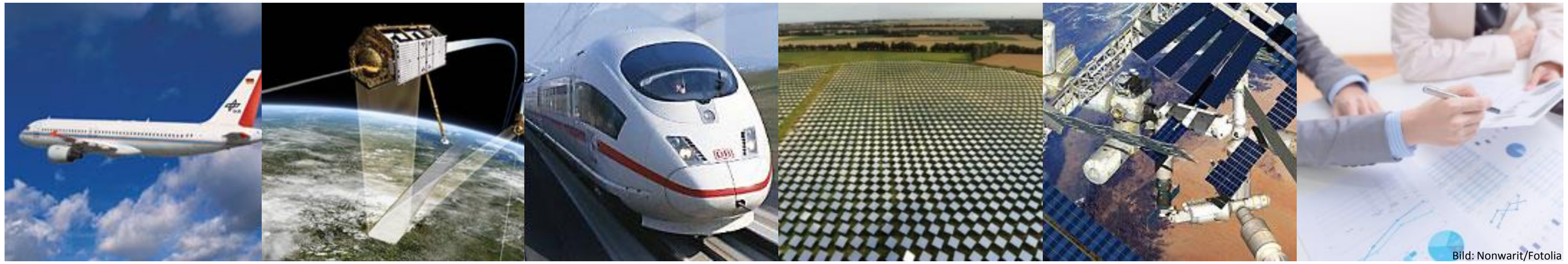
Data Science Introduction

Dr.-Ing. Achim Basermann
German Aerospace Center (DLR)
Institute for Software Technology
Department Head “High-Performance Computing”

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

DLR

German Aerospace Center



- Research Institution
  - Research and development in aeronautics, space, energy, transportation, digitalization and security
  - National und international cooperations

- Space Agency
  - Planing and implementation of German space activities

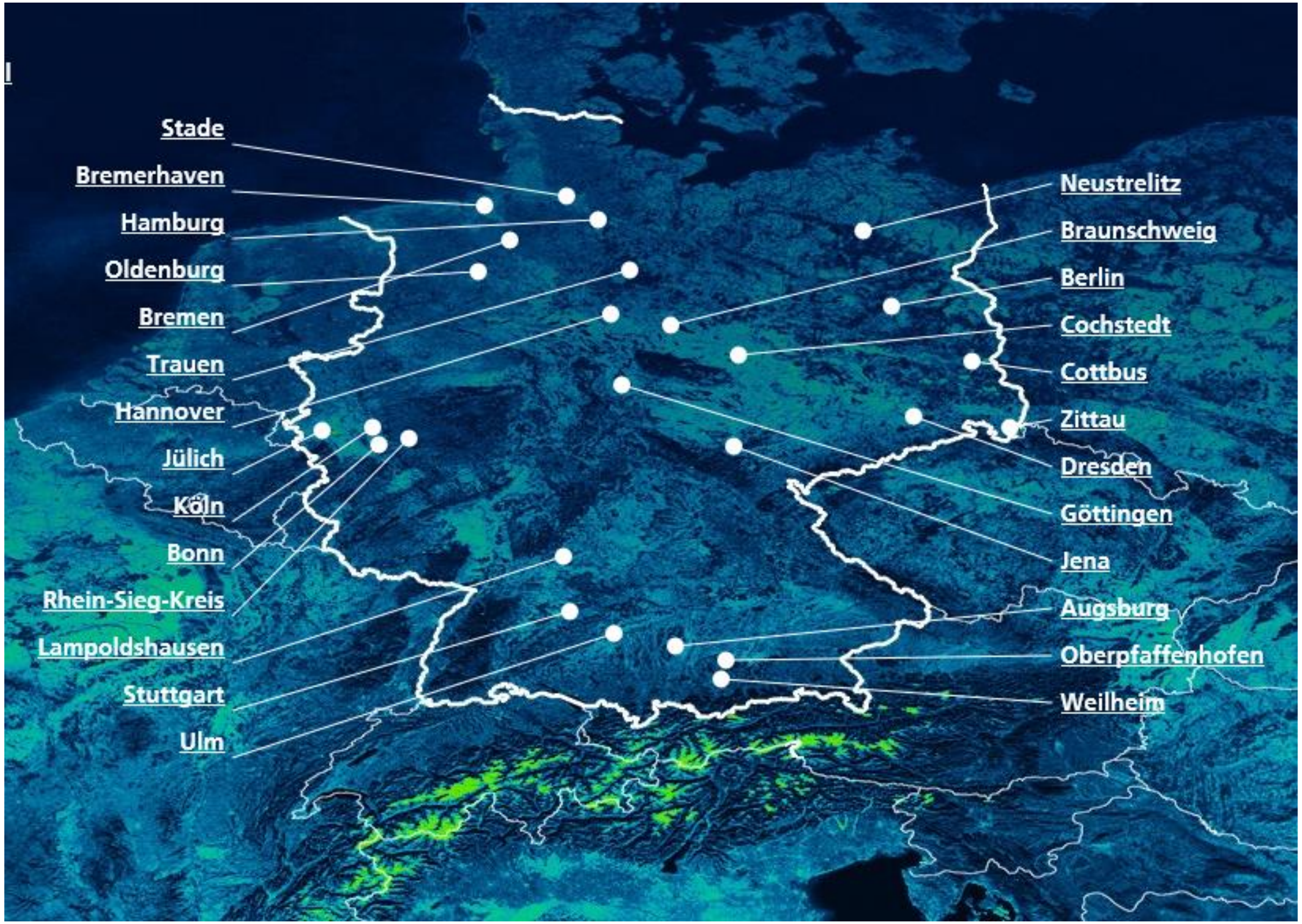
- Project Management Agency
  - Research promotion

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

DLR Locations and Employees

Approx. 9000 employees across
51 institutes and facilities at 28 sites.

Offices in Brussels, Paris,
Tokyo and Washington.



HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

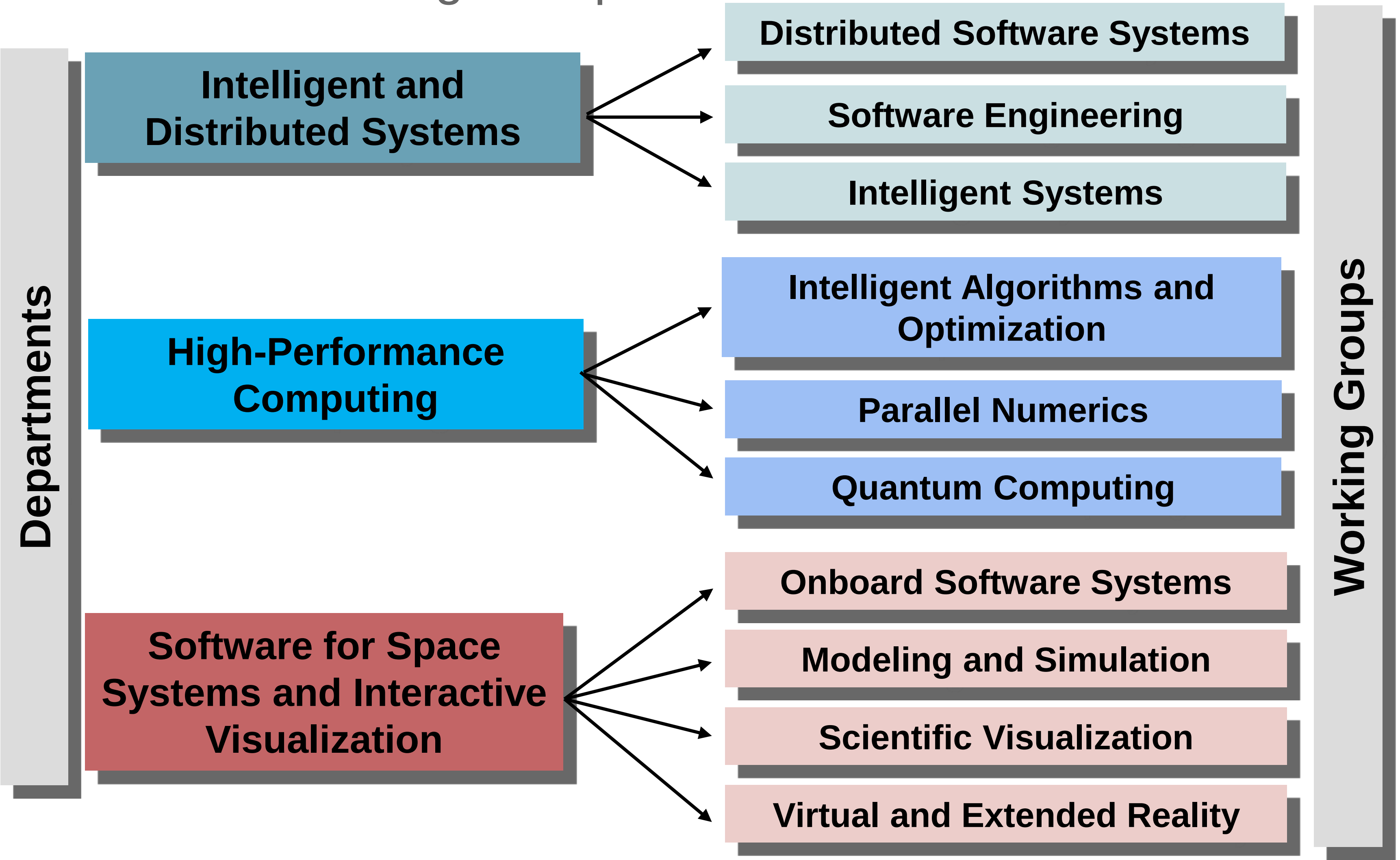
Institute for Software Technology

- stands for **innovative software engineering**,
- develops **challenging individual software solutions** for DLR, and
- is partner in **scientific projects** in the area of simulation and software technology.

DLR

DLR Institute for Software Technology

Scientific Themes and Working Groups



HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

High Performance Computing
Teams

Department
High Performance Computing
Head: Dr. Achim Basermann
Deputy: Dr. Margrit Klitz

- **Intelligent Algorithms & Optimiziation**
  Dr. Martin Siggel
- **Quantum Computing**
  Dr. Tobias Stollenwerk
- **Parallel Numerics**
  Dr. Jonas Thies

Go to www.menti.com and use the code 96 37 79 (SciML: 88 94 91)

- **Use your mobile**
- **Use your laptop**
- **Use your tablet**
- **Use your desktop**

for input to *mentimeter* questions

Survey

- **What characterizes a Data Scientist?**
- **Data Science Basics**
- **The Data Science Process**
- **AI and ML**
- **Common Applications**
- **Scientific Machine Learning**

Source: https://bit.ly/30dekJB

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

What characterizes a Data Scientist?

DLR

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

What characterizes a Data Scientist?

A data scientist is a person who has the knowledge and skills to conduct sophisticated and systematic analyses of data. A data scientist extracts insights from data sets for product development, and evaluates and identifies strategic opportunities.

- “There’s a joke running around on Twitter that the definition of a data scientist is ‘a data analyst who lives in California’,” — Malcolm Chisholm

- “Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others,” — Mike Loukides

- “A data scientist is a rare hybrid, a computer scientist with the programming abilities to build software to scrape, combine, and manage data from a variety of sources and a statistician who knows how to derive insights from the information within. S/he combines the skills to create new prototypes with the creativity and thoroughness to ask and answer the deepest questions about the data and what secrets it holds,” — Jake Porway

What characterizes a Data Scientist? (cont.)

- Data scientists are "analytically-minded, statistically and mathematically sophisticated data engineers who can infer insights into business and other complex systems out of large quantities of data," — Steve Hillion
- "A data scientist is someone who blends math, algorithms, and an understanding of human behavior with the ability to hack systems together to get answers to interesting human questions from data," — Hilary Mason
- Data scientist is a "change agent." "A data scientist is part digital trendspotter and part storyteller stitching various pieces of information together." — Anjul Bhambhri
- "The definition of "data scientist" could be broadened to cover almost everyone who works with data in an organization. At the most basic level, you are a data scientist if you have the analytical skills and the tools to 'get' data, manipulate it and make decisions with it." — Pat Hanrahan
- "By definition all scientists are data scientists. In my opinion, they are half hacker, half analyst, they use data to build products and find insights. It's Columbus meet Columbo – starry eyed explorers and skeptical detectives." — Monica Rogati

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

What characterizes a Data Scientist?

- “A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.” — Daniel Tunkelang

- An ideal data scientist is “someone who has both the engineering skills to acquire and manage large data sets, and also has the statistician’s skills to extract value from the large data sets and present that data to a large audience.” — John Rauser

- Data scientist is “someone who can bridge the raw data and the analysis – and make it accessible. It’s a democratising role; by bringing the data to the people, you make the world just a little bit better,” — Simon Rogers

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

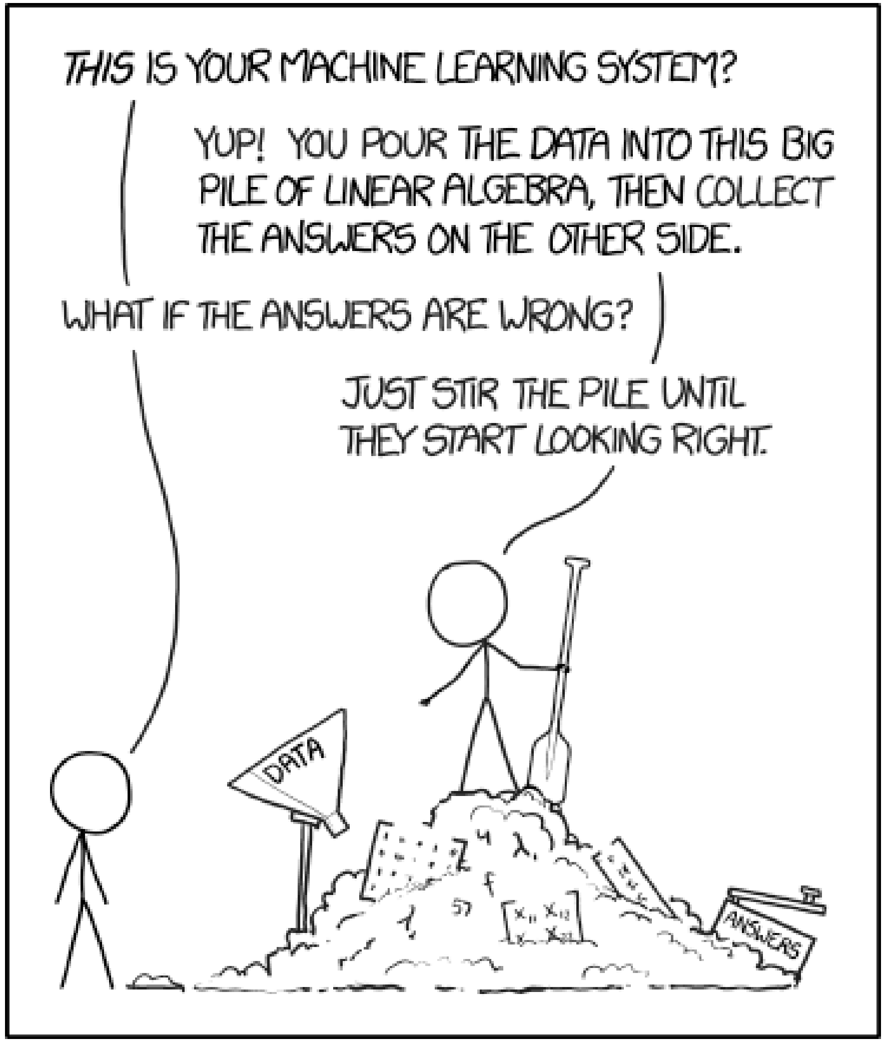
What characterizes a Data Scientist? (cont.)

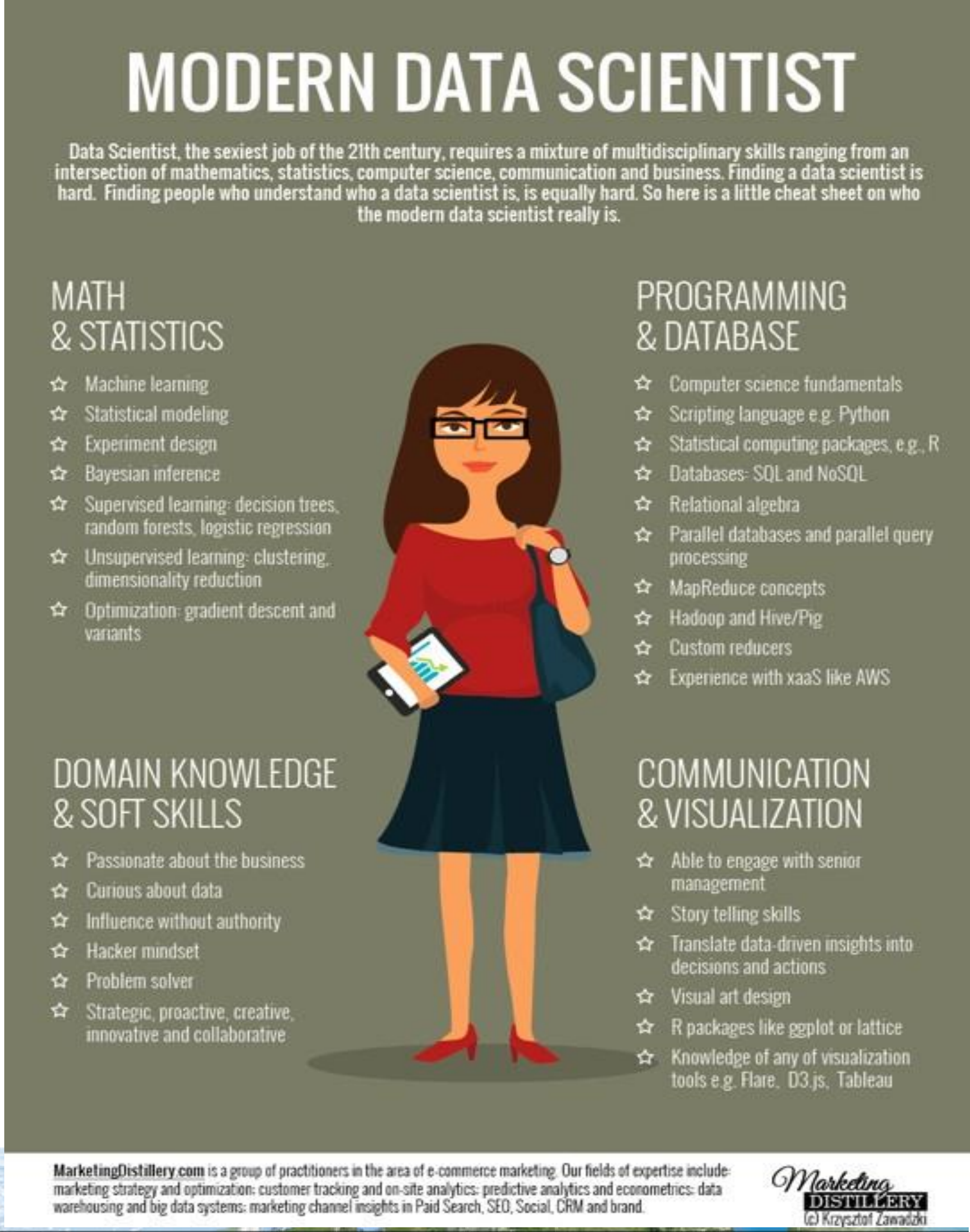
- “A data scientist is an engineer who employs the scientific method and applies data-discovery tools to find new insights in data. The scientific method — the formulation of a hypothesis, the testing, the careful design of experiments, the verification by others — is something they take from their knowledge of statistics and their training in scientific disciplines. The application (and tweaking) of tools comes from their engineering, or more specifically, computer science and programming background. The best data scientists are product and process innovators and sometimes, developers of new data-discovery tools,” — Gil Press

