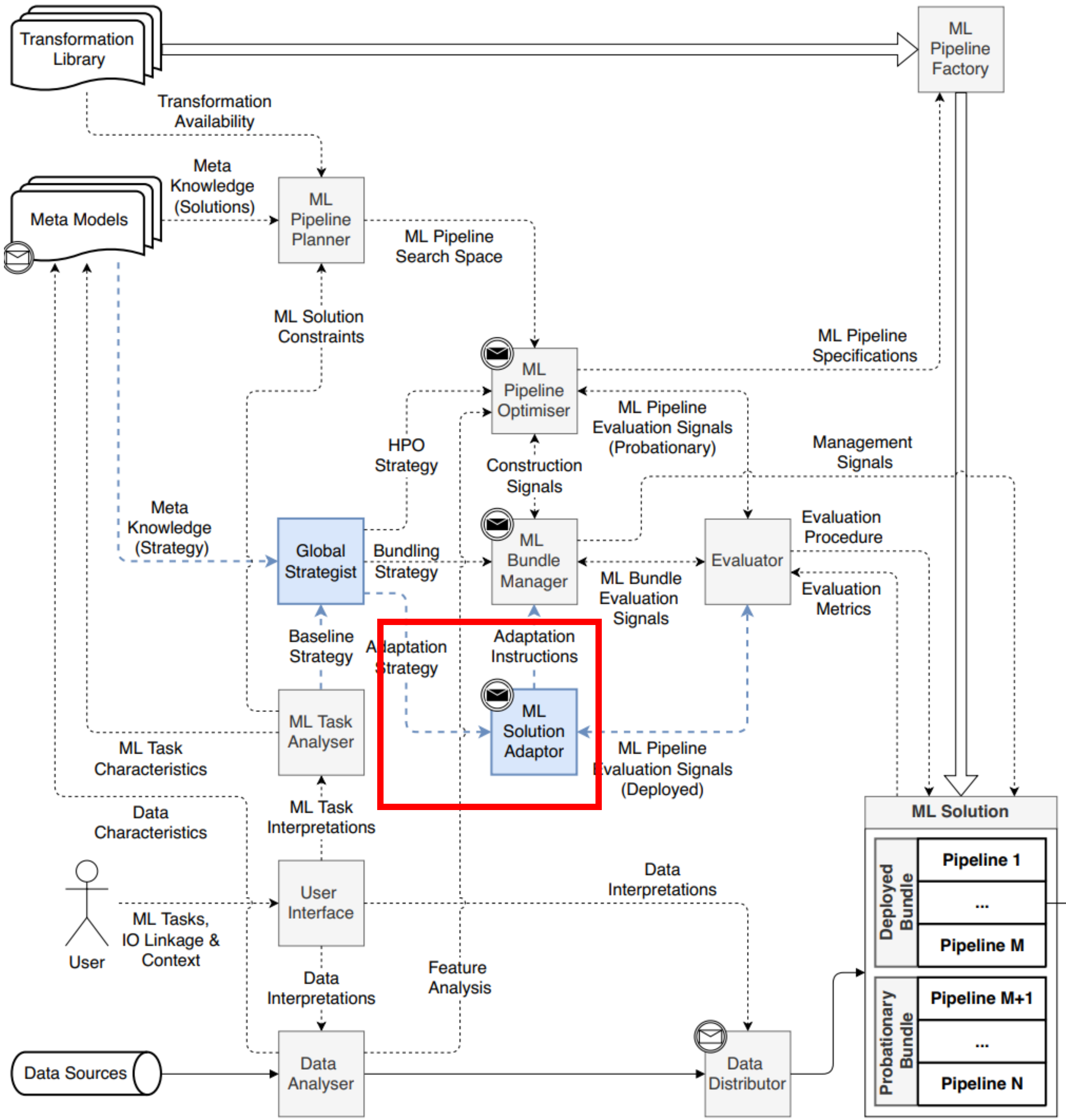
D. J. Kedziora, K. Musial, and B. Gabrys, "Autonoml: Towards an integrated framework for autonomous machine learning," *Foundations and Trends® in Machine Learning*, vol. 17, no. 4, pp. 590-766, 2024.

AutonoML Framework

ML solution adaptor module:

- Monitors the performance of individual deployed ML pipelines
- Polling the evaluator with some frequency over the lifetime of the ML solution
- When the threshold for drift is passed, it instructs the bundle manager to deal with the failing pipeline accordingly.

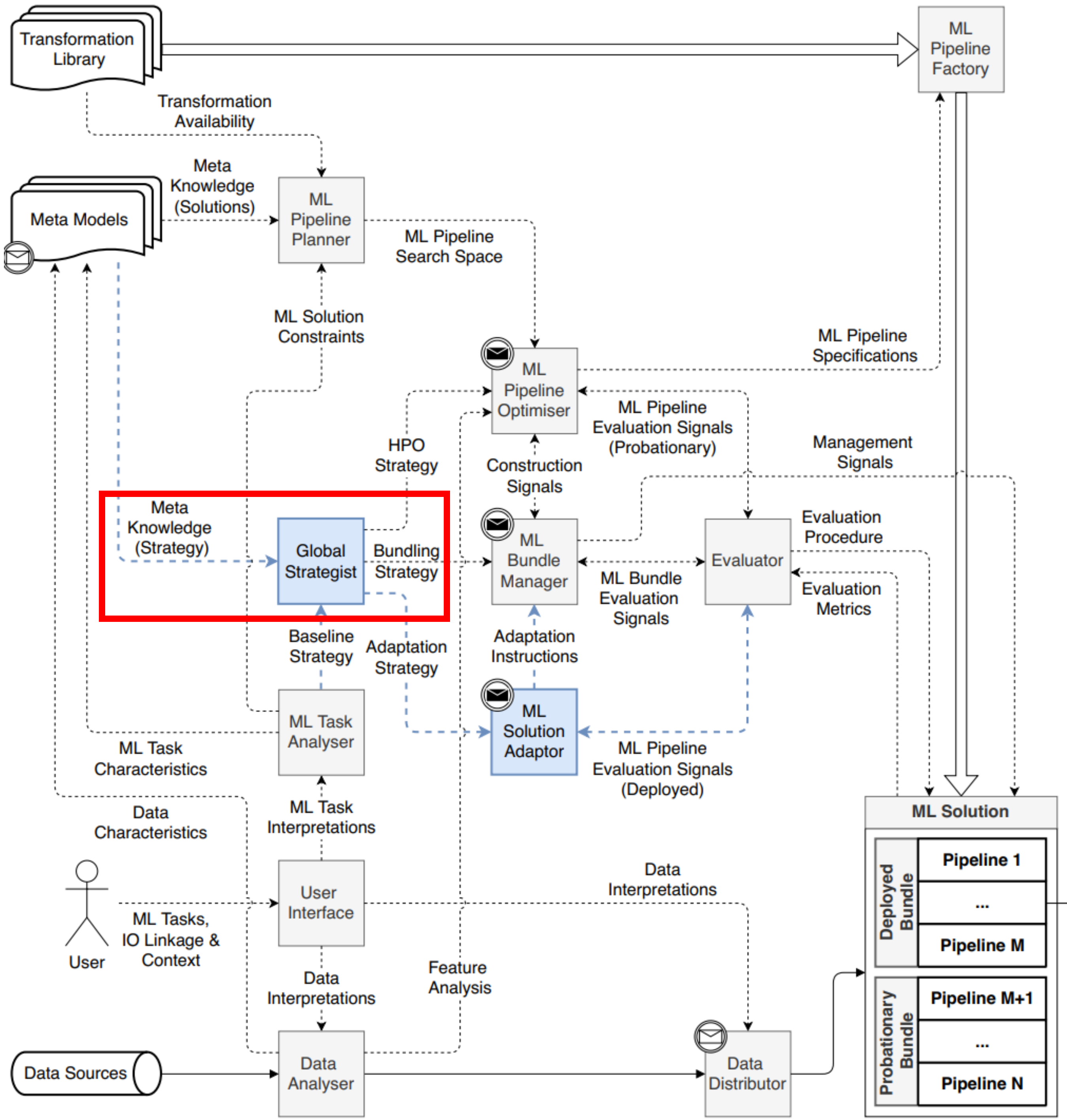
D. J. Kedziora, K. Musial, and B. Gabrys, "Autonoml: Towards an integrated framework for autonomous machine learning," *Foundations and Trends® in Machine Learning*, vol. 17, no. 4, pp. 590-766, 2024.