- “A data scientist represents an evolution from the business or data analyst role. The formal training is similar, with a solid foundation typically in computer science and applications, modeling, statistics, analytics and math. What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems, they will pick the right problems that have the most value to the organization,” — IBM researchers

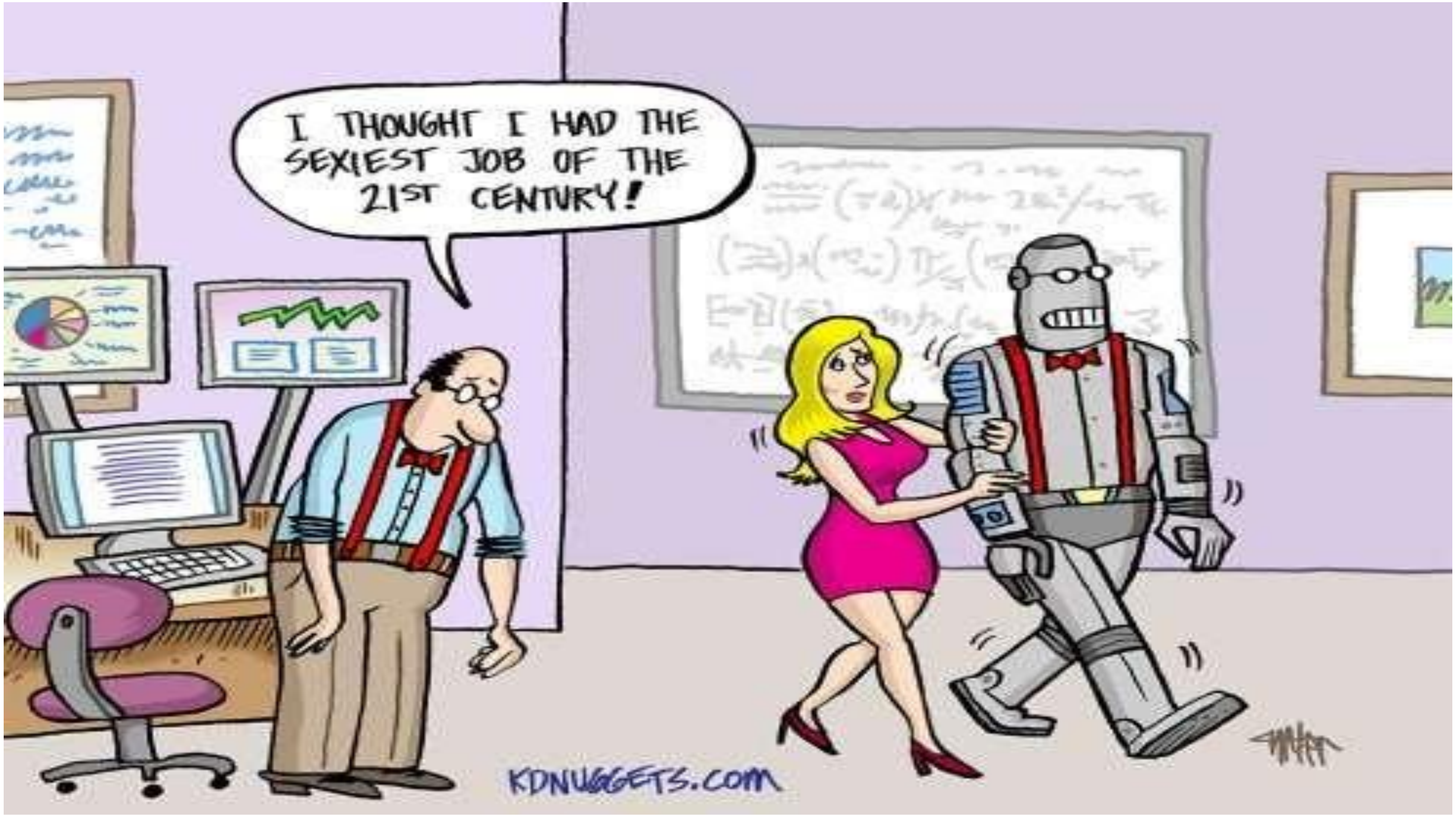
https://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

What characterizes a Data Scientist? (cont.)





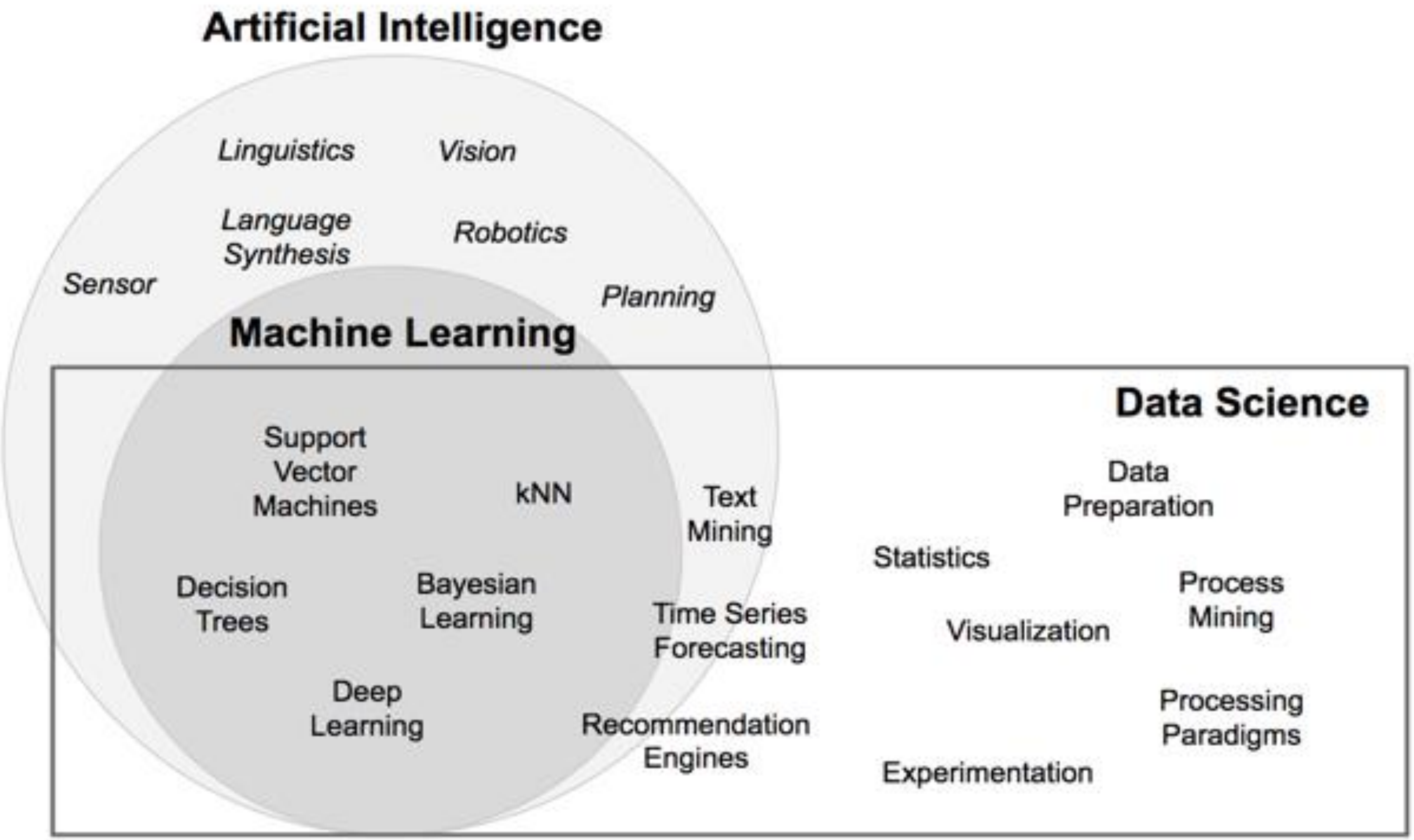
What characterizes a Data Scientist? (cont.)



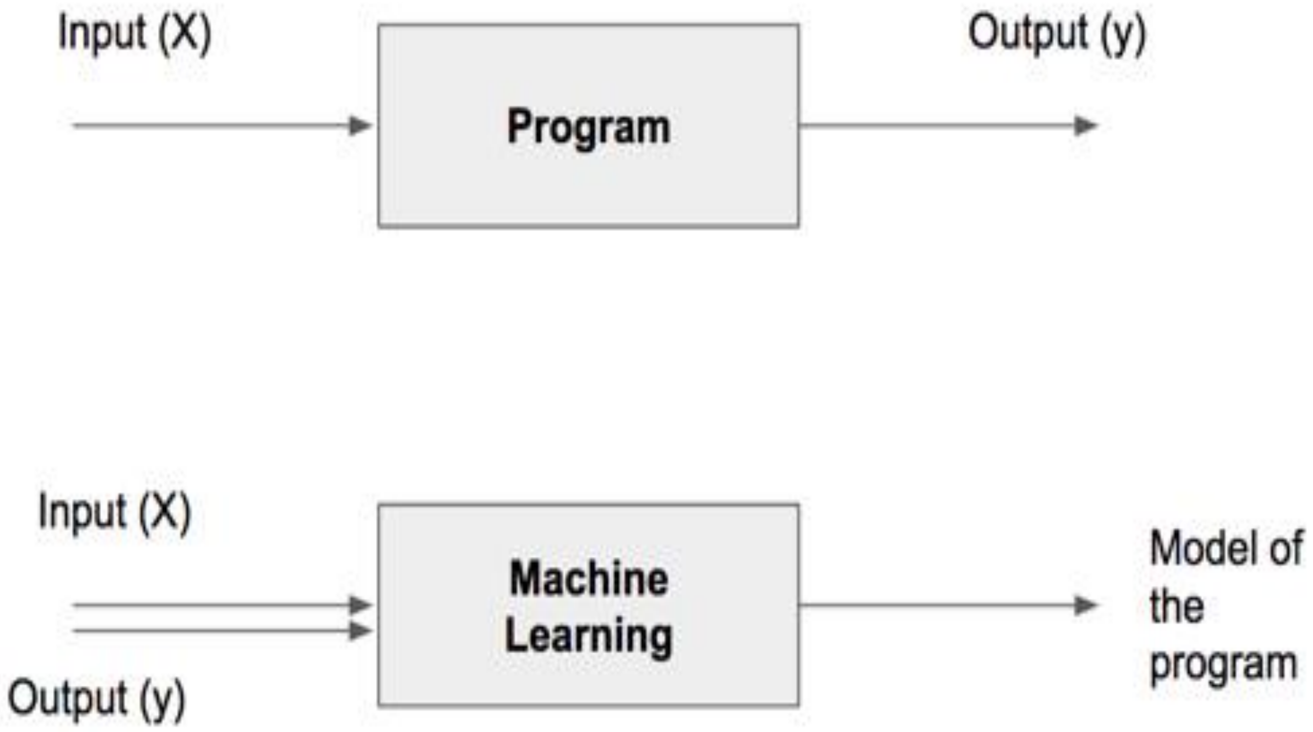
Data Science Basics

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

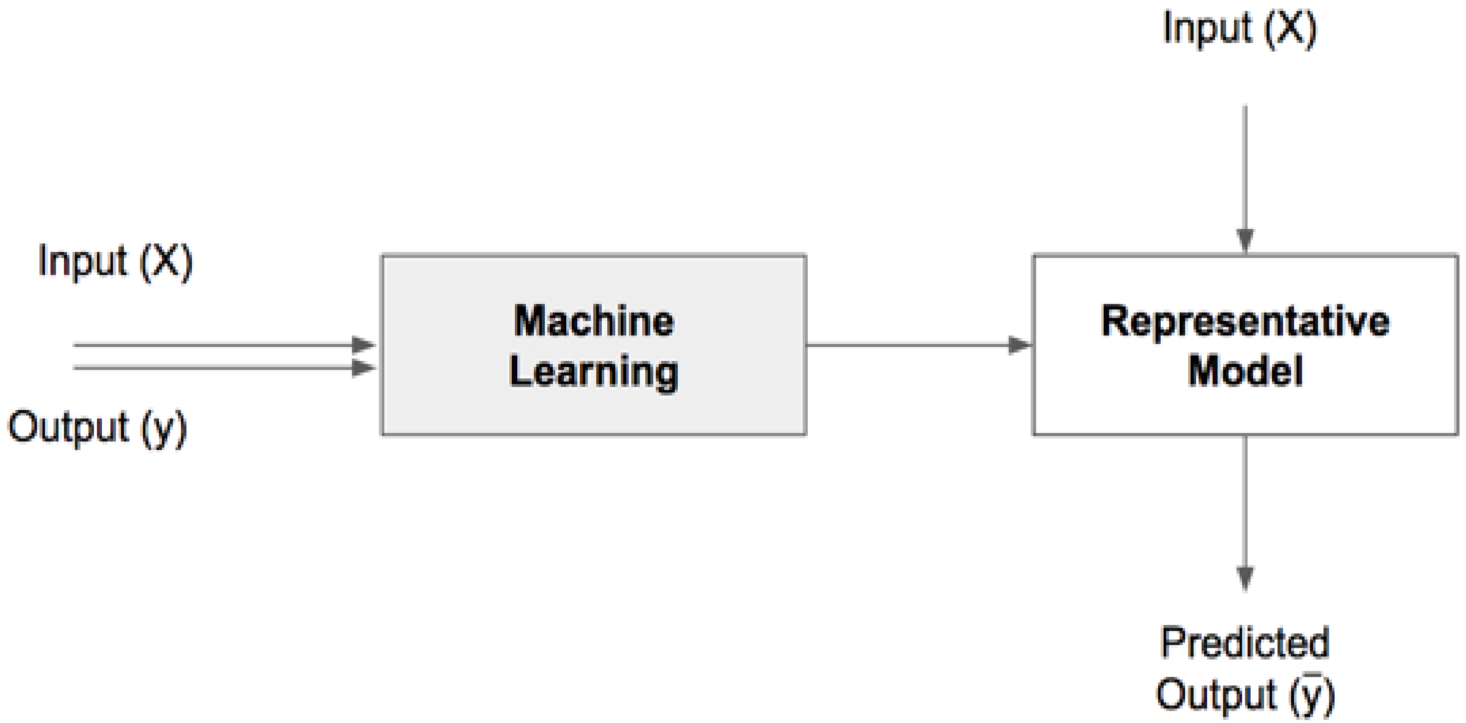
What is Data Science?



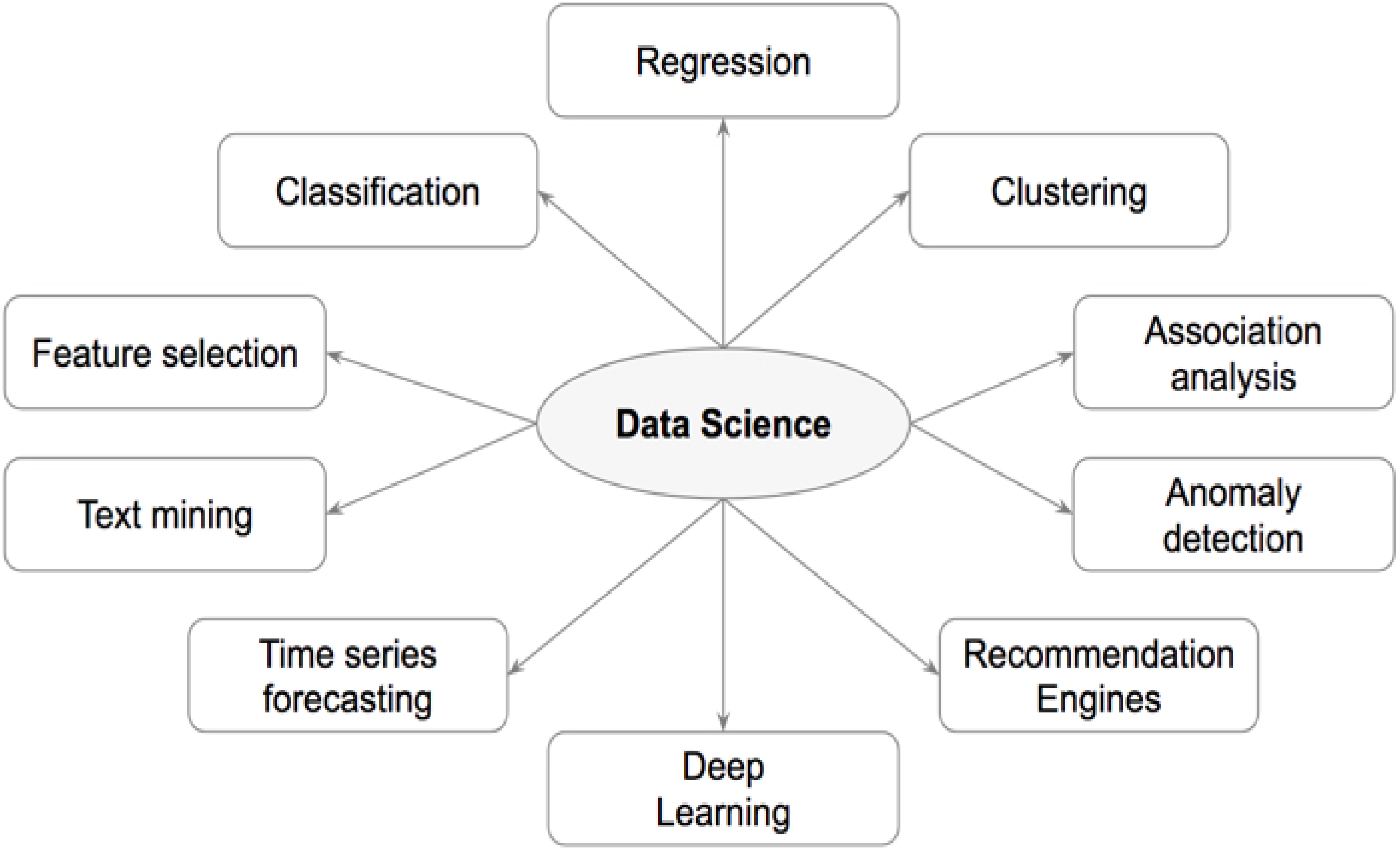
Models


Input (X)
Program
Output (y)
Input (X)
Output (y)
Machine Learning
Model of the program

Models (cont.)


Input (X)
Output (y)
Machine Learning
Representative Model
Input (X)
Predicted Output ($\bar{y}$)

Data Science Tasks



Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an item set based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailer based on transaction purchase history.

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

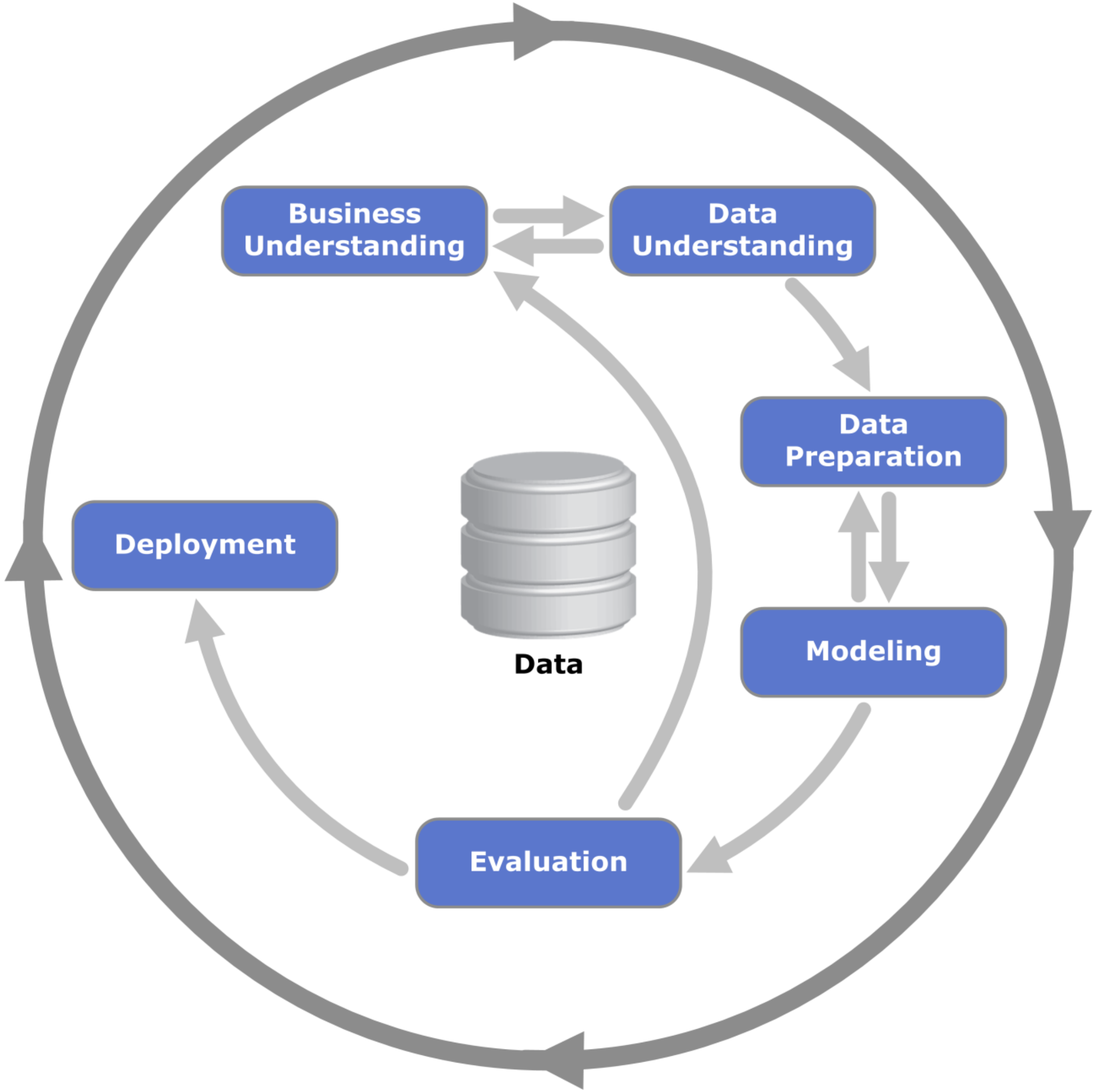
The Data Science Process

DLR

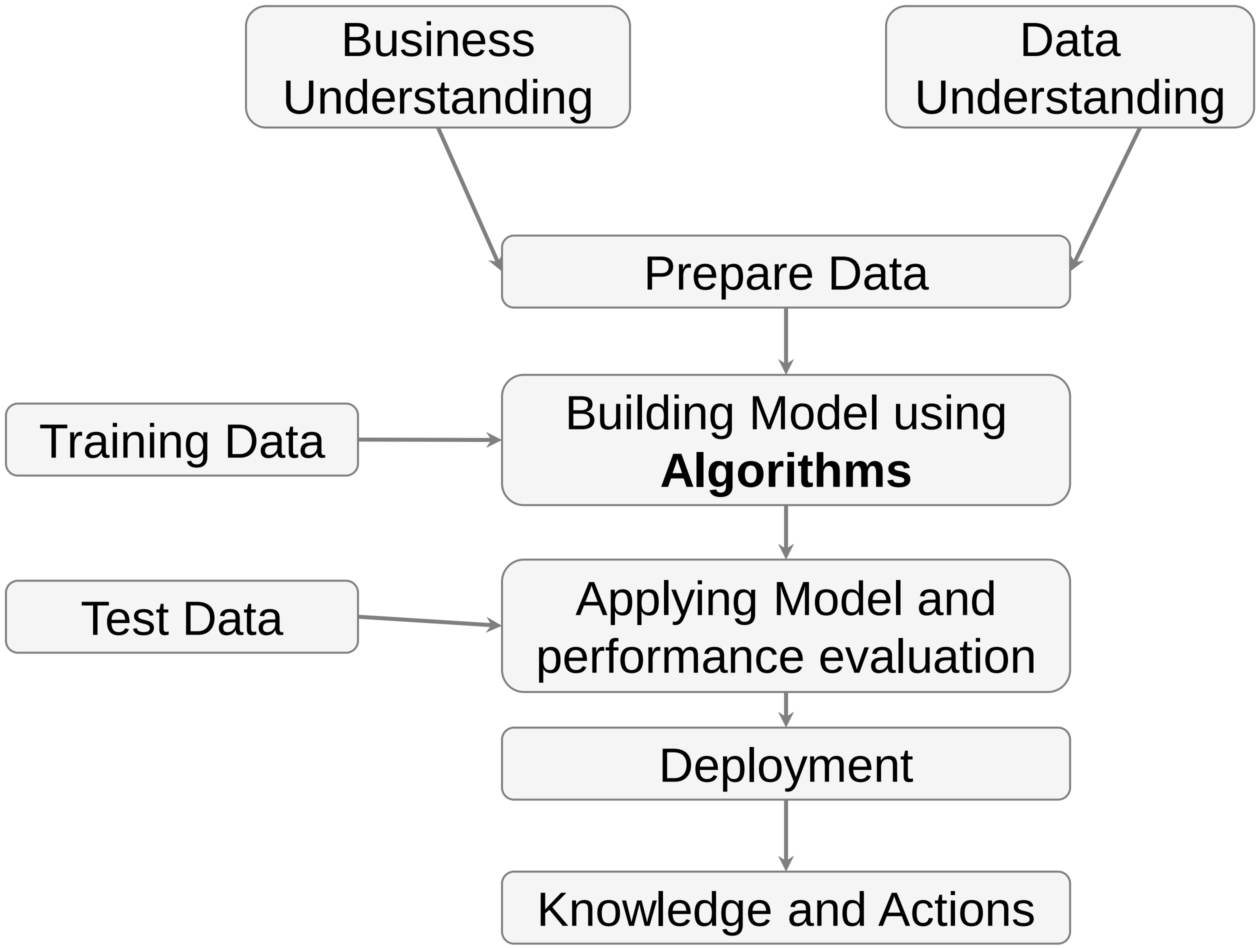
HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Data Science Process

- **CRISP-DM**



Process



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

1. Prior Knowledge

Gaining information on:

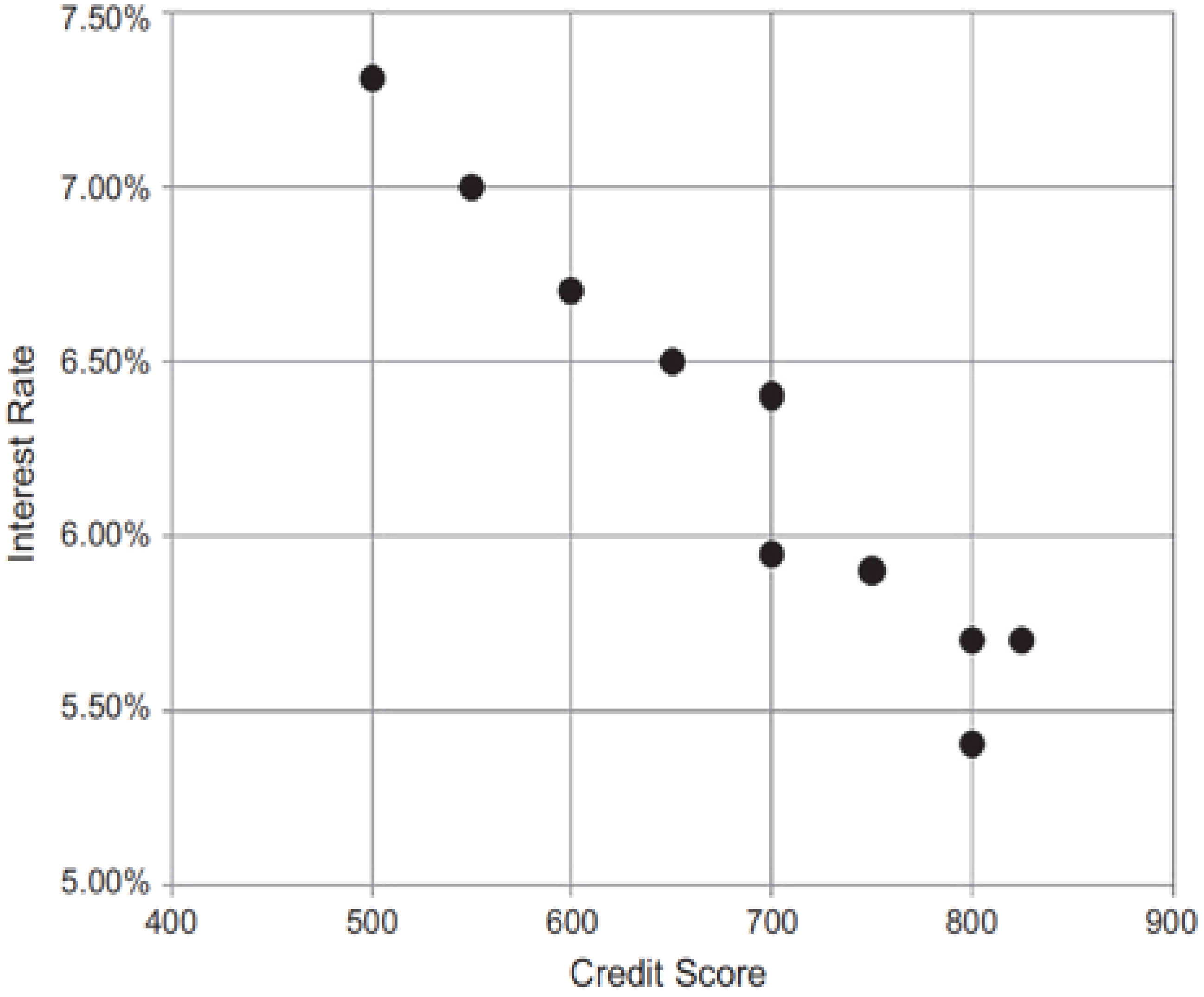
- Objective of the problem
- Subject area of the problem
- Data

Table 2.1 Data Set

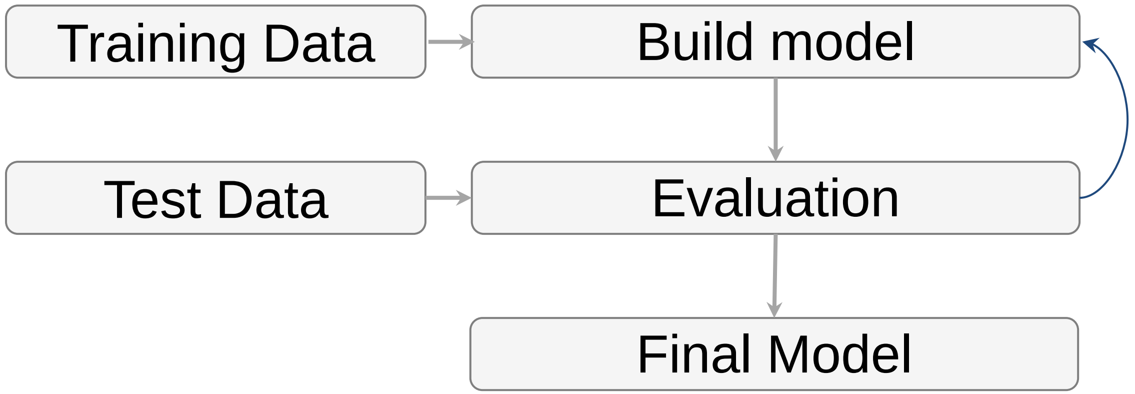
Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

2. Data Preparation

- Data Exploration
- Data quality
- Handling missing values
- Data type conversion
- Transformation
- Outliers
- Feature selection
- Sampling



3. Modeling


Training Data
Build model
Test Data
Evaluation
Final Model

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

3. Modeling (cont.)

Spliting training and test data sets

Table 2.3 Training Data Set

Borrower	Credit Score (X)	Interest Rate (Y)
01	500	7.31%
02	600	6.70%
03	700	5.95%
05	800	5.40%
06	800	5.70%
08	550	7.00%
09	650	6.50%

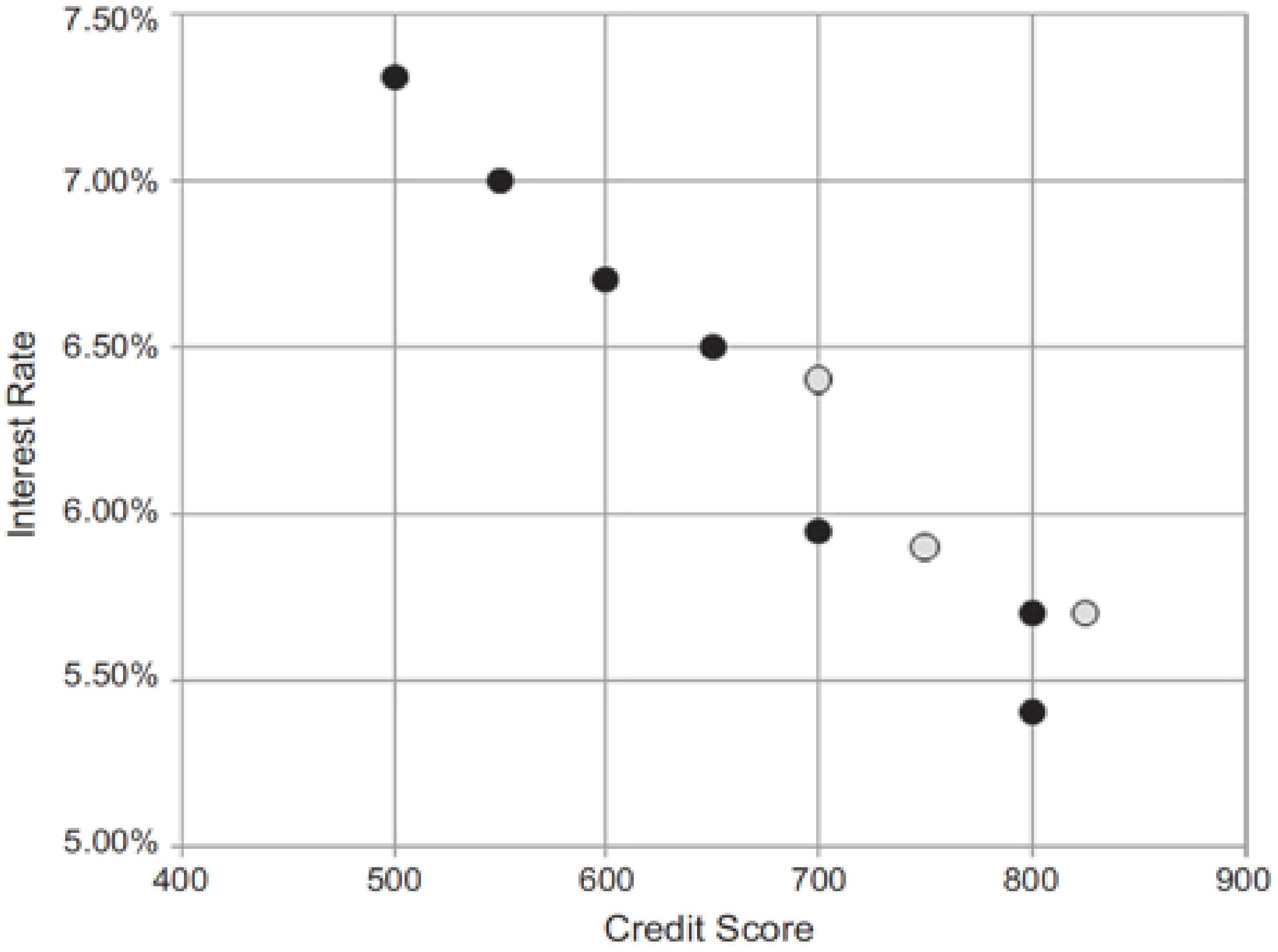
Table 2.4 Test Data Set

Borrower	Credit Score (X)	Interest Rate (Y)
04	700	6.40%
07	750	5.90%
10	825	5.70%