AutonoML Framework

The global strategist module:

- Responsible for prescribing individual tactics regarding HPO, pipeline ensembling, and adaptation.
- Meta-models feed their strategic knowledge into the global strategist for further dissemination.

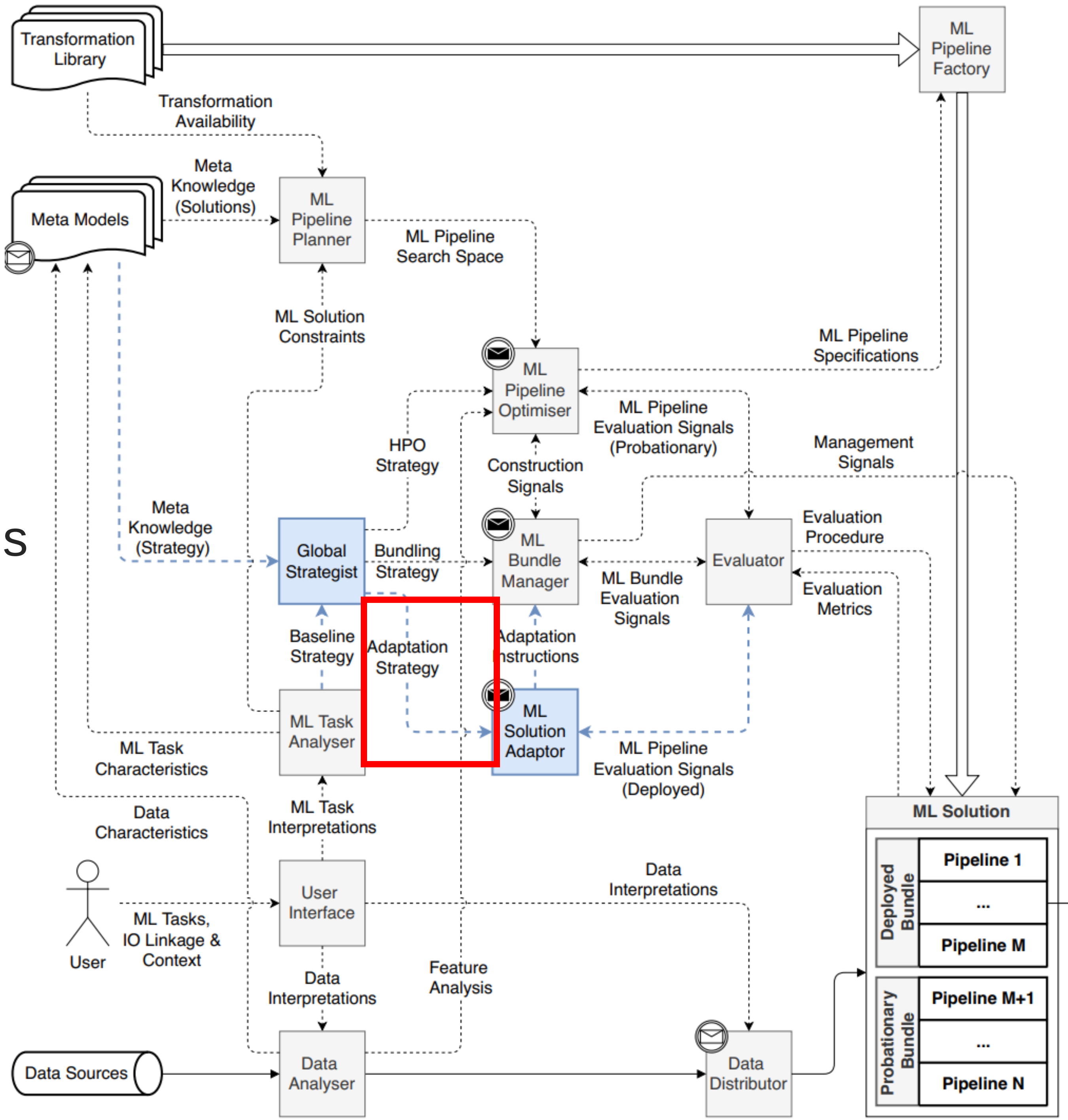


D. J. Kedziora, K. Musial, and B. Gabrys, "Autonoml: Towards an integrated framework for autonomous machine learning," *Foundations and Trends® in Machine Learning*, vol. 17, no. 4, pp. 590-766, 2024.