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

3. Modelling (cont.)

Spliting training and test data sets

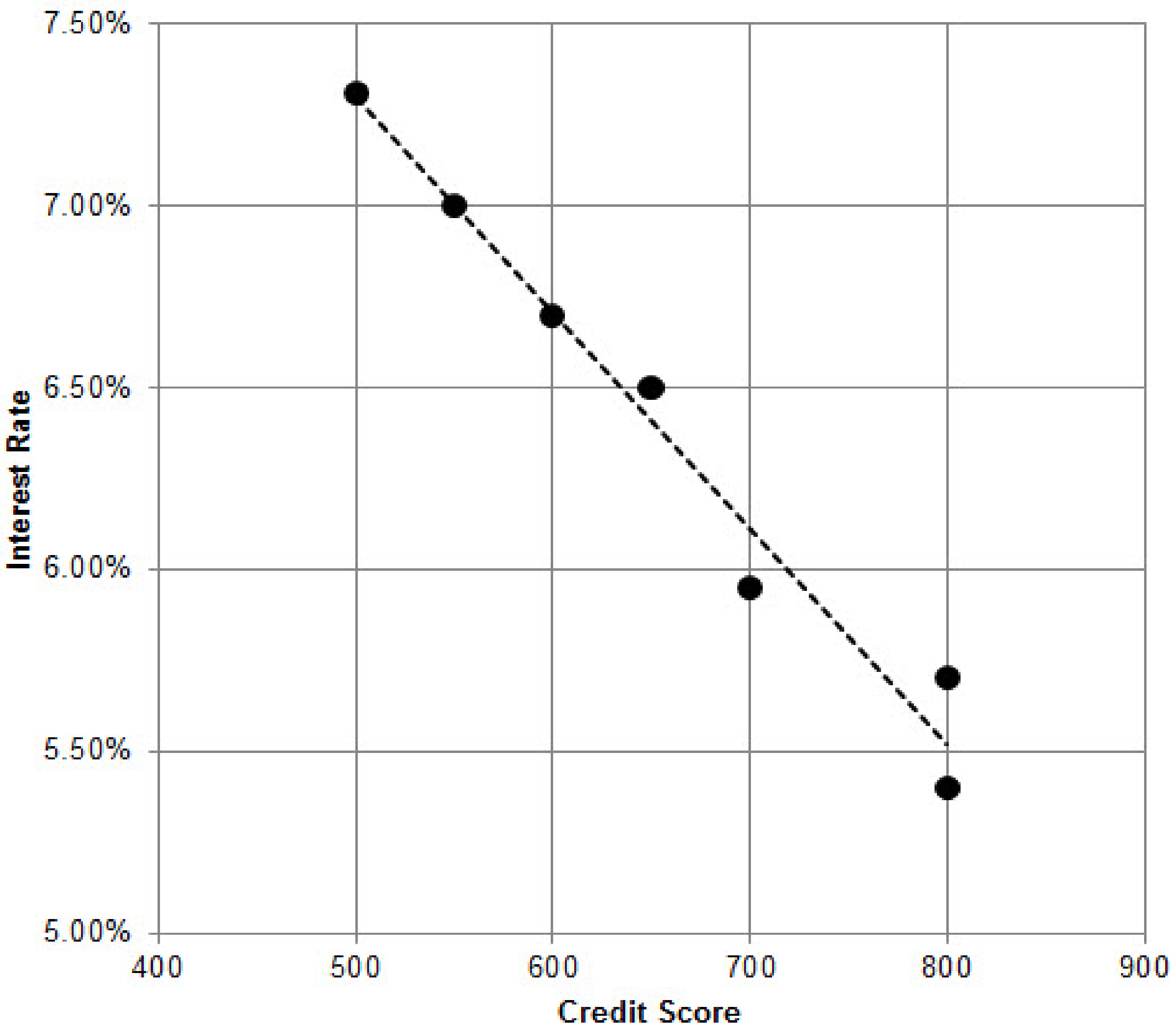


- Training Data
- Test Data

DLR

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

3. Modeling (cont.)



$$y = 0.1036 - \frac{6.1}{100{,}000}x$$

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

3. Modeling (cont.)

Evaluation of test dataset

Table 2.5 Evaluation of Test Data Set

Borrower	Credit Score (X)	Interest Rate (Y)	Model Predicted (Y)	Model Error
04	700	6.40%	6,09%	-0,31%
07	750	5.90%	5,79%	-0,11%
10	825	5.70%	5,33%	-0,37%

4. Application

Product readiness

Technical integration

Model response time

Remodeling

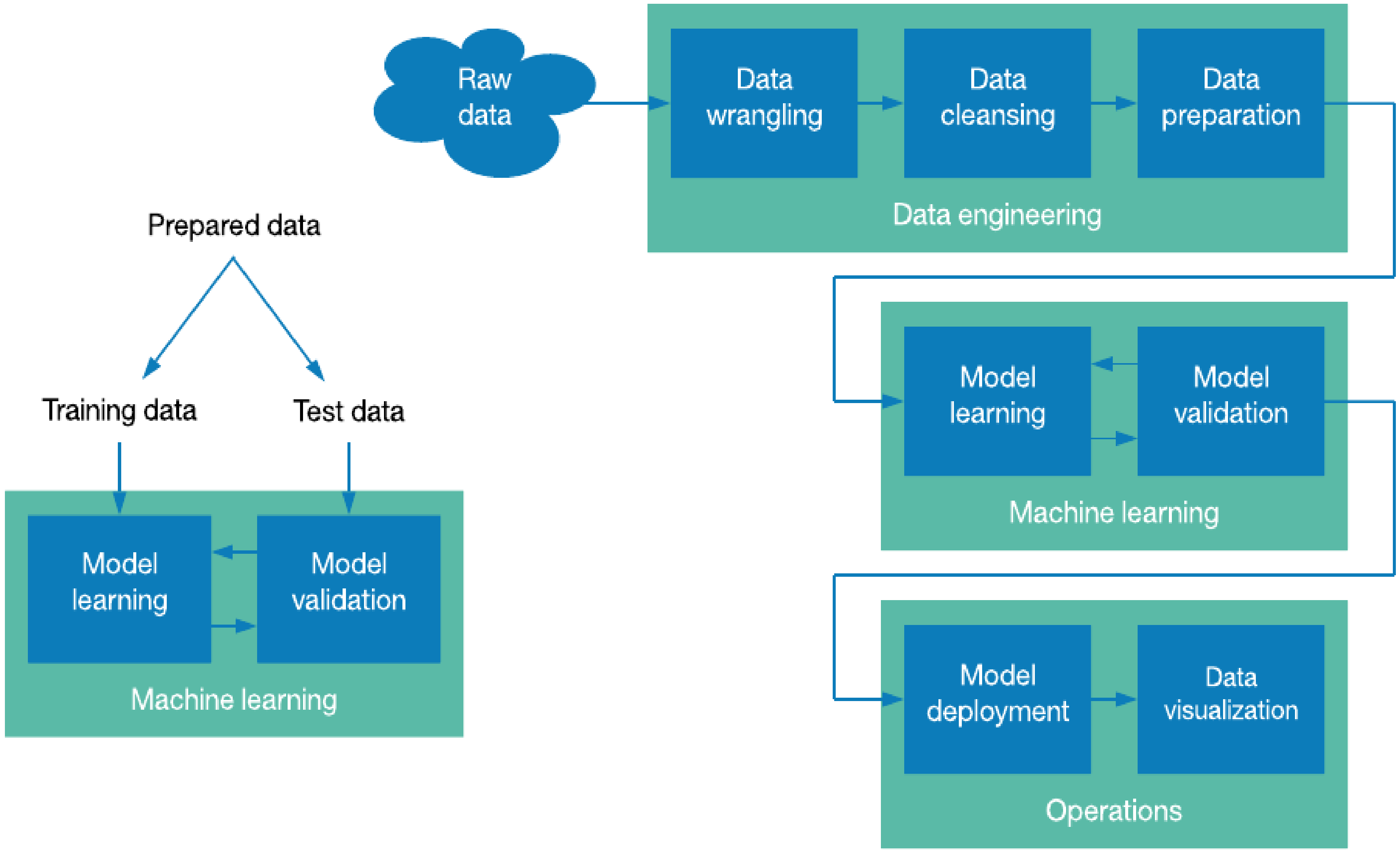
Assimilation

5. Knowledge

Posterior knowledge

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

Data Science Pipeline

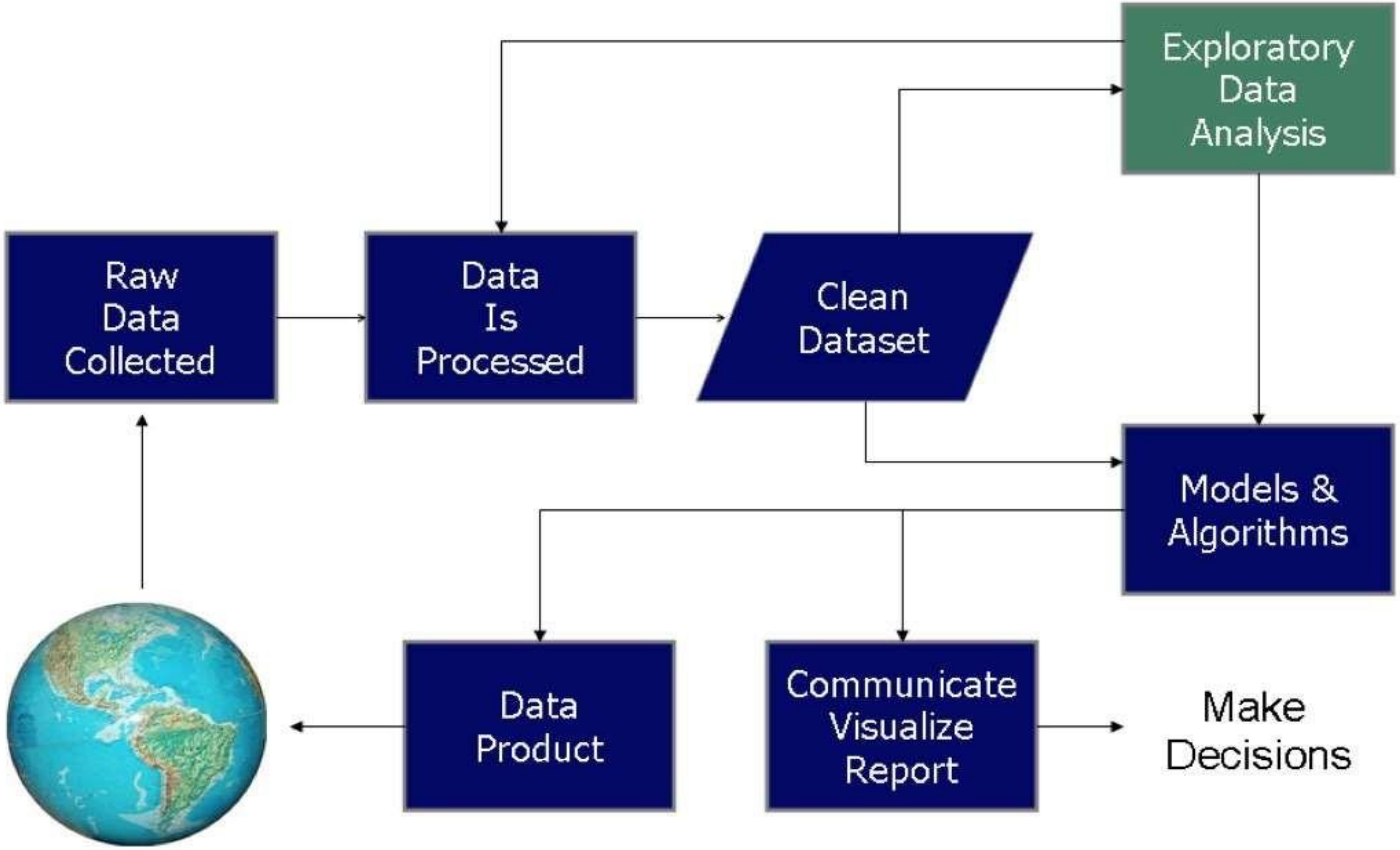


HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Models of Data

Structured data	Semi-structured data	Unstructured data
Databases	XML / JSON data	Audio
	Email	Video
	Web pages	Image data
		Natural language
		Documents

Data Science Workflow



AI and ML

Artificial Intelligence (AI)

- Artificial Intelligence is the science of getting machines to think and make decisions like human beings do.
- Theory and development of computer systems

Importance of AI

- AI automates Repetitive Learning and discovery through data.
- AI adds intelligence to existing products.
- AI adapts through progressive learning algorithms to let the data do the programming.
- AI analyzes more and deeper data using neural networks that have many hidden layers.
- AI achieves incredible accuracy through deep neural networks.

Applications of AI

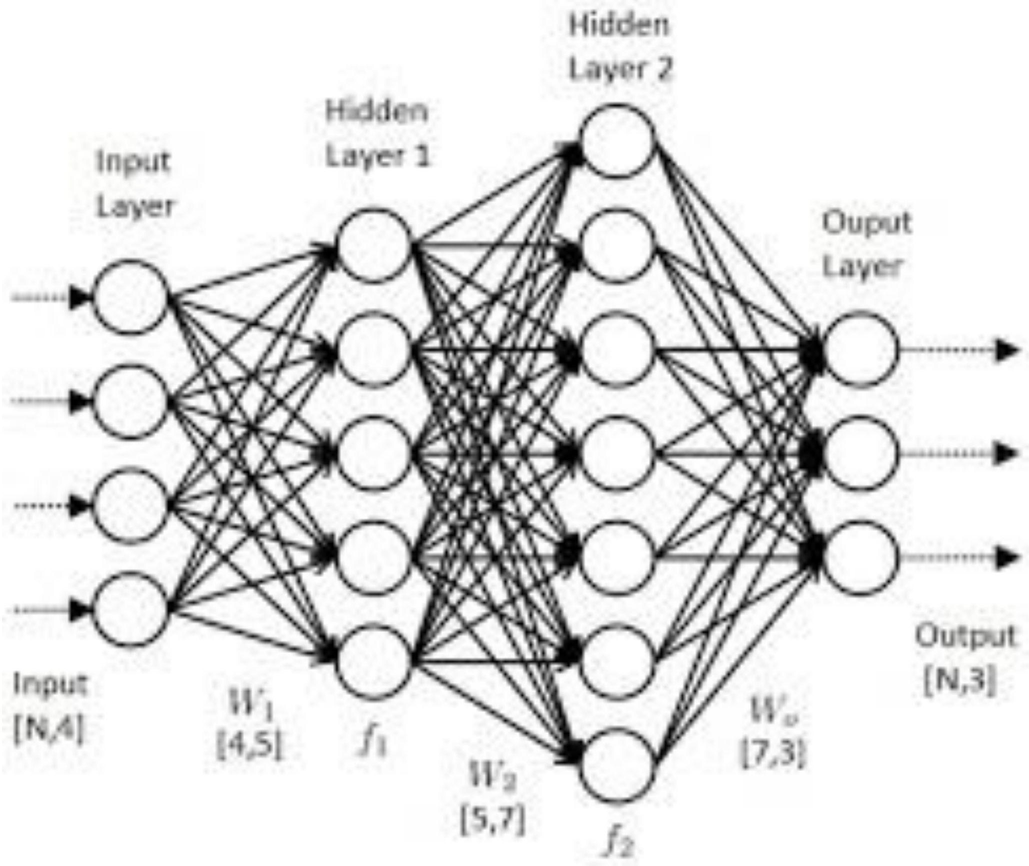
- **AI in Health Care**

Applications of AI (cont.)

- **AI in Agriculture**

Selected Domains of AI

- **Neural Networks**
  - Neural Networks are a class of models within the general machine learning literature.



Selected Domains of AI (cont.)

- **Robotics**
  - A branch of AI, which is composed of different branches and applications of robots

Selected Domains of AI (cont.)

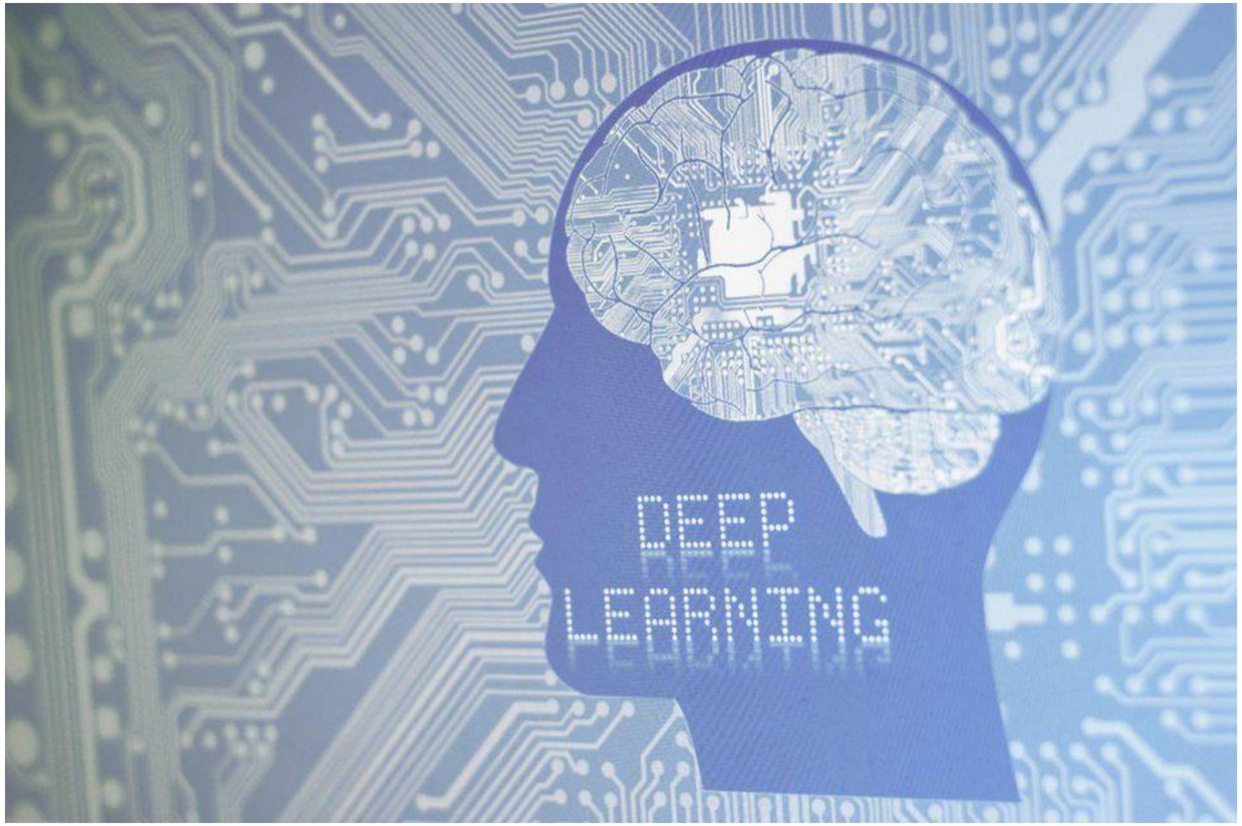
- **Expert System**
  - A computer system that emulates the decision-making ability of a human expert

Selected AI Technologies

- **Machine Learning (ML)**

Selected AI Technologies (cont.)

- **Deep Learning Platforms**



Selected AI Technologies (cont.)

- Virtual Agents / Virtual Assistants

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Classification of AI

- **Weak AI**
  - AI system that is designed and trained for a specific type of task
  - Also known as Narrow AI

Classification of AI (cont.)

- **Strong AI**
  - AI system with generalized human cognitive abilities so that when presented with an unfamiliar task, it has enough intelligence to find a solution
  - Also known as Artificial General Intelligence

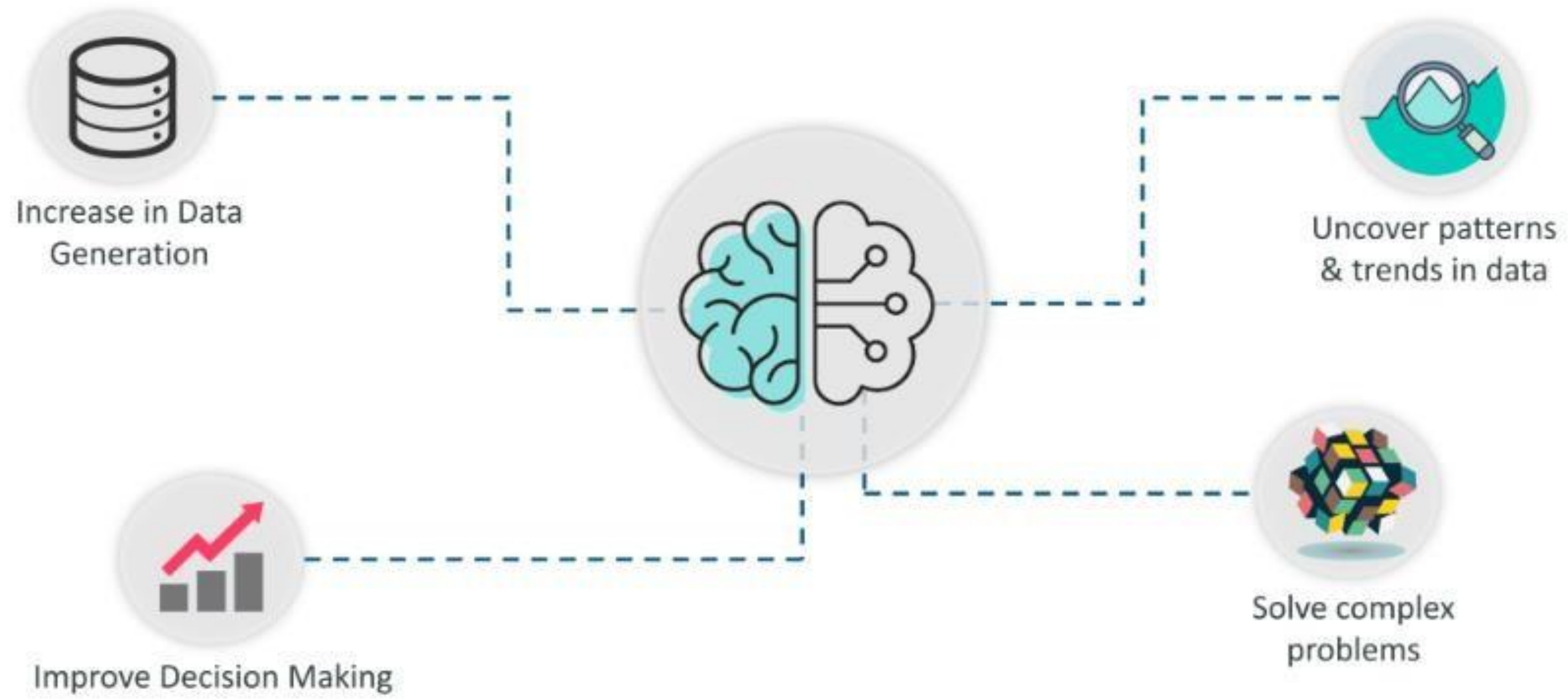
AI, ML and DL

- It covers anything which enables the computers to behave like humans.