AutonoML Framework

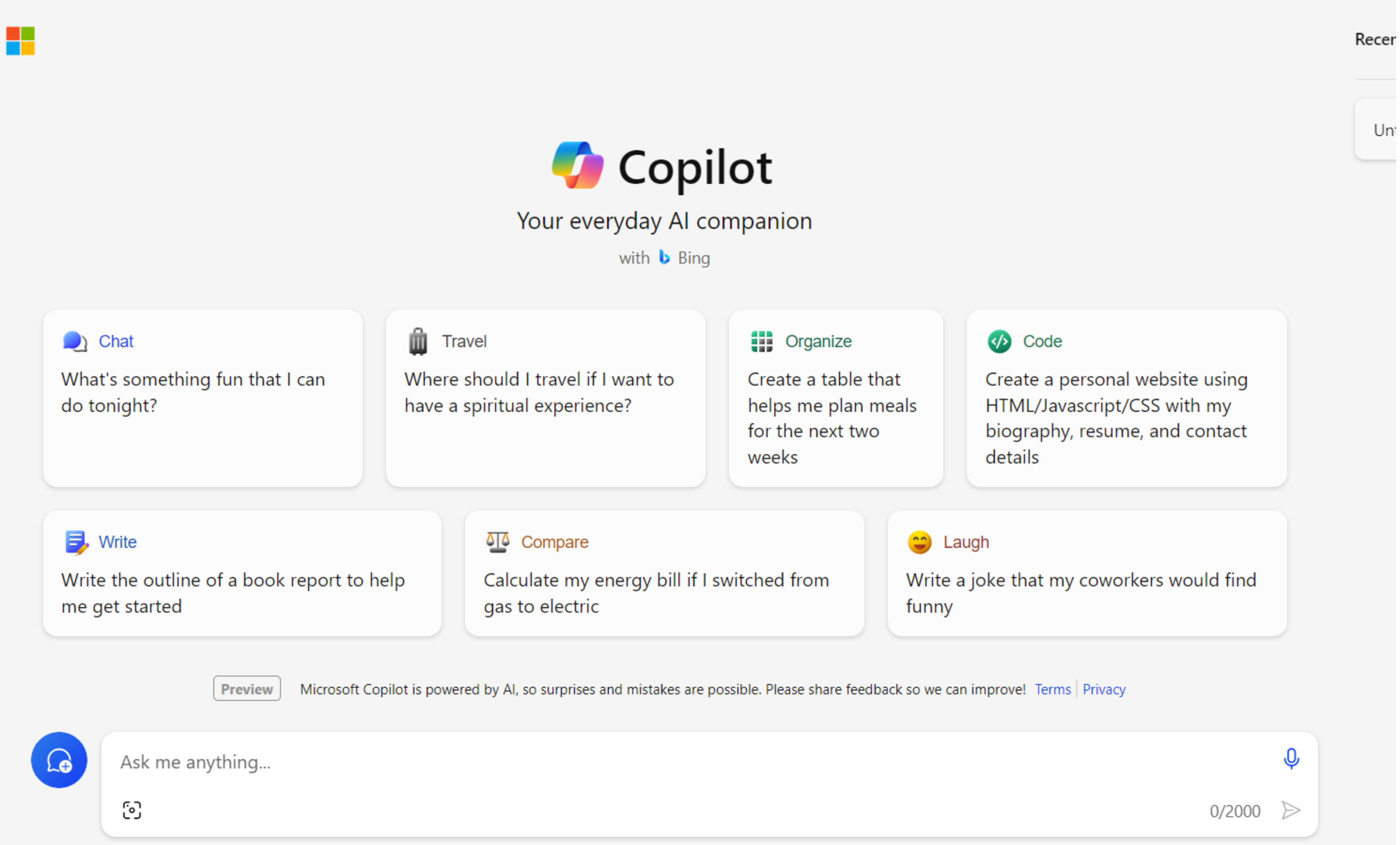
Strategies for adaptation:

- Using neuro-fuzzy theories, online-learning algorithms, instance-based incremental learners
- Employ generic adaptation mechanisms that work irrespective of base learner, e.g., Dynamic Weighted Majority and Paired Learners.
- Composite of multiple adaptive mechanisms, e.g. batch-based retraining or adding a new expert.



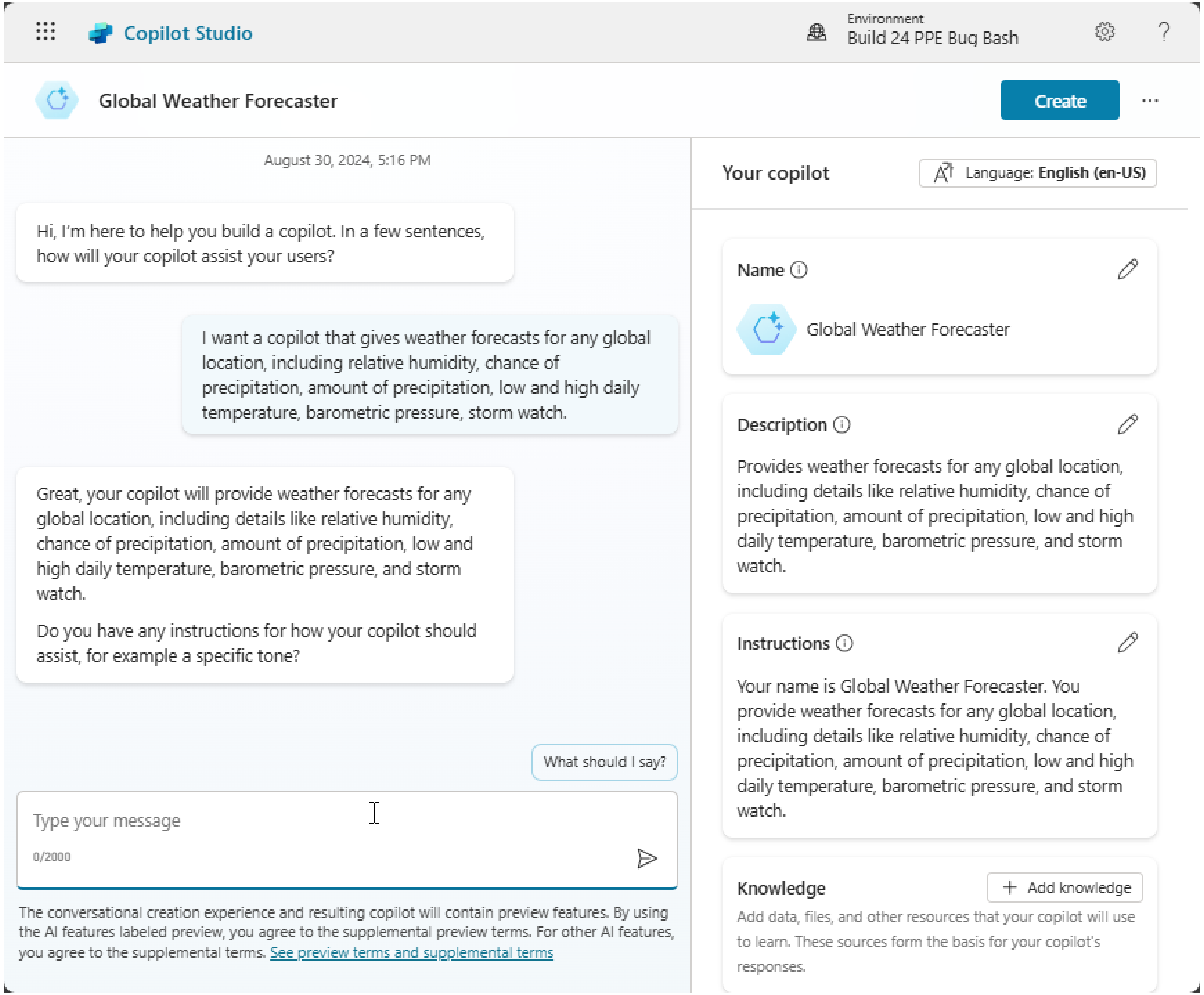
AutonoML System Example

ARC Digital Bioprocess Development Hub



AutonoML System Example

ARC **Digital Bioprocess Development** Hub



Creating a copilot agent for our specific business needs by defining requirements, functions, knowledge sources, working contexts and constraints, *etc*

Evolution of Human's Roles

ARC Digital Bioprocess Development Hub

- **Shift focus of technical and business stakeholders to problem formulation and context understanding**
  - AutonoML frameworks are capable of fully mechanising all ML workflow phases to find ML solutions
  - Free technical workers from tedious tasks, e.g., feature selection, hyper-parameters configuration
  - Human resources can be redirected to more innovative zones:
    - Refining ML task definitions/constraints
    - Sourcing more informative data sources
    - Better understanding how problem and context interrelate
    - Building the best sandbox in which AutoML will proceed to play, e.g., *constrain radiological imaging systems to specific disease types and scan formats*
  - Stakeholders no longer instruct the automated algorithms on how to solve a problem
    - Merely dedicate time and effort to ensuring that definitions/constraints are sound and watertight

Evolution of Human’s Roles

- The strengths and weaknesses of operators and AutonoML systems become complementary
  - AutonoML
    - Augmenting the capabilities of users, countering the limitations of human cognition
  - Domain experts: precisely define/constrain a problem and its working context
    - *NOT* to directly interfere with the internal mechanisms of AutonoML assistant
  - Collaborative decision-making
    - Examples: medical diagnostics, car driving
- Humans become mentors and supervisors to AutonoML systems
  - Ensure the effectiveness, safety and trustworthiness of ML systems
  - Review an entire ML workflow to understand why specific models and hyperparameters were selected and how predictive outcomes were derived

Evolution of Human's Roles

ARC **Digital Bioprocess Development** Hub

- All stakeholders become capable of improving AutonoML performance
  - A legitimate role for technicians to continue guiding AutonoML search, especially if they are privy to useful external information
  - Encoding their rich expertise to be used by AutonoML systems
    - via suggestions of promising model architectures or the refining of constraints
- Continuing legal responsibility forces humans to ensure AutonoML systems are transparent
  - Not accept black-box designs for AutonoML
  - Human operators will expect transparency to justify their usage as appropriate and legally defensible
  - Knowing to what extent model outcomes can be trusted and in which contexts they can be appropriately applied

Evolution of Human’s Roles

ARC Digital Bioprocess Development Hub

- Learning new insights from outcomes of AutonoML systems
  - AutonoML teaches humans new tricks
  - Exploring how the AutonoML system performs its job
  - Insights for how to conduct ML applications better
  - The integration of explainability mechanisms will presumably further enhance this collection of insight
  - **Examples**: find new ways to better play Go and Chess games, discover and verify new drugs

What Lies Beyond the Constraints

ARC Digital Bioprocess Development Hub

- Many real-world problems are fluid and cannot be neatly constrained
- In semi-structured and unstructured environments, rare or undesirable events and unforeseen changes in goals may substantially distort the solution
- Hard to collect sufficient data for many other ML applications
  - ML models need a large amount of data for training
- Humans can learn effectively from limited amount of data

→ AutonoML systems need the ability to learn in open-ended environments with limited amount of data, new events and situations which do not exist in training data

ARC Digital Bioprocess Development Hub

PART 3:

QUESTIONS & ANSWERING SESSION

(5 mins)

BREAK

PART 4:

INTERACTING WITH AUTOML SYSTEMS: OPEN-ENDED ENVIRONMENTS

(25 mins – Dr. Khuat)

The Prospective Evolution of Automation in ML

ARC Digital Bioprocess Development Hub

Stage in a ML workflow	Need for human involvement		
	AutoML: Current practices	AutoML: Constrained but fully automated	AutoML: Open-ended environments
Problem formulation & context understanding	Yes	Yes	**Partly**
Data engineering	Yes	No	No
Model development	Yes	No	No
Deployment	Yes	No	No
Monitoring and maintenance	Yes	No	No

The Challenges of Learning in an Open World

ARC **Digital Bioprocess Development** Hub

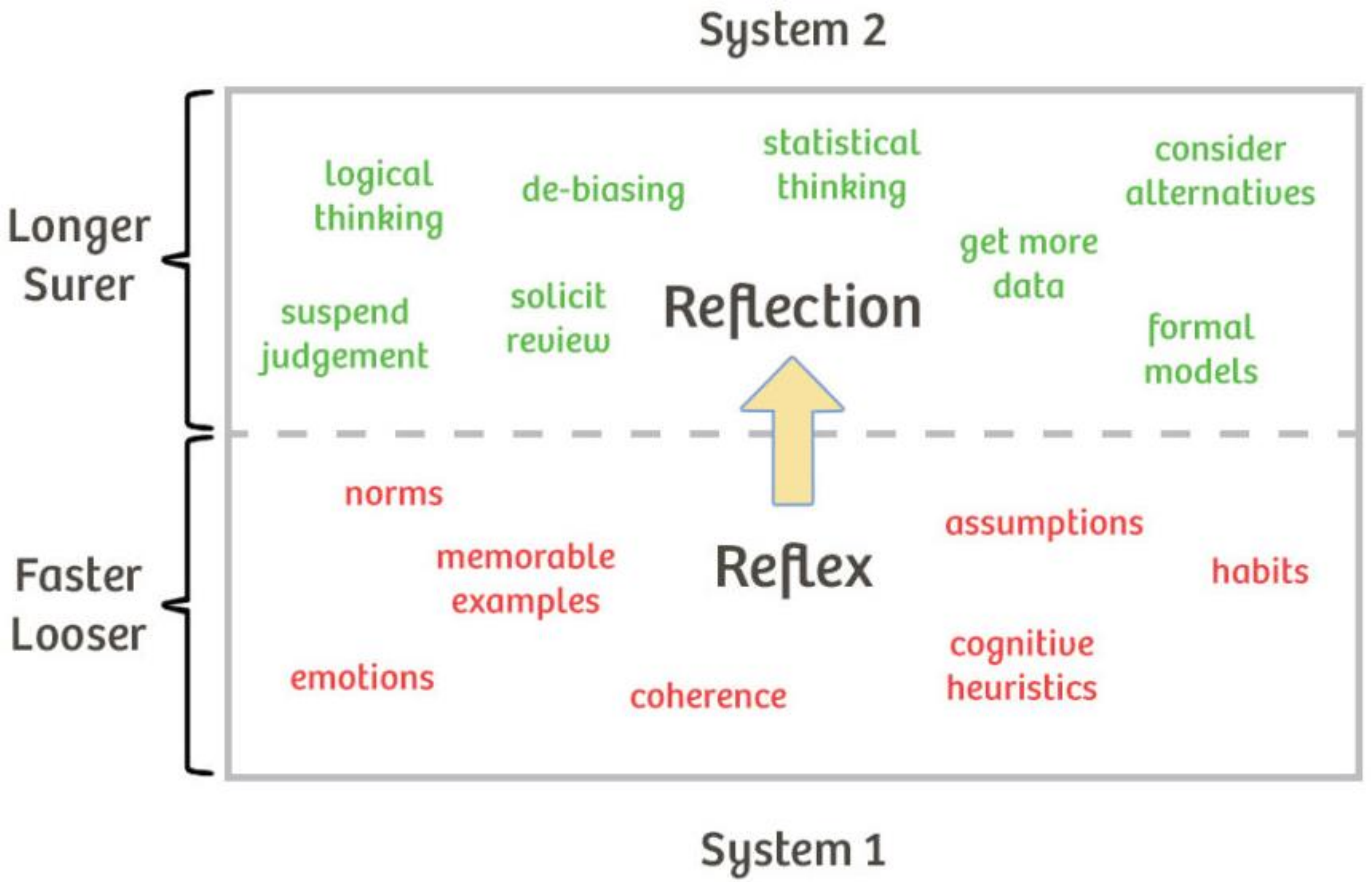
- Require the ability to identify and learn unknowns in order to become more knowledgeable
  - Goes beyond simple pattern recognition based on many previously determined tasks
  - Examples: *the ways toddlers learn: observe the world, generate a mental model regarding how it works, take action, using results to refine the mental model, repeat these steps until sufficiently interpreting the world*
- Require the ability to generate information from raw data and systematically organise it into high-level abstract concepts
- Require the ability to make decisions with incomplete and uncertain information