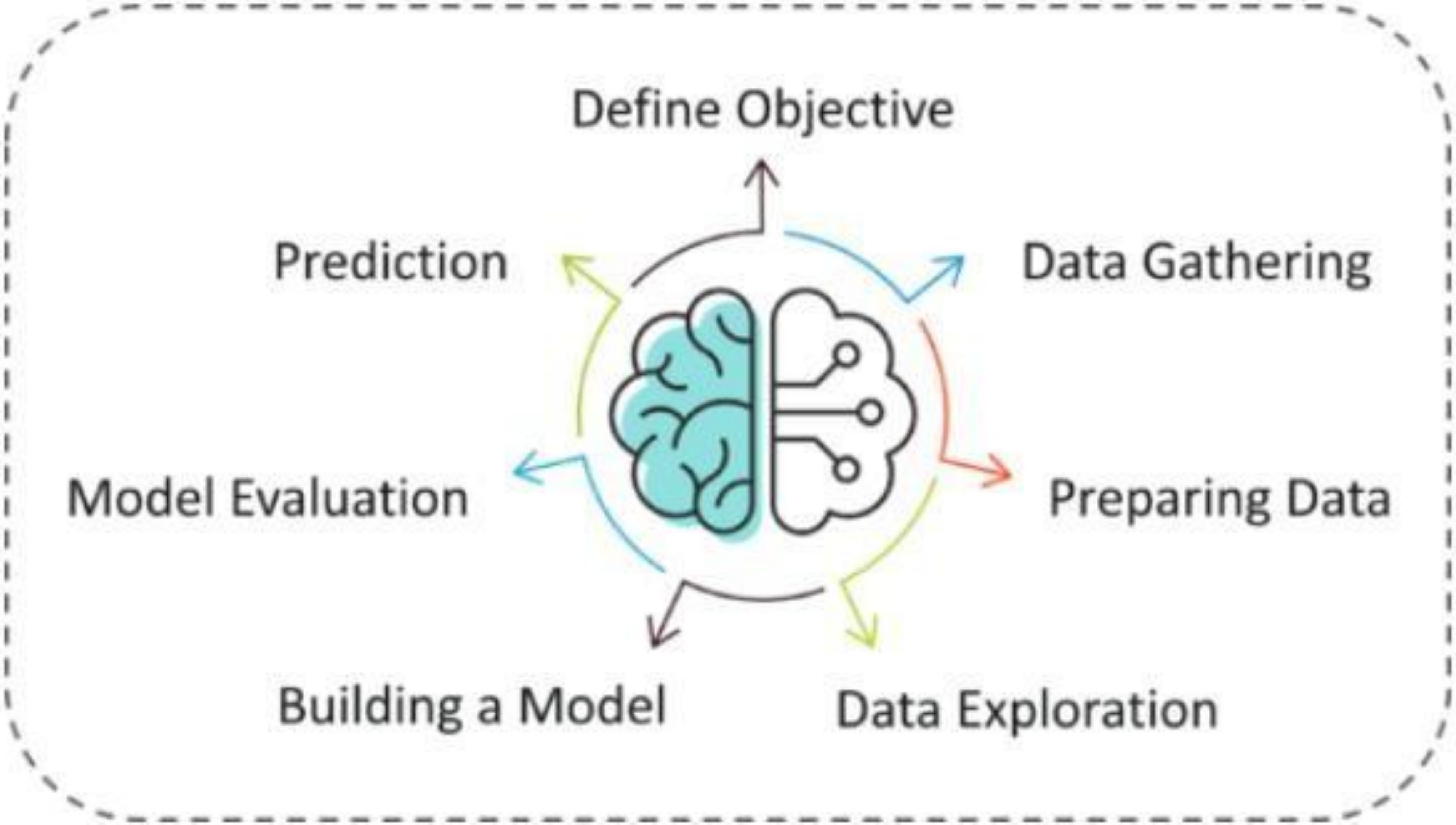
- It deals with the extraction of patterns from large data sets.

- It deals to train Deep Neural Networks so as to achieve better accuracy in those cases where former technologies were not performing up to the mark.

Need for Machine Learning (ML)



ML Process


Define Objective
Data Gathering
Preparing Data
Data Exploration
Building a Model
Model Evaluation
Prediction

ML Types

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

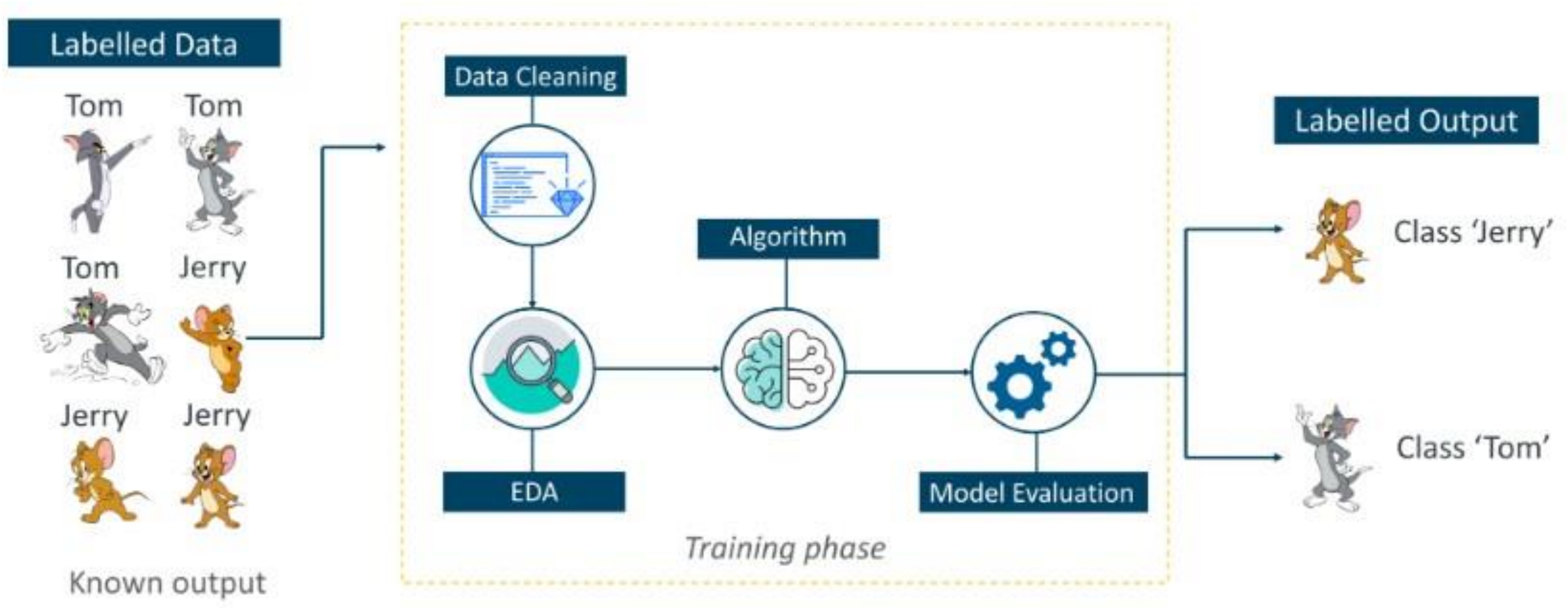
HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Understanding Supervised Learning

- A technique in which we teach or train the machine using data which is well labeled

DLR

Understanding Supervised Learning (cont.)

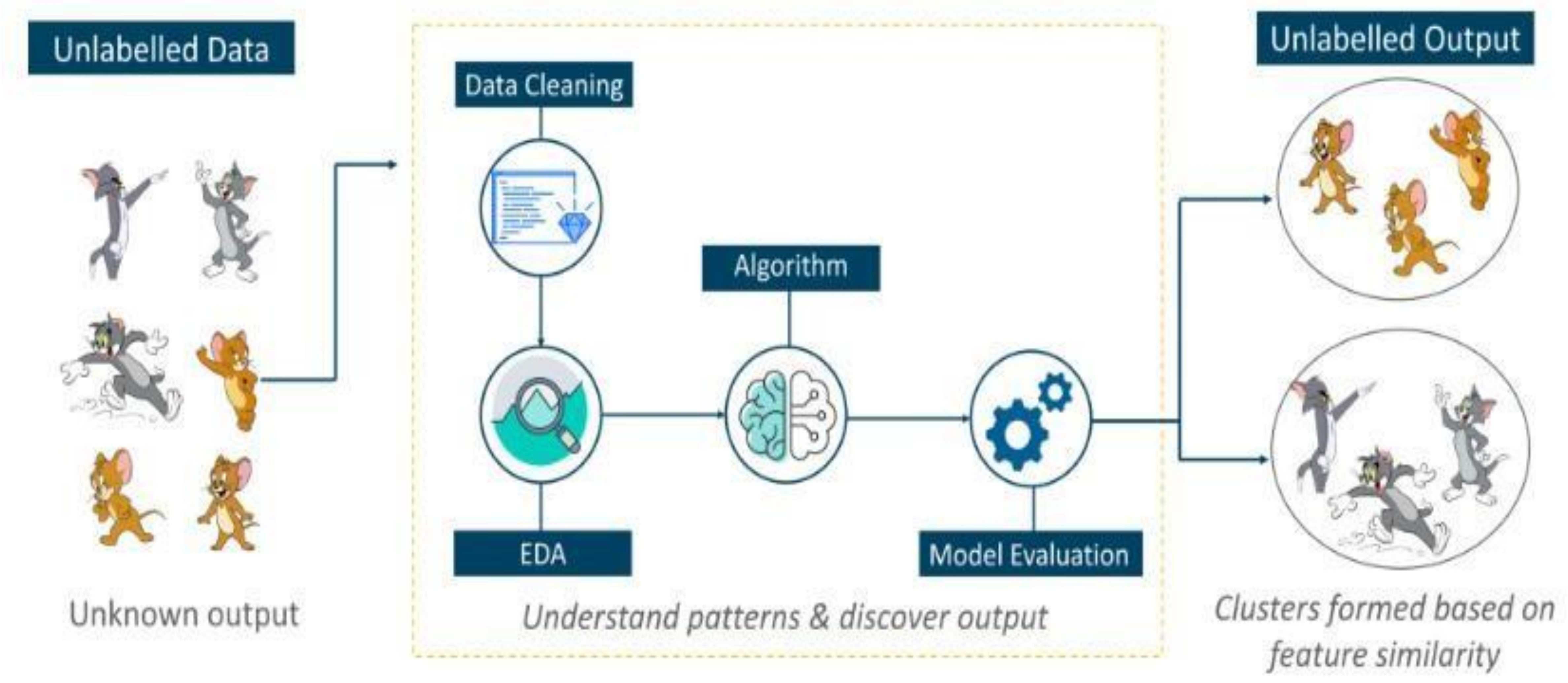


HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Understanding Unsupervised Learning

- Involves training by using unlabeled data and allowing the model to act on that information without guidance

Understanding Unsupervised Learning (cont.)

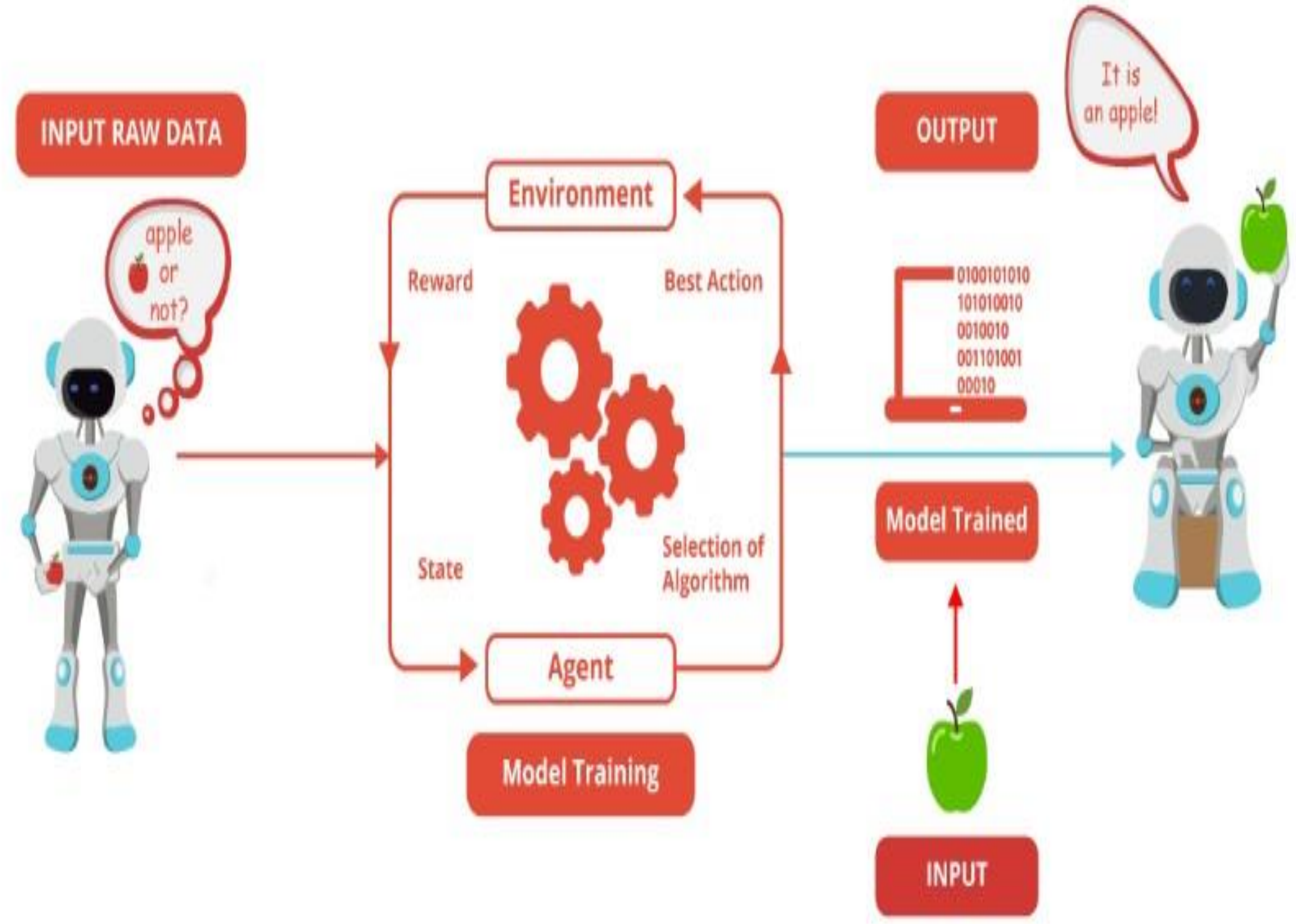


Understanding Reinforcement Learning

- A part of ML where an agent is put in an environment and he learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions

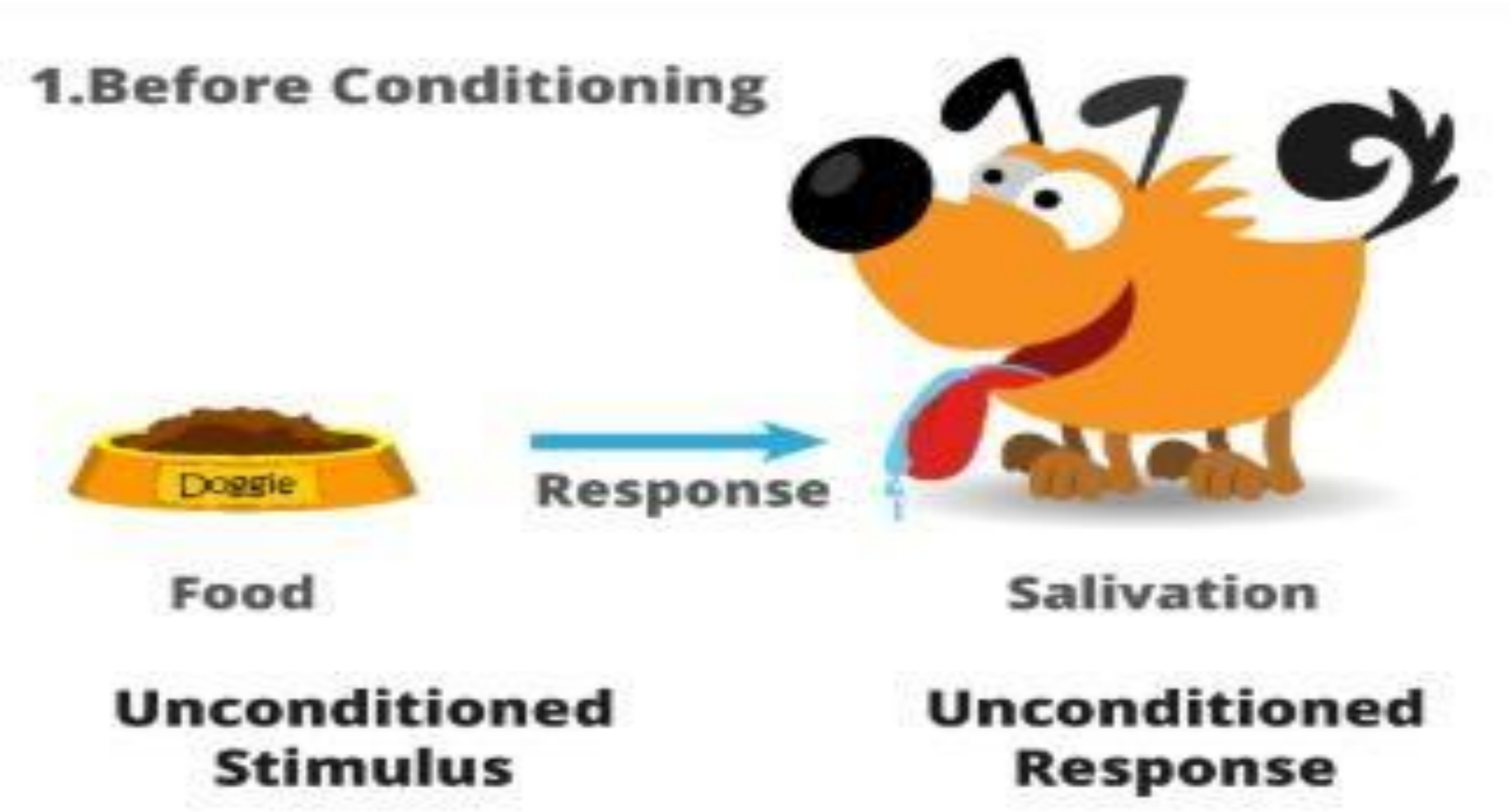
HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Understanding Reinforcement Learning (cont.)