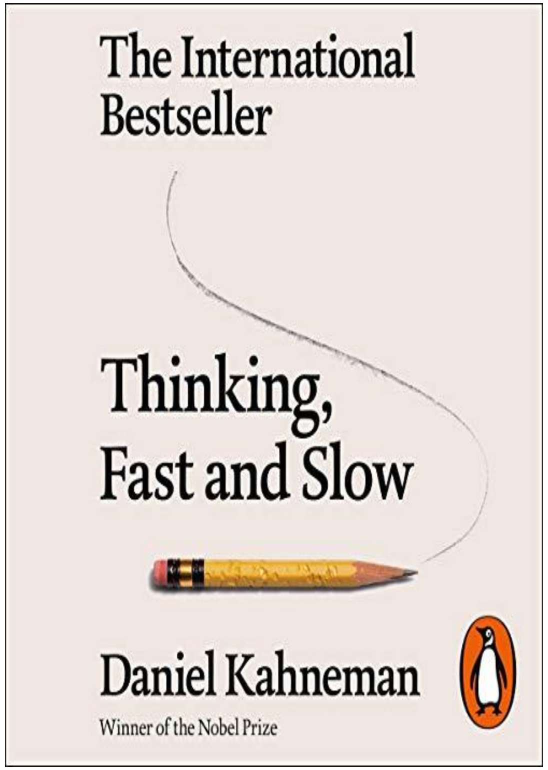
One Possible Solution: AutonoML and Reasoning

ARC Digital Bioprocess Development Hub

- How humans make decisions?

Dual-process model of cognition: Fast and Slow thinking

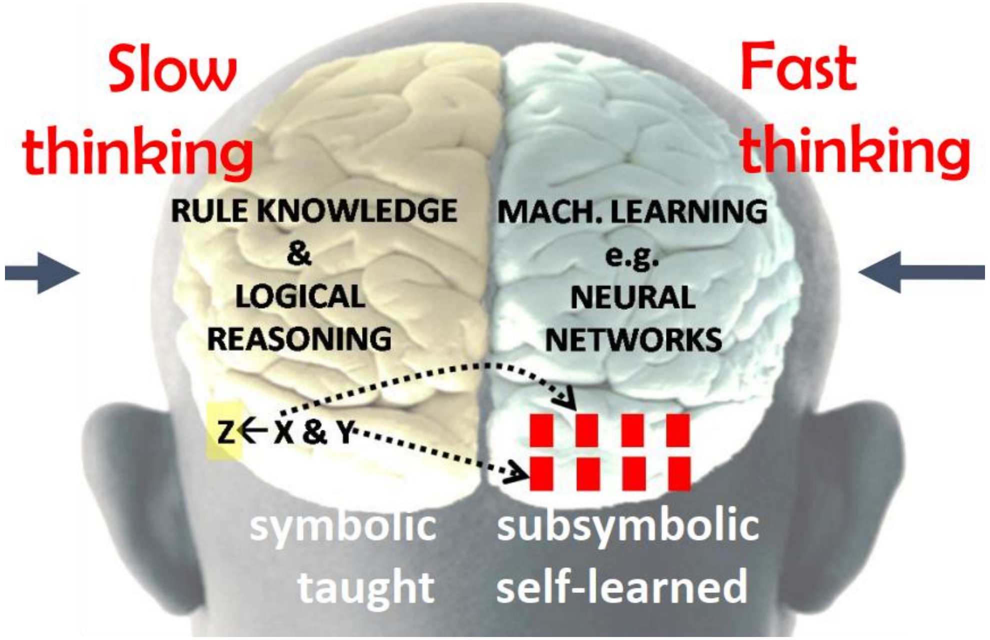




One Possible Solution: AutonoML and Reasoning

ARC Digital Bioprocess Development Hub

- The Paradigm Shift from Data-driven to Knowledge-driven
  - A hybrid learning paradigm
  - Discover, extract and generalise abstract knowledge from large and noisy datasets
  - Reasoning with this learned knowledge to better tackle new ML tasks
  - *An ideal open-world AutonoML system should lean on deductive reasoning to fill in the gaps, making assumptions, inferences and deliberation based on prior conceptual knowledge, updating and revising where need be.*



Human roles in Autonomous Open-world Systems

ARC Digital Bioprocess Development Hub

- **Humans launch ML applications in a more general manner**
  - Humans will no longer need to be as narrow and precise with definitions of an ML problem
  - Users can provide the system with terms regarding the problem
  - Example: *A user simply provides a current multi-faceted portfolio and tells the AutonoML system, “maximise money in one year”*

Human roles in Autonomous Open-world Systems

ARC Digital Bioprocess Development Hub

- **Shifting human-machine interaction from direct instruction to collaboration**
  - Humans have strengths in leadership, teamwork, creativity, and social skills
  - Automated systems are capable in speed, scalability, and quantitative computation
  - Should **form a hybrid decision-making** collective by **reaching consensus** or **compromise**
  - Stakeholders may preemptively **identify** which **tasks** and **conditions** are **safe** for **independent operation** and ***which require human interventions***
  - AutonoML systems **learn from human feedback**

Human roles in Autonomous Open-world Systems

ARC Digital Bioprocess Development Hub

- **Novel roles arise to support collaborative interactions**
  - Some humans may have to work along the interface between a team and an AutonoML system
    - Explainers
      - Translate the behaviours of smart autonomous systems to other users
    - Sustainers
      - Prevent any harmful outcomes of an autonomous system
      - Review risk analyses compiled by explainers
      - Checking/validating data, outcomes and mechanical behaviours
- **Humans monitor and validate AutonoML systems frequently and in real-time**
  - Computational systems can make decisions which conflict with human rules and value.

Human roles in Autonomous Open-world Systems

ARC Digital Bioprocess Development Hub

- **Emergency procedures are established to override undesirable AutonoML behaviours**
  - Example: Boeing 737 MAX crashes
- **Humans are forced to engineer more rigorous tests for AutonoML**
  - Powerful testing suite serves as a proactive approach to avoid disaster.
  - Continual self-evaluation and error diagnosis.

Human roles in Autonomous Open-world Systems

- **Interactions are designed to instill human/social values into AutonoML systems**
- **Society begins to debate what responsibility and trust mean concerning AutonoML**
- **Regulatory and governance mechanisms are established for AutonoML**
  - Require the learning algorithms have to be transparent, fair, accountable, and concordant with values shared by all stakeholders.

PART 4:

QUESTIONS & ANSWERING SESSION

(5 mins)

ARC Digital Bioprocess Development Hub

PART 5:

CRITICAL DISCUSSIONS AND FUTURE DIRECTIONS

(30 mins – Dr. Khuat)

Critical Discussions

ARC **Digital Bioprocess Development** Hub

- Evolution of Industry Revolution
  - Industry 4.0
    - Interconnected world
    - Construction of smart autonomous systems fuelled by big data and ML
    - An increasing level of automation and digitalisation, mainly manifested via AI-driven technologies
  - Industry 5.0
    - **Bringing** human, social and environmental factors **back** into the bigger picture
    - A collection of values including **human-centricity**, **sustainability**, **resilience**, and **ecological/social benefits**
    - **Human demands** and **interests** become **core** to the **development process** of technologies
    - The notion of *co-working* becomes very important
    - Merging human and machine abilities
    - The **integration** of **human cognitive abilities** into AI systems, and collaborative interactions

Critical Discussions

ARC Digital Bioprocess Development Hub

- Humans need to remain in the loop for AI systems
  - “We will need well-thought-out interactions of humans and computers to solve our most pressing problems” (Michael Jordan)
  - Ensuring AutoML usage remains socially responsible is also a strong motivation for efficient and productive human-machine communications
  - Even if an AutonoML system can solve a problem ‘perfectly’, human involvement is indispensable for ensuring that form of perfection accords with the definition a stakeholder chooses.
- ML research should not only prioritise a mimicry of human thinking in the pursuit of AGI
  - Human intelligence may not be the perfect standard for problem-solving
  - Automation of context understanding may not be achieved without installing AutonoML with the ability to reason, transfer and generalize knowledge on a higher level than data??