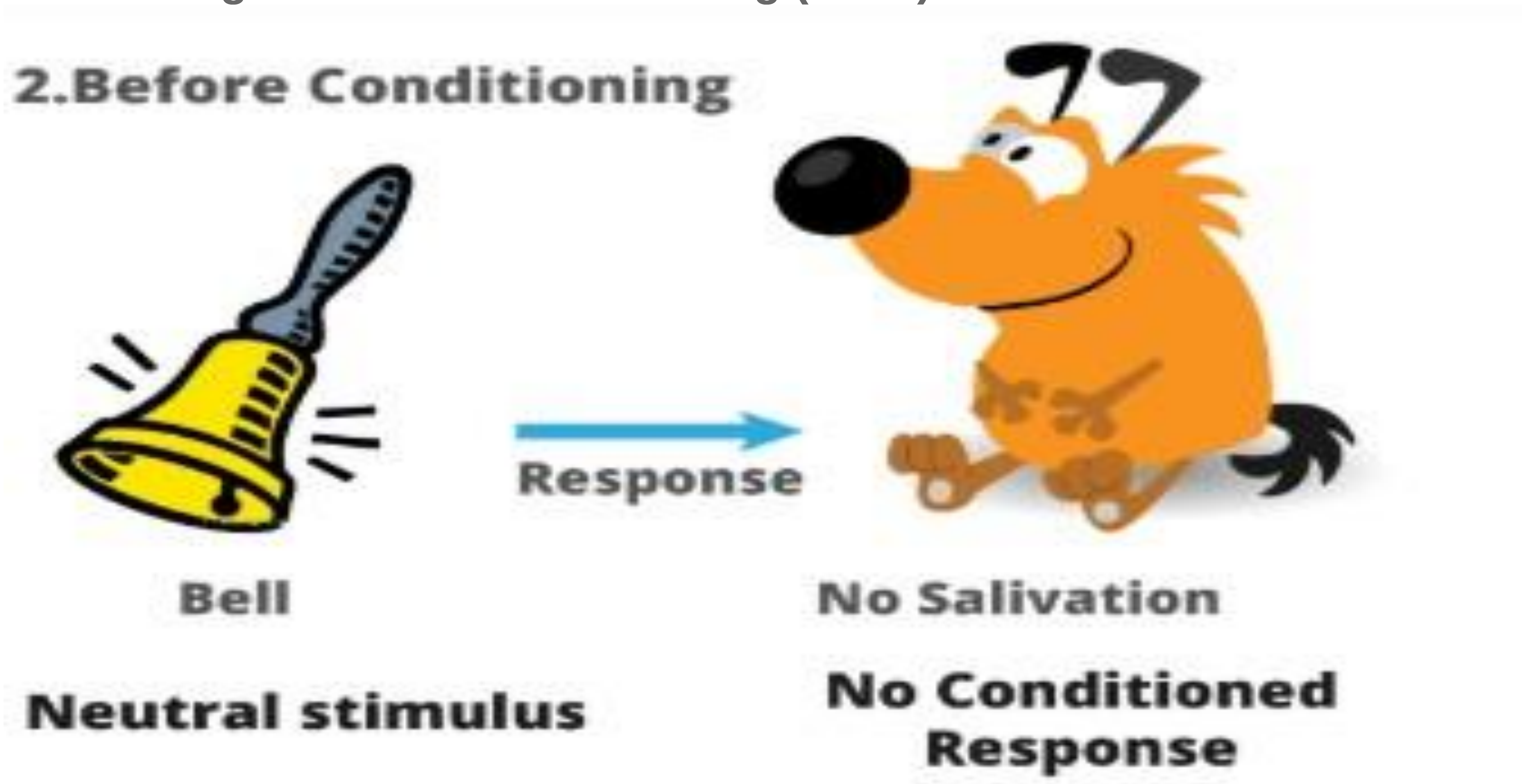


- Hit and trial method of learning
- All about the interaction between the environment and the learning agent
- Exploration and exploitation

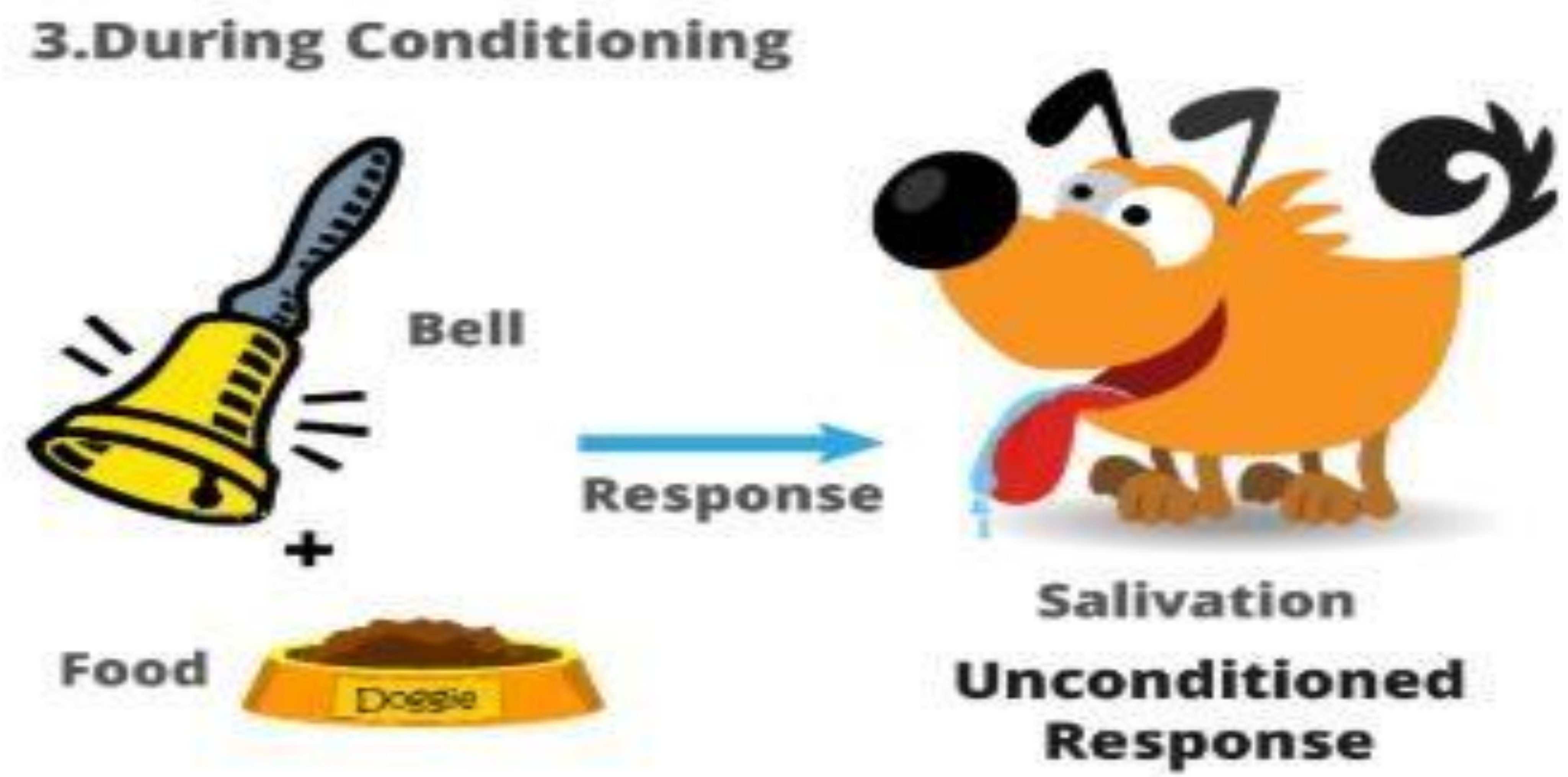
Understanding Reinforcement Learning (cont.)



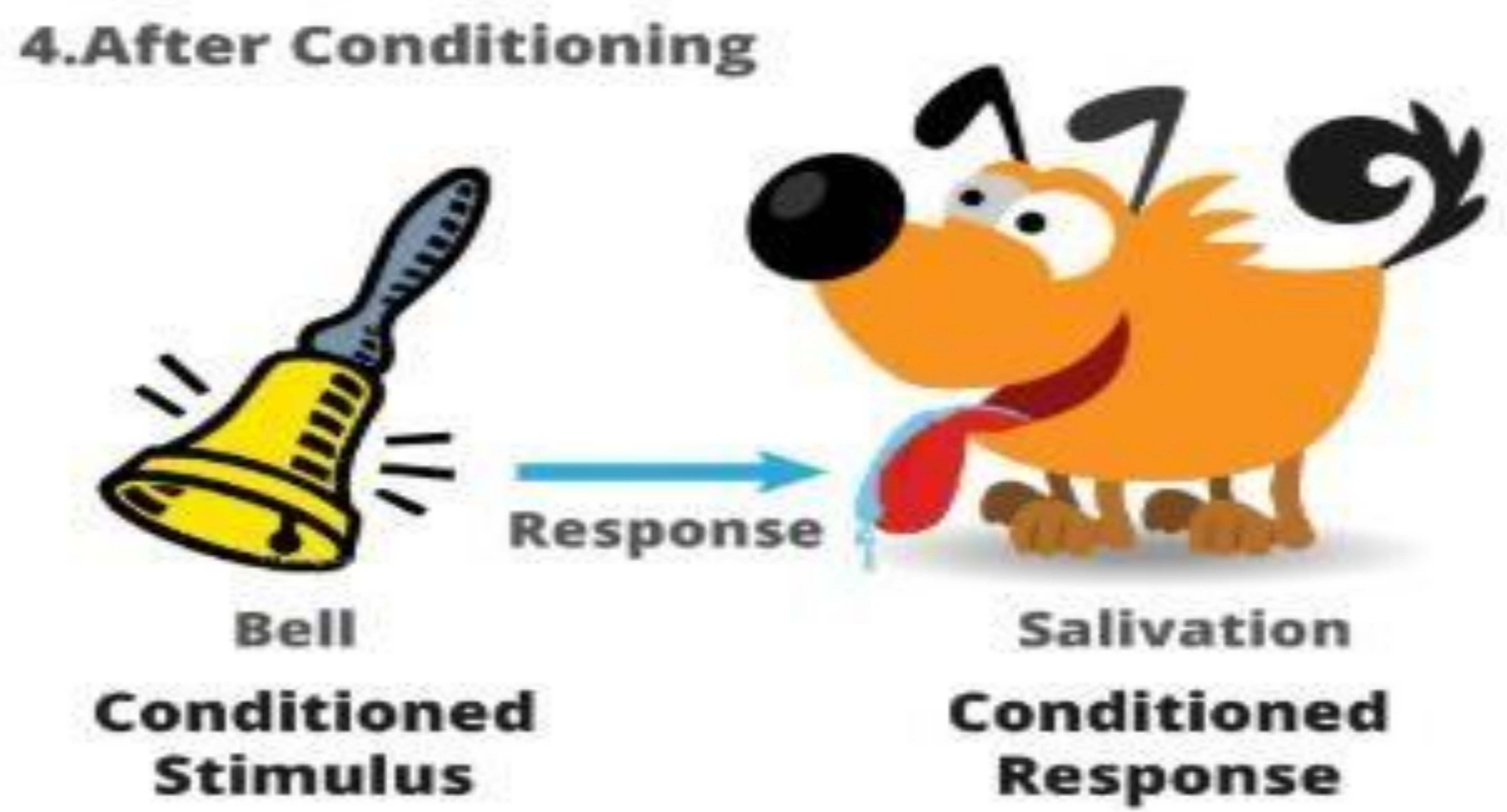
Understanding Reinforcement Learning (cont.)



Understanding Reinforcement Learning (cont.)

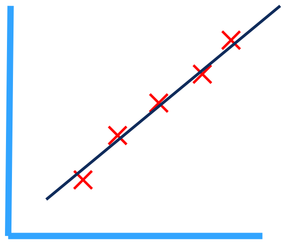


Understanding Reinforcement Learning (cont.)

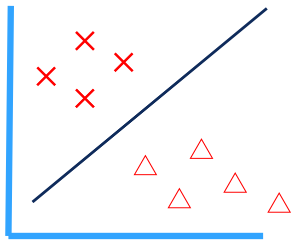


HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

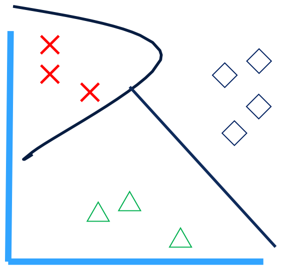
ML Tasks from a Statistical View



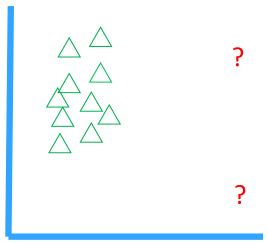
Regression – Looking for a statistical relationship across variables that may give us an estimate of a particular outcome.



Classification – Similar to regression but looking for separations in the data given predefined classes. (Supervised)



Clustering – Do not have predefined classes but trying to find groups or sets based upon data at hand. (Unsupervised)



Anomaly Detection – Identification of outliers based upon expected ranges of data.

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Common Applications

DLR

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science

- **Security**

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science (cont.)

- **Sports**

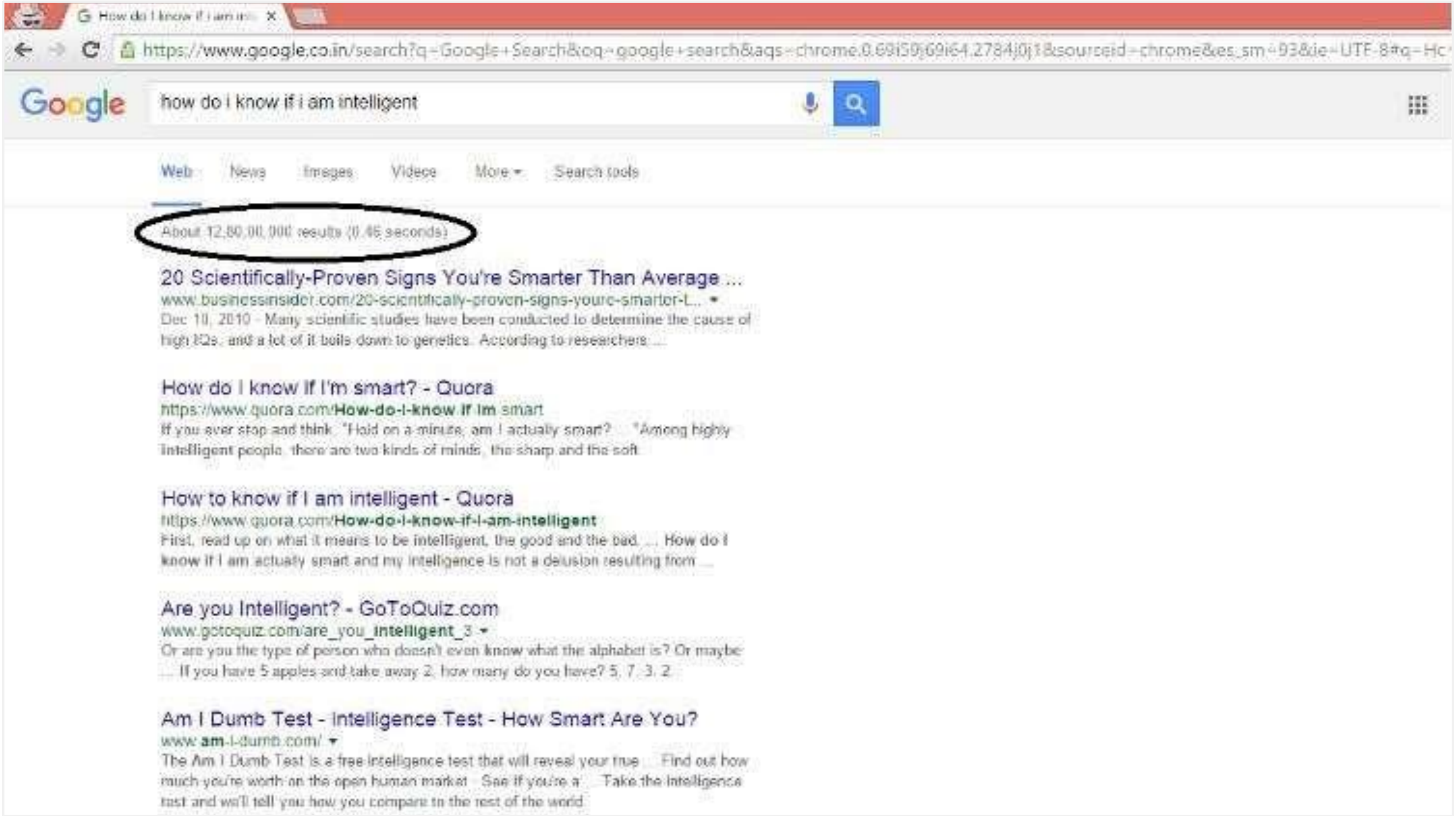
DLR

Applications of Data Science (cont.)