Critical Discussions

- If a fully autonomous open-world system in the future completely does not need inputs from humans:
  - We need to define explanatory mechanisms that allow stakeholders.
    - Understand system behaviours,
    - The rationale behind decision-making processes
  - We need to allow users to at least determine whether it is worth accepting an ML solution.
  - Explainability is critical to support understanding and build confidence, e.g. allowing for the diagnosis of odd behaviour.

Potential Research Directions

ARC Digital Bioprocess Development Hub

- The most challenging issues
  - How to leverage the full power of autonomous machines
    - While simultaneously supporting seamless coordination with humans, generating decisions that align with sophisticated values and needs.
  - How to perform a transition in thinking from a purely technical focus to one that
    - Ponders human-centric solutions
    - Mirrors gravitation in the broader data science literature towards the concept of Industry 5.0.

Potential Research Directions

ARC Digital Bioprocess Development Hub

- Generating human-centred design principles for the selection of optimal ML solutions
    - The specification and comprehension of contexts and requirements
    - Developing a solid understanding of
        - Who the relevant stakeholders are.
        - What they care about.
        - How they are expected to interact with an AutoML system.
        - What the working environments and constraints for the framework are.
- Promoting transparency and explainability in AutoML systems
    - Moving towards transparency with explanatory User Interfaces and white-box system design.

Potential Research Directions

ARC Digital Bioprocess Development Hub

- Integrating human values in terms of fairness and ethics into AutoML systems
  - Ensure processes and outcomes of AutonoML systems adhere to human requirements.
  - Devise realistic solutions to define fairness, bias, and discrimination within the context of AI in a simple fashion.
  - Provide normative justification for how and why the outcomes of AutonoML systems are consistent with human values.
- Establishing concepts of operation for human-machine interaction in AutoML/AutonoML systems
  - The determination of tasks, the responsibilities of stakeholders, the degree of interrelation among operations between human operators and AutonoML systems.
  - How to support collaborative interactions with UX design and other protocols.

Potential Research Directions

ARC Digital Bioprocess Development Hub

- Building common protocols and design principles for open-world ML
  - Contemplate how best to incrementally open up the constraints of ML problems, ensuring theory and practice can handle relevant tasks before advancing even further.
- Embedding situational and context awareness into AutonoML
  - An autonomous system must construct and maintain a model of itself and its operational environment, at least for relevant factors of interest.
  - Facilitate dynamic information flow and integrate low-level data into the cognitive model of AutonoML systems.
  - The situational awareness model must be able to actively learn from new insights captured by the system.
  - The model needs an effective well-structured representation of existing situations for quick recall, in real-time, whenever a similar situation is encountered.

Potential Research Directions

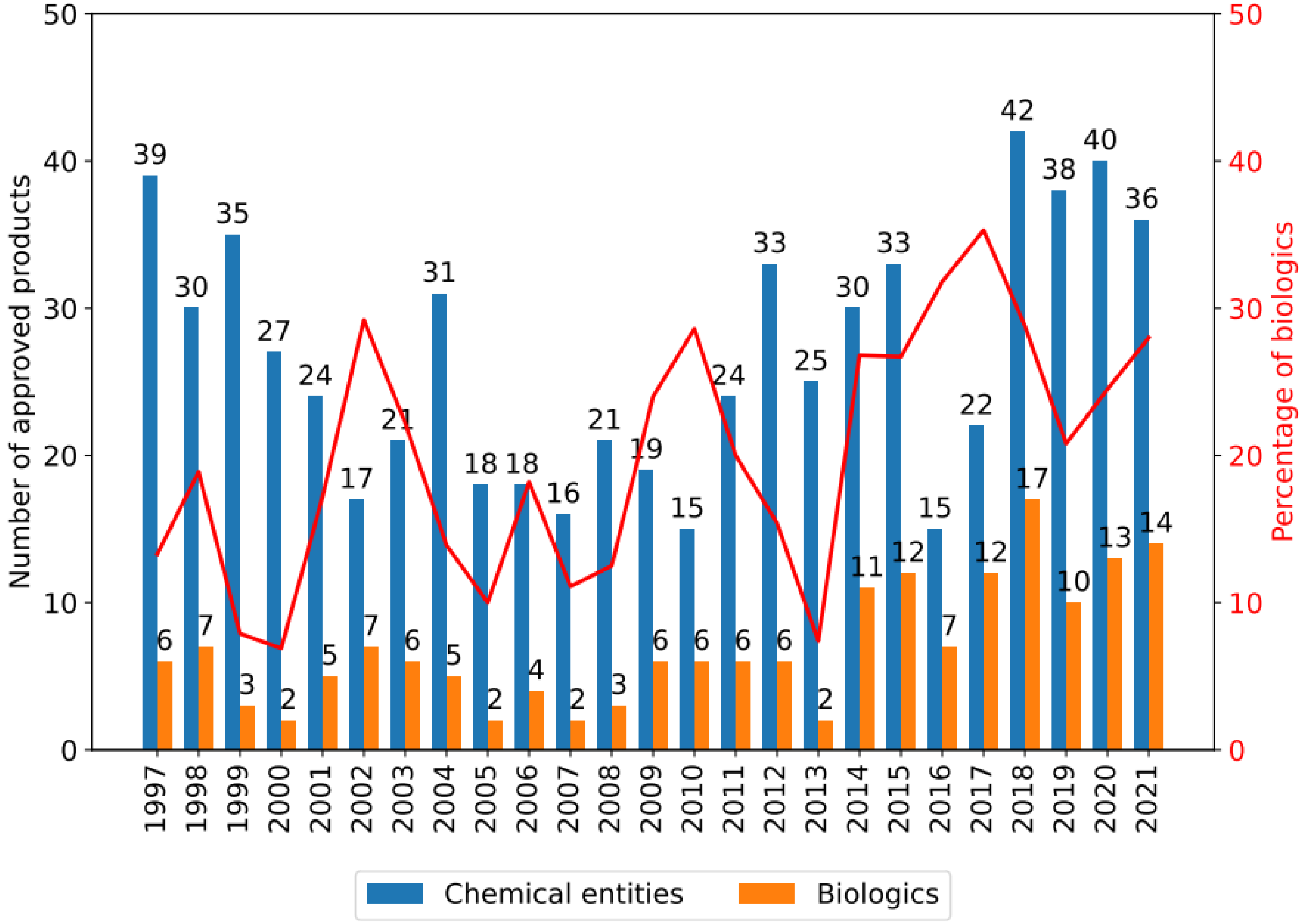
ARC Digital Bioprocess Development Hub

- Fusing data-driven search methods with higher-level reasoning
  - Humans can transfer knowledge and insight among tasks, adapt to changes in a dynamic environment, reason causally and with abstraction, and logically generalise concepts rather than simply recognise patterns and detect associations.
  - Advancing the capabilities of AutonoML beyond situational awareness may require finding ways of integrating knowledge-driven learning and reasoning approaches with data-driven search algorithms.

A Potential Application of AutonoML in an End-to-End Digital Twin of Bioprocesses for Monoclonal Antibody (mAb) Development and Manufacturing

An Introduction to Biopharma

ARC Digital Bioprocess Development Hub



New drugs approved by FDA over the last 25 years and the percentage of biologics out of the total drug approved each year

The market for **biologics** has explosively grown with a percentage of new biological products approved by FDA every year since 2014 for treating various human diseases including cancers, autoimmune, metabolic and infectious diseases, always exceeding **20%** of the total number of new approved drugs

An Introduction to Biopharma

ARC Digital Bioprocess Development Hub

Top ten drugs by sales globally in 2023. The biological products are highlighted in red.

Drug name	Manufacturer(s)	Sales 2023 ($US Billion)
Keytruda (pembrolizumab)	Merck	$25.011
Humira (adalimumab)	AbbVie	$14.404
Ozempic (semaglutide)	Novo Nordisk	$13.892
Eylea (aflibercept)	Regeneron, Bayer	$12.876
Eliquis	Bristol Myers Squibb	$12.206
Dupixent (dupilumab)	Sanofi	$11.590
Biktarvy	Gilead Sciences, Inc.	$11.850
Comirnaty	Pfizer, BioNTech	$15.305
Stelara (ustekinumab)	Johnson & Johnson	$10.858
Darzalex (daratumumab) & Darzalex Faspro	Johnson & Johnson	$9.744

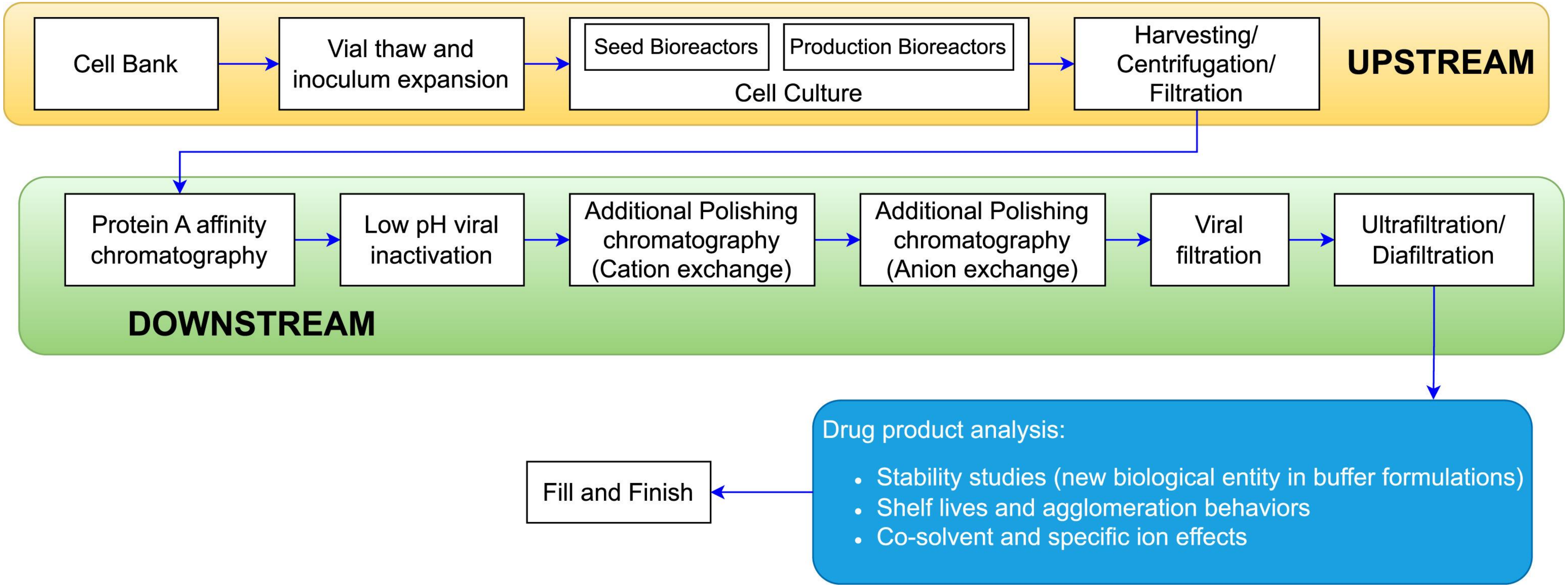
Monoclonal antibodies (mAbs) emerge as the leading product in the rapidly growing market of high-valued biologics

- **The global mAb market is expected to reach a revenue of around US$500 billion by the end of 2030**

An Introduction to Biopharma

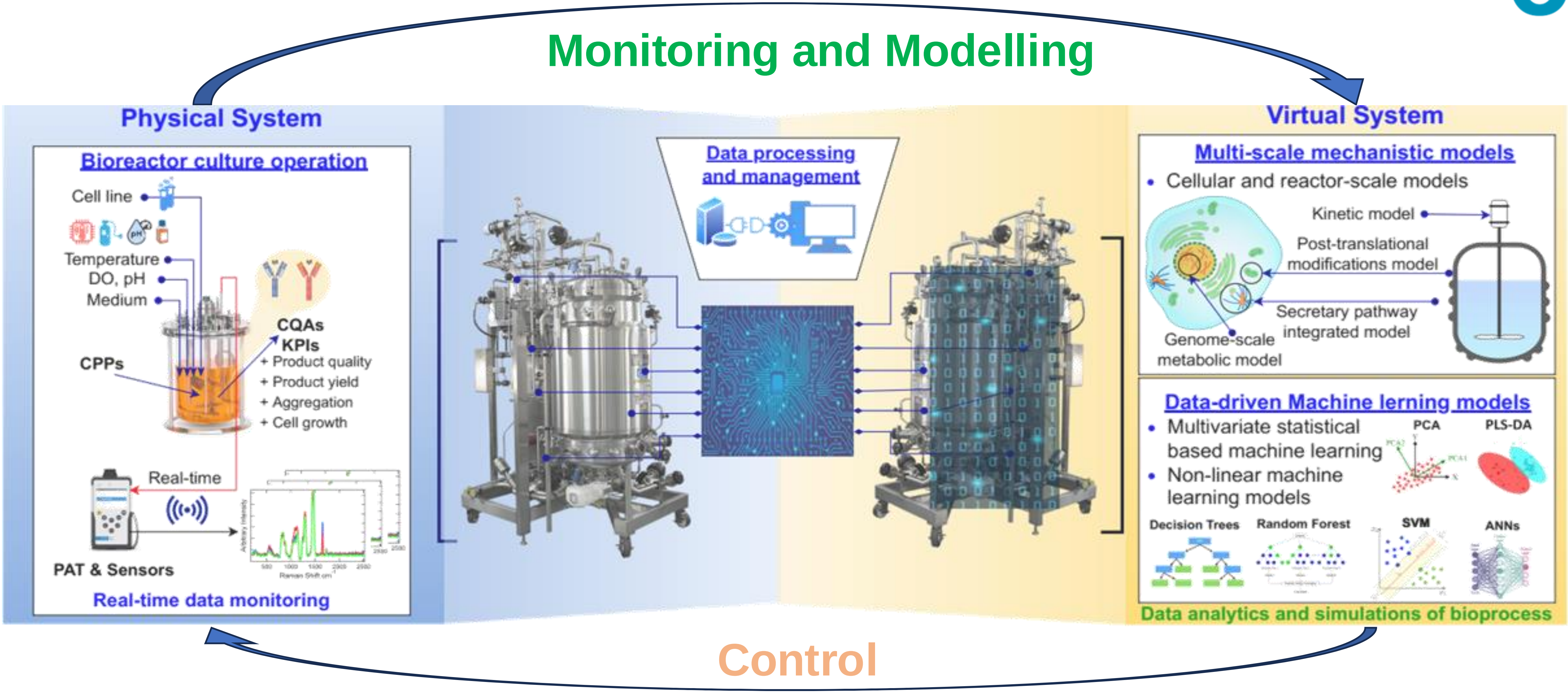
ARC **Digital Bioprocess Development** Hub