- Banking and Finance

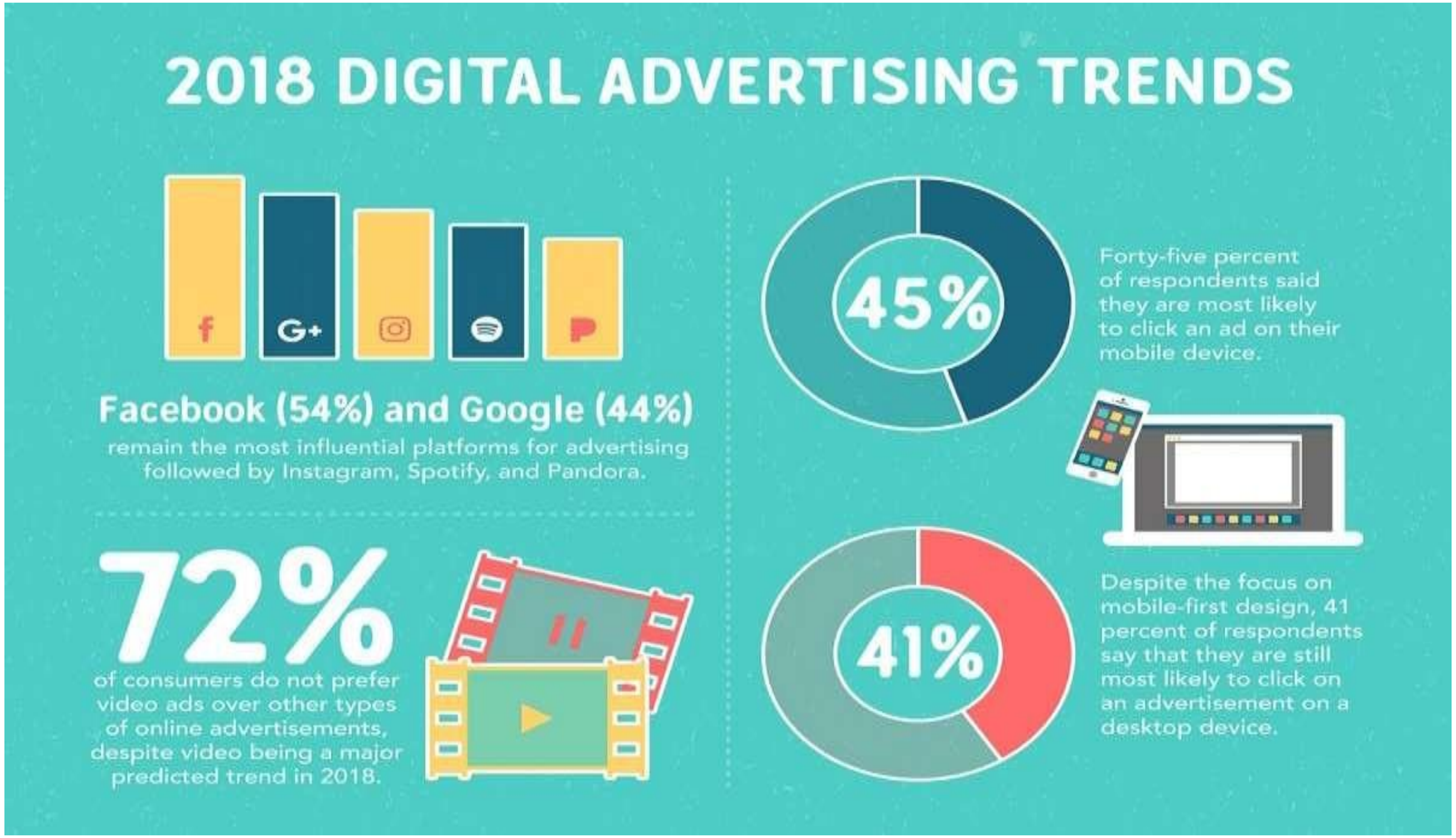
Applications of Data Science (cont.)

- Internet Search



Applications of Data Science (cont.)

- **Digital Advertisements**



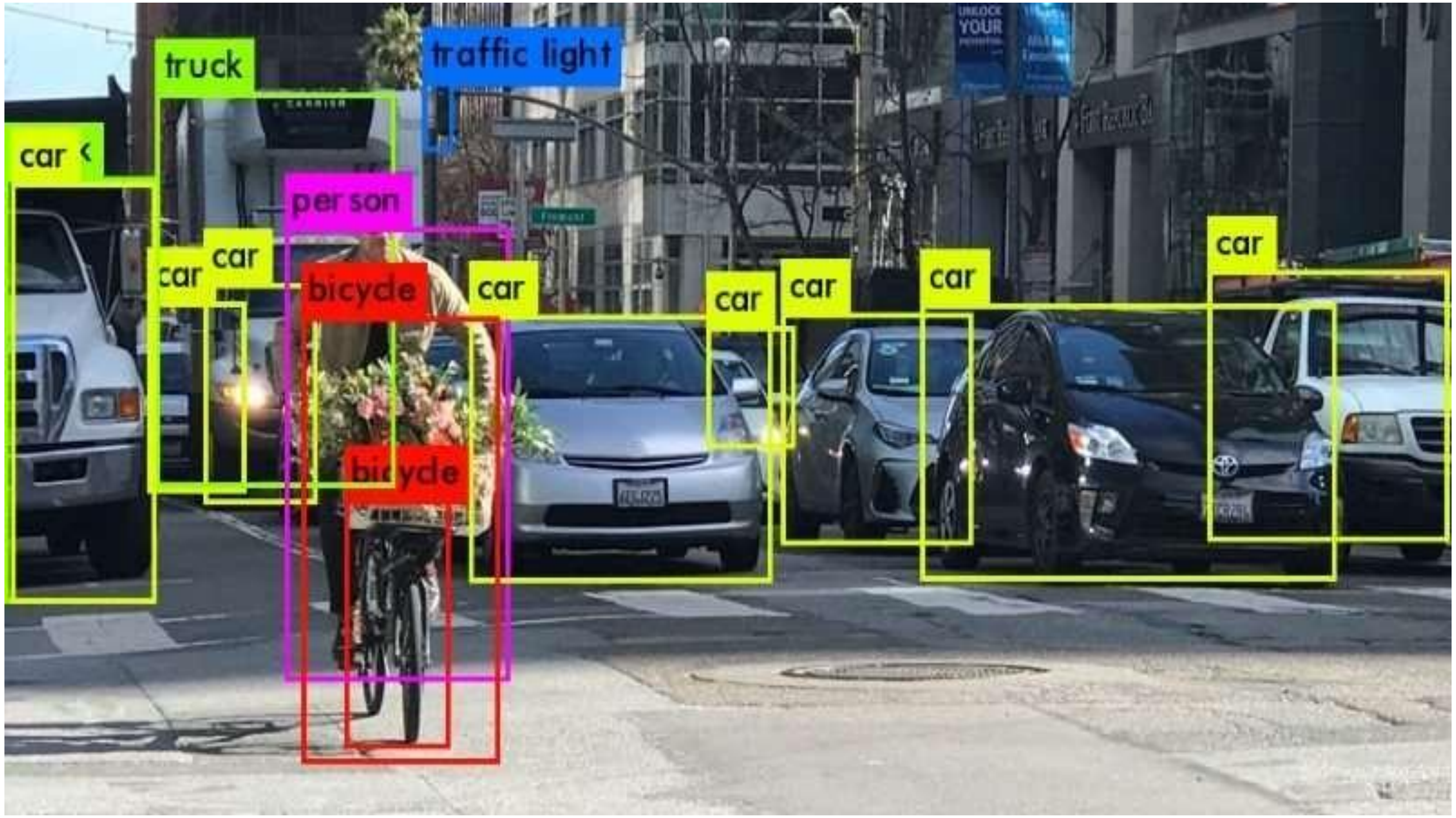
Applications of Data Science (cont.)

- **Recommender Systems**

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science (cont.)

- **Image Processing**



Applications of Data Science (cont.)

- **Speech Recognition**

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science (cont.)

- **Gaming**

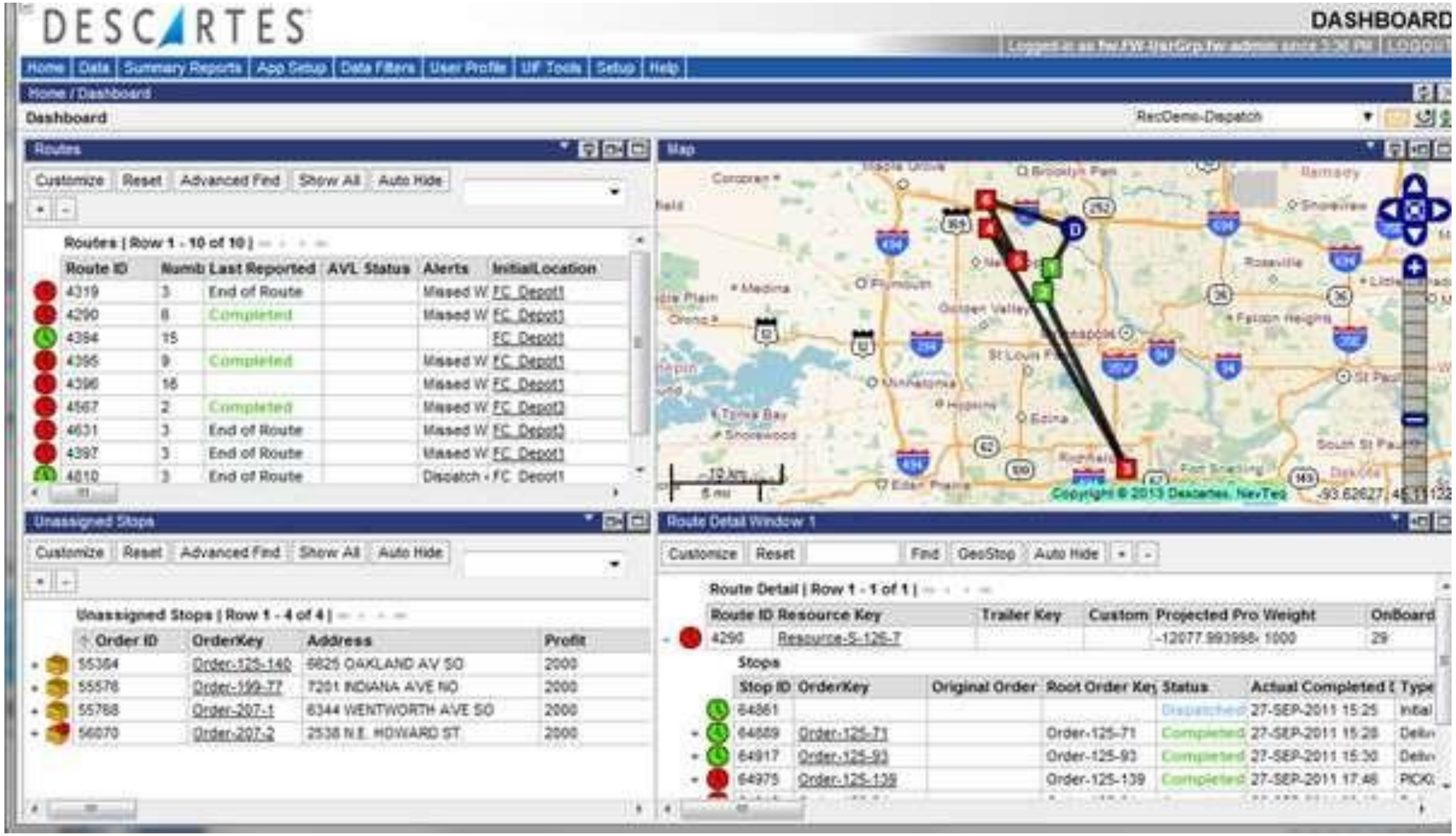
HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science (cont.)

- **Price Comparison Websites**

Applications of Data Science (cont.)

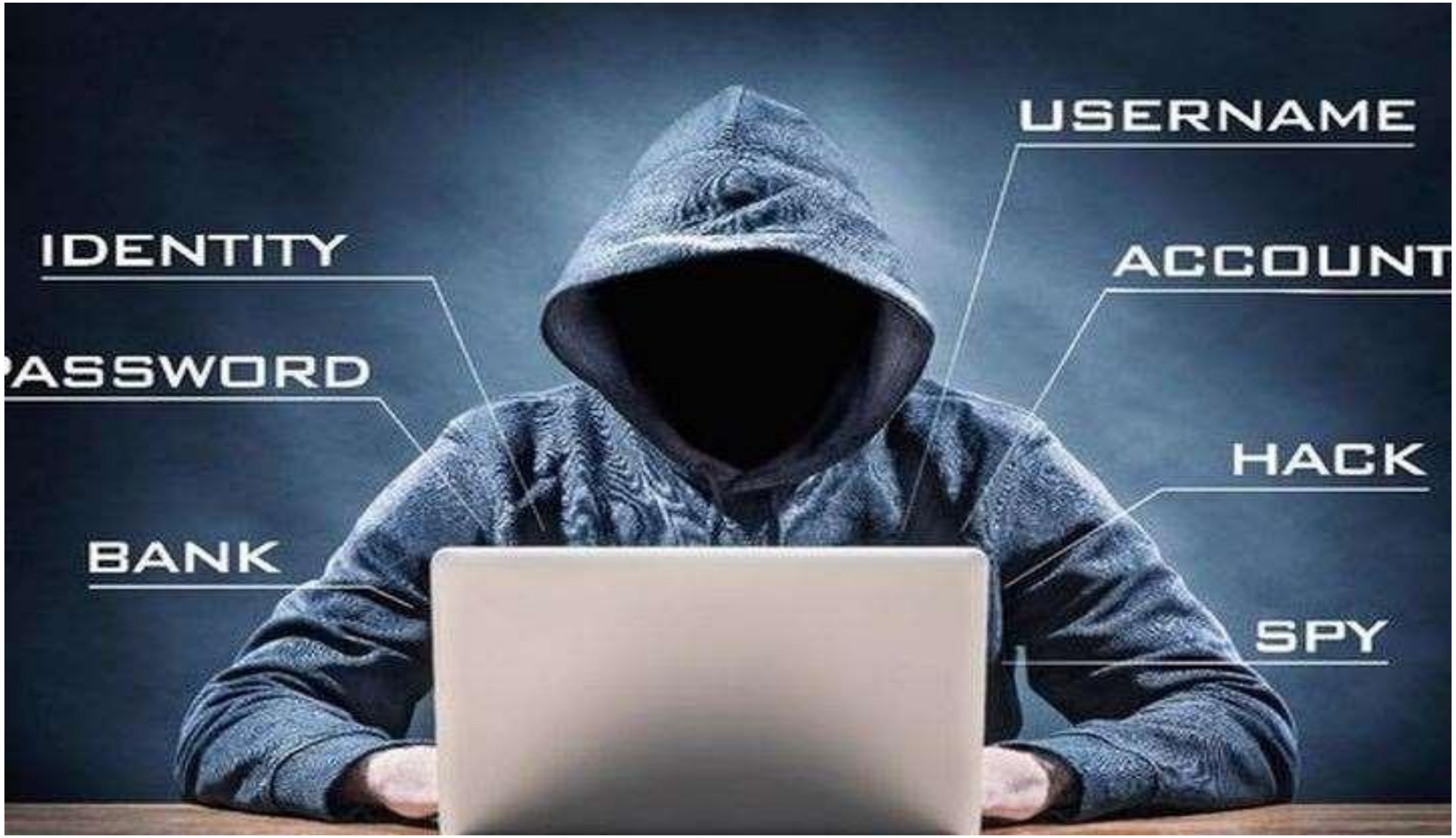
- **Airline Routing Planning**



HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science (cont.)

- **Fraud and Risk Detection**



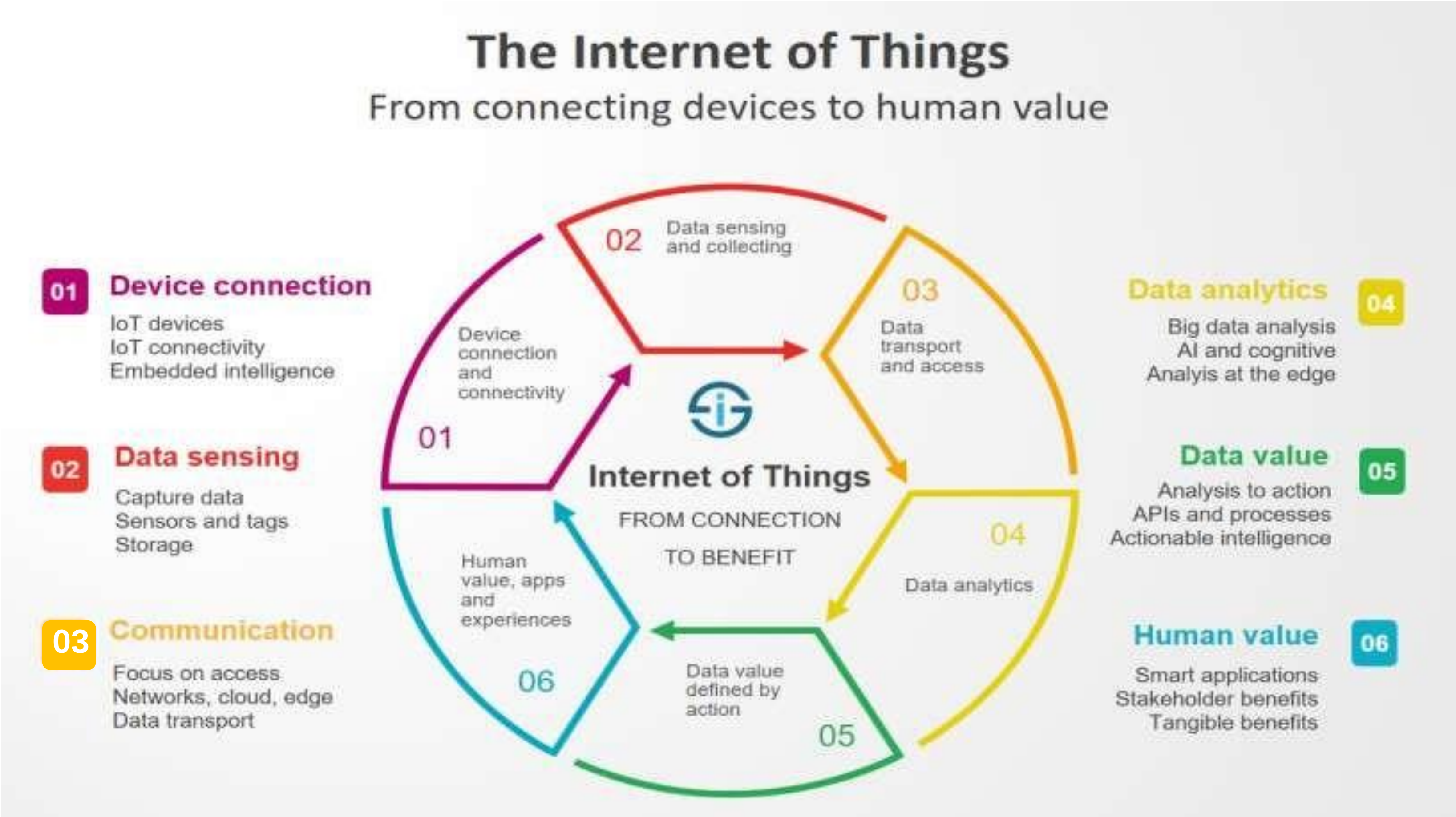
Applications of Data Science (cont.)

- **Delivery Logistics**

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science (cont.)

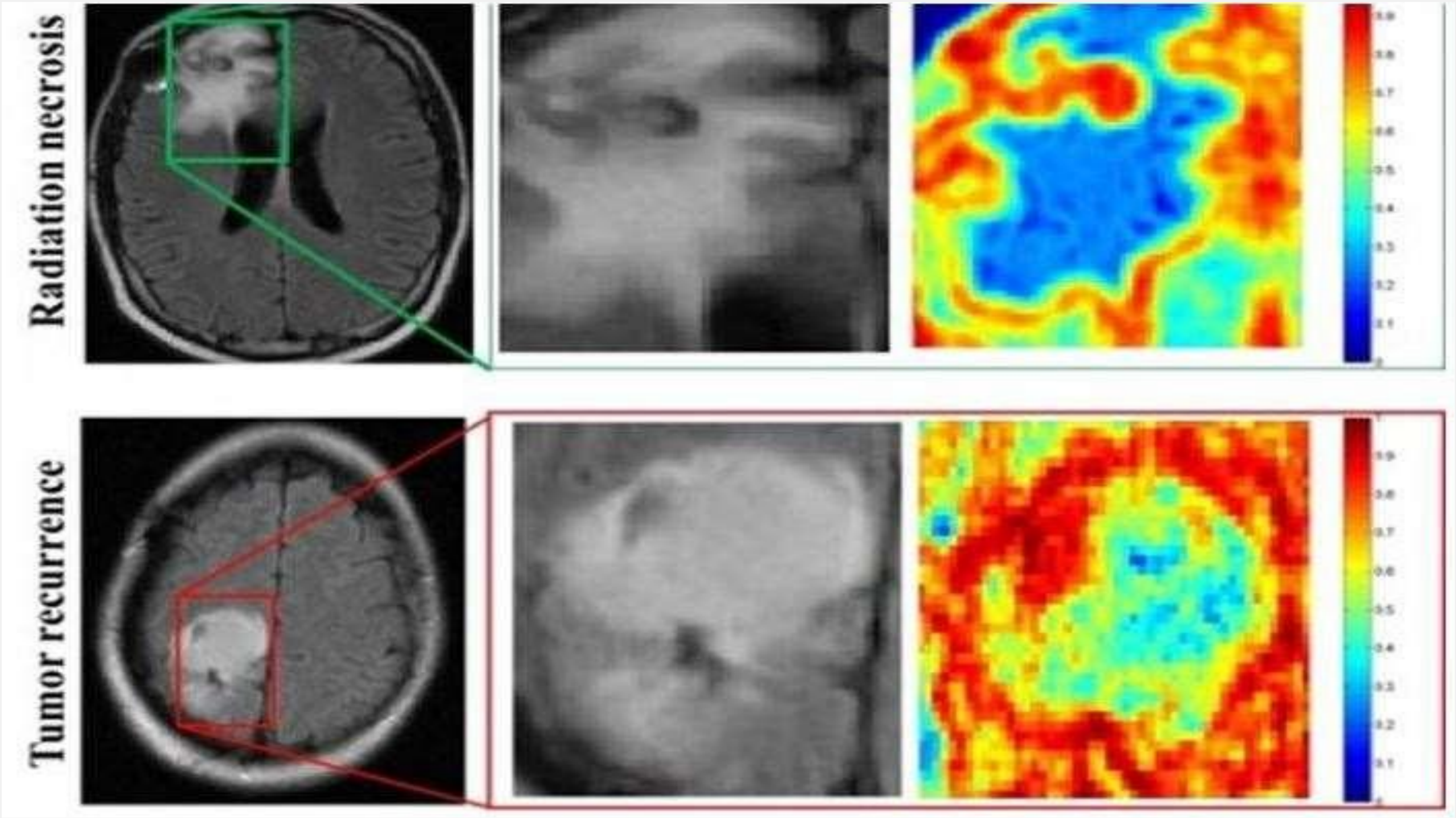
- **Internet of Things (IoT)**



HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Applications of Data Science (cont.)

- **Health Care**



Applications of Data Science (cont.)

- **Augmented Reality**

Applications of Data Science (cont.)

- **Self-Driving Cars**

Applications of Data Science (cont.)

- **Robots**

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Impact of Data Science on Society

- **Saving Energy**

Impact of Data Science on Society (cont.)

- **Data-Driven Hospitals**

Impact of Data Science on Society (cont.)

- **A Cleaner Environment**

Data Analytics Trends

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Scientific Machine Learning

DLR

Scientific Machine Learning

"Scientific machine learning (SciML) is a core component of artificial intelligence (AI) and a computational technology that can be trained, with scientific data, to augment or automate human skills.

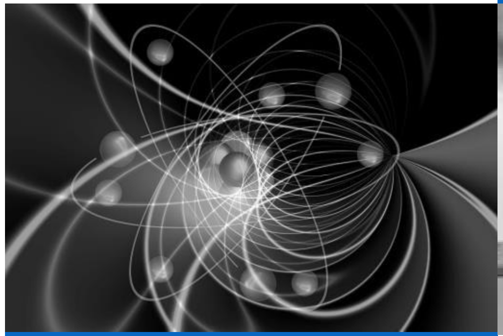
Across the Department of Energy (DOE), SciML has the potential to transform science and energy research. Breakthroughs and major progress will be enabled by harnessing DOE investments in massive data from scientific user facilities, software for predictive models and algorithms, high-performance computing platforms, and the national workforce."

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

SciML: Role for Model Reduction

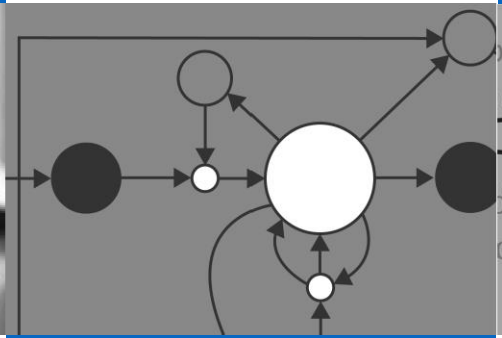
What role for model reduction?

1 reduce the cost of training **2** foundational shift in ML perspectives



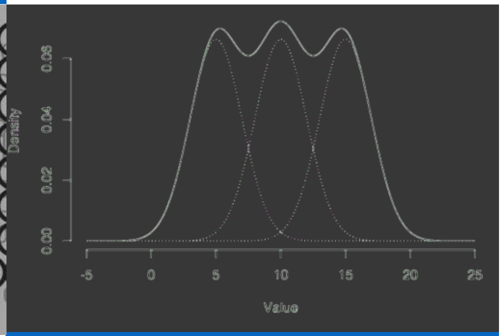
Respect physical constraints

Embed domain knowledge



Bring interpretability to results

Integrate heterogeneous, noisy & incomplete data



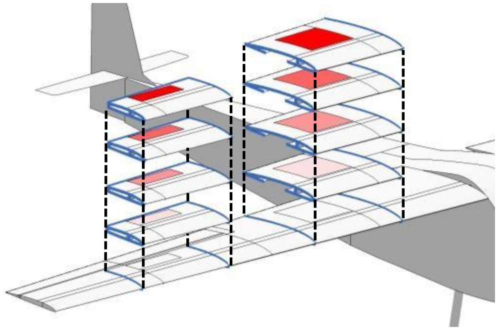
Get predictions with quantified uncertainties

SciML: Predictive Digital Twin

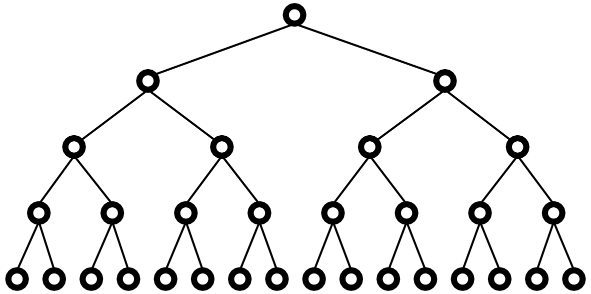
via component-based ROMs and interpretable machine learning

ROMs embed predictive modeling and reduce the cost of training

Offline:



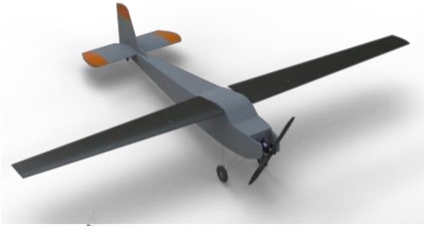
Construct library of ROMs representing different asset states

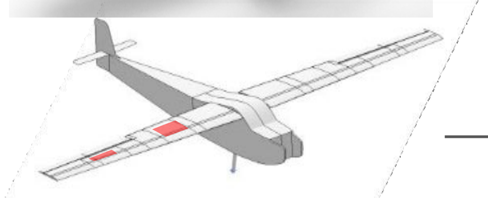


Use model library to train a classifier that predicts asset state based on sensor data

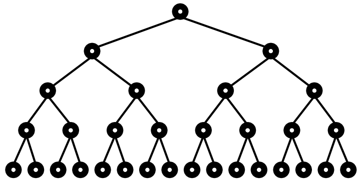
Online:

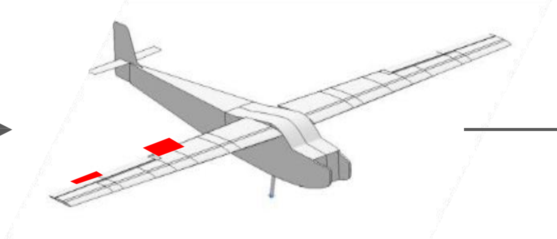
sensor data





current Digital Twin





updated Digital Twin

Analysis
Prediction
Optimization

[Kapteyn, Knezevic, W. AIAA Scitech 2020]

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

SciML: ML and MOR

Machine learning

“The scientific study of algorithms & statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns & inference instead.” [Wikipedia]

Reduced-order modeling

“Model order reduction (MOR) is a technique for reducing the computational complexity of mathematical models in numerical simulations.” [Wikipedia]

What is the connection between reduced-order modeling and machine learning?

Model reduction methods have grown from Computational Science & Engineering, with focus on ***reducing* high-dimensional models** that arise from physics-based modeling, whereas machine learning has grown from Computer Science, with a focus on ***creating* low-dimensional models** from black-box data streams. [Swischuk et al., *Computers & Fluids*, 2019]

SciML: ML and MOR (cont.)

Machine learning

"The scientific study of algorithms & statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns & inference instead." [Wikipedia]

Reduced-order modeling

"Model order reduction (MOR) is a technique for reducing the computational complexity of mathematical models in numerical simulations." [Wikipedia]

Reduced-order modeling and machine learning:

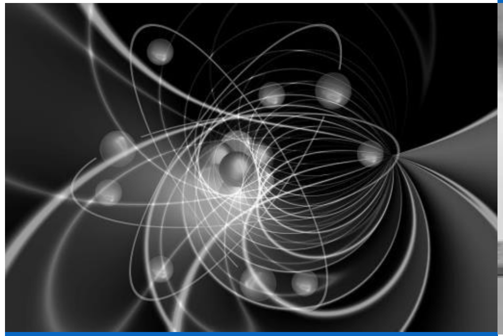
Can we get the best of both worlds?

Discover hidden structure	Embed governing equations
Non-intrusive implementation	Structure-preserving
Black-box & flexible	Predictive (error estimators)
Accessible & available	Stability-preserving

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

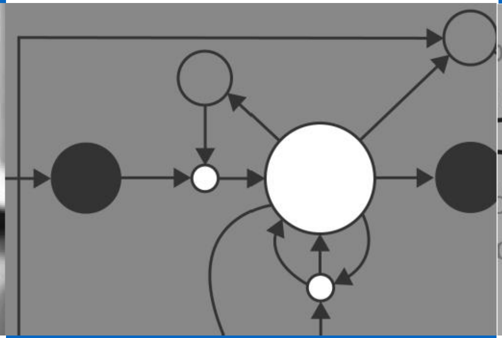
SciML: Strategy

Learning from data through the lens of models is a way to exploit structure in an otherwise intractable problem



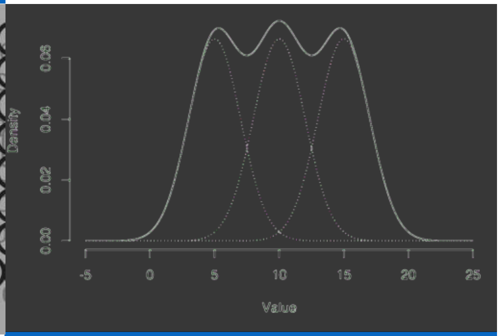
Respect physical constraints

Embed domain knowledge



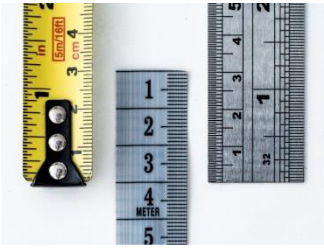
Bring interpretability to results

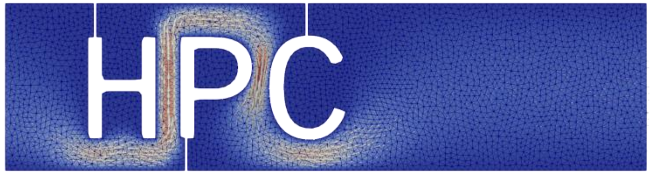
Integrate heterogeneous, noisy & incomplete data



Get predictions with quantified uncertainties

SciML@DLR: Space Applications of Statistics and Machine Learning





Tools:

- →Clustering, statistics
- →Neural networks + feedback loops
- → Optimal control theory
- →Uncertainty quantification

Datafusion
- Measurements
- Simulation

ML →

Dimensional Reduction
- Feature Extraction
- Reduced Order Modelling

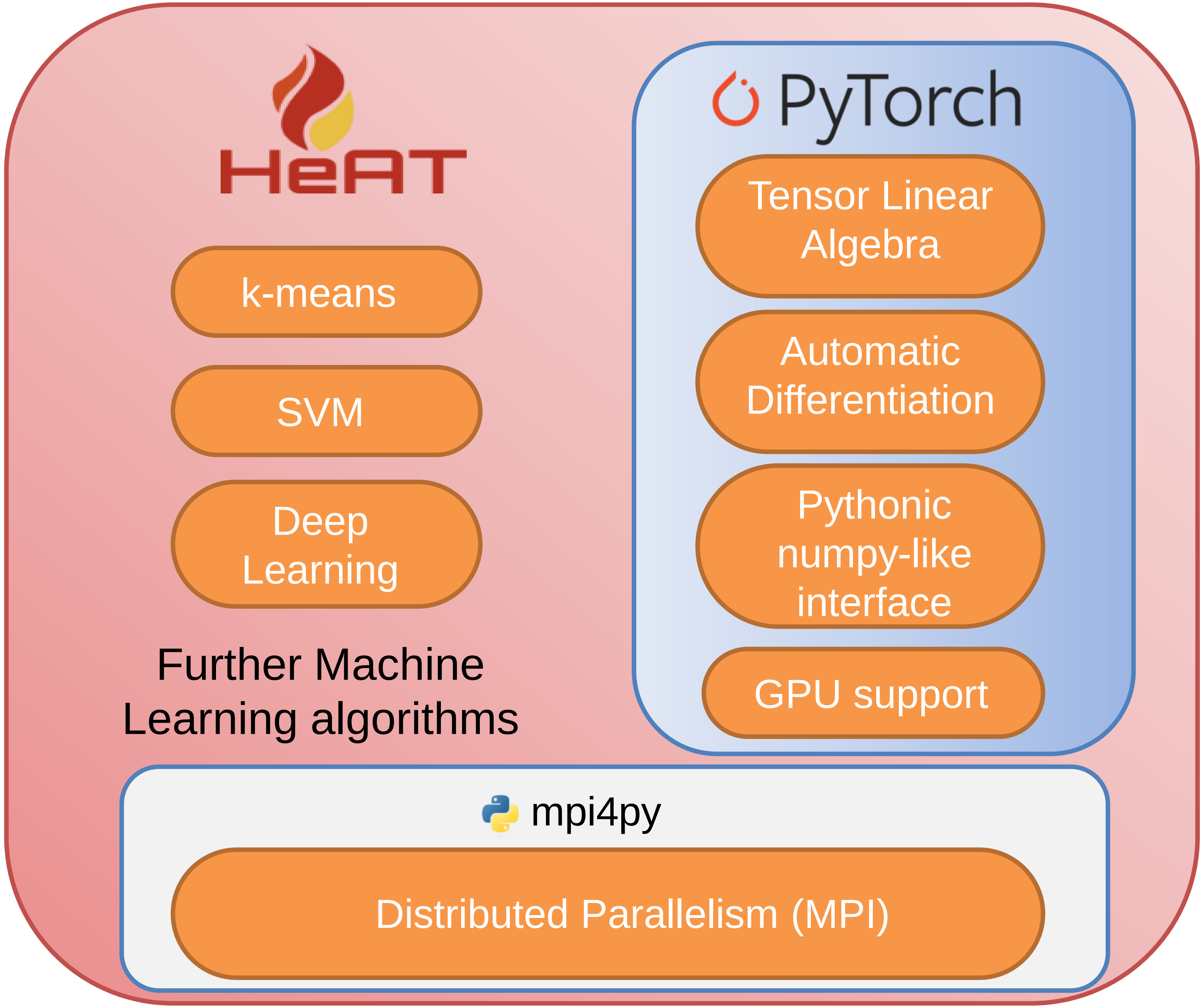
ML →

Real-time Prediction & Feedback
- Health Monitoring
- Robust Optimal Control

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

SciML@DLR: Software HeAT

- HeAT = Helmholtz Analytics Toolkit
- A framework for data analysis and Machine Learning, jointly developed by six Helmholtz Centers.
- Open Source with MIT License
- Available at https://github.com/helmholtz-analytics



HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

Many thanks for your attention!

Questions?

Dr.-Ing. Achim Basermann

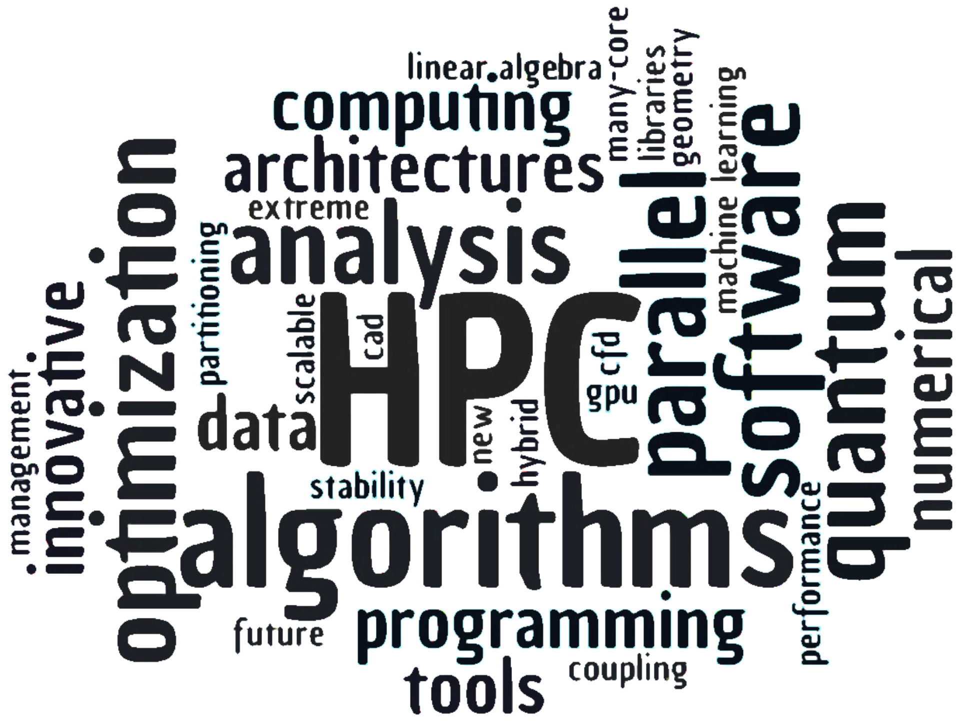
German Aerospace Center (DLR)

Simulation and Software Technology

Department Head High-Performance Computing

Achim.Basermann@dlr.de

http://www.DLR.de/sc



We are hiring: http://www.dlr.de/jobs/

DLR