- A typical bioprocess for production of mAbs



Potential applications of AutonoML towards BioPharma 4.0

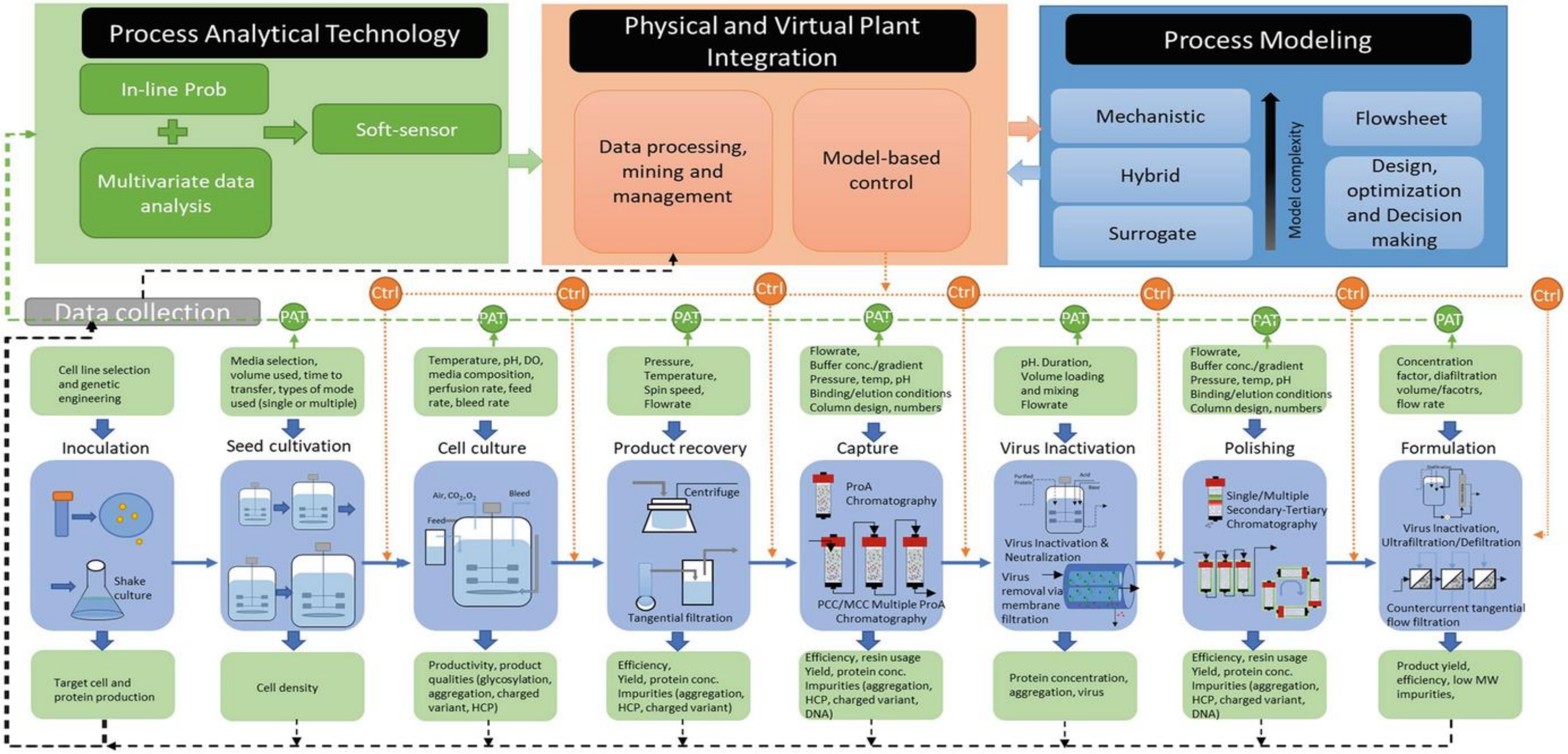
ARC **Digital Bioprocess Development** Hub



An **Upstream** Cell Culture Digital Twin

Potential applications of AutonoML towards BioPharma 4.0

ARC **Digital Bioprocess Development** Hub



A Fully integrated Digital Twin framework for the whole biopharmaceutical manufacturing process

Source: Towards Digital Twin for Biopharmaceutical Processes: Concept and Progress

ARC Digital Bioprocess Development Hub

PART 5:

QUESTIONS & ANSWERING SESSION

(5 mins)

ARC Digital Bioprocess Development Hub

Reference

1. T. T. Khuat, D. J. Kedziora, and B. Gabrys, "The roles and modes of human interactions with automated machine learning systems: A critical review and perspectives," *Foundations and Trends® in Human-Computer Interaction*, vol. 17, no. 3-4, pp. 195-387, 2023. http://dx.doi.org/10.1561/1100000091
2. D. J. Kedziora, K. Musial, and B. Gabrys, "Autonoml: Towards an integrated framework for autonomous machine learning," *Foundations and Trends® in Machine Learning*, vol. 17, no. 4, pp. 590-766, 2024. http://dx.doi.org/10.1561/2200000093
3. A. Scriven, D. J. Kedziora, K. Musial, and B. Gabrys, "The technological emergence of automl: A survey of performant software and applications in the context of industry," *Foundations and Trends® in Information Systems*, vol. 7, no. 1-2, pp. 1-252, 2023. http://dx.doi.org/10.1561/2900000030
4. X. Dong, D. J. Kedziora, K. Musial, and B. Gabrys, "Automated deep learning: Neural architecture search is not the end," *Foundations and Trends® in Machine Learning*, vol. 17, no. 5, pp. 767-920, 2024. http://dx.doi.org/10.1561/2200000119
5. T.T. Khuat, R. Bassett, E. Otte, A. Grevis-James, and B. Gabrys. "Applications of machine learning in antibody discovery, process development, manufacturing and formulation: Current trends, challenges, and opportunities," *Computers & Chemical Engineering*, vol, 182, pp. 108585. https://doi.org/10.1016/j.compchemeng.2024.108585

ACKNOWLEDGEMENTS

ARC Digital Bioprocess Development Hub

THANK YOU FOR JOINING THIS TUTORIAL

Australian Government
Australian Research Council

THE UNIVERSITY OF MELBOURNE

UTS

RMIT UNIVERSITY

CSL

cytiva

patheon by Thermo Fisher Scientific

SARTORIUS

Mass Dynamics

INSILICO

UNIVERSITY OF TARTU

Utrecht University

University of Nottingham UK | CHINA | MALAYSIA

WE EXPECT FURTHER DISCUSSIONS