Data Science

Evgeny Burnaev, Associate Professor
Skoltech

Skoltech
Skolkovo Institute of Science and Technology

Key Factors

- **Data Analysis is an old topic:**
  - expensive data storing, limited access to data, one-time data usage

- **Recent progress:**
  - Efficient capabilities to convert different types of information (texts, signals, images, video, etc.) into digital representation
  - Capabilities to store large volumes of digital data and to perform search/retrieval
  - Capabilities to fast transform
  - Fast transmission via the communication channels of large volumes of data (remote data access including simultaneous data access of a large number of users)
  - Computational capabilities for fast processing of big data (+ High Performance/ Distributed Computing + ...)

Data Science

Large amounts of data + New processing capabilities

⬇⬇⬇

Capabilities to pose and efficiently solve new scientific and applied problems statements

⬆⬆⬆

Scientific basis is elaborated in a new **multidisciplinary area of knowledge,** evolved in XXIst century in a new academic and university discipline called

«Data Science»

Skoltech
Skolkovo Institute of Science and Technology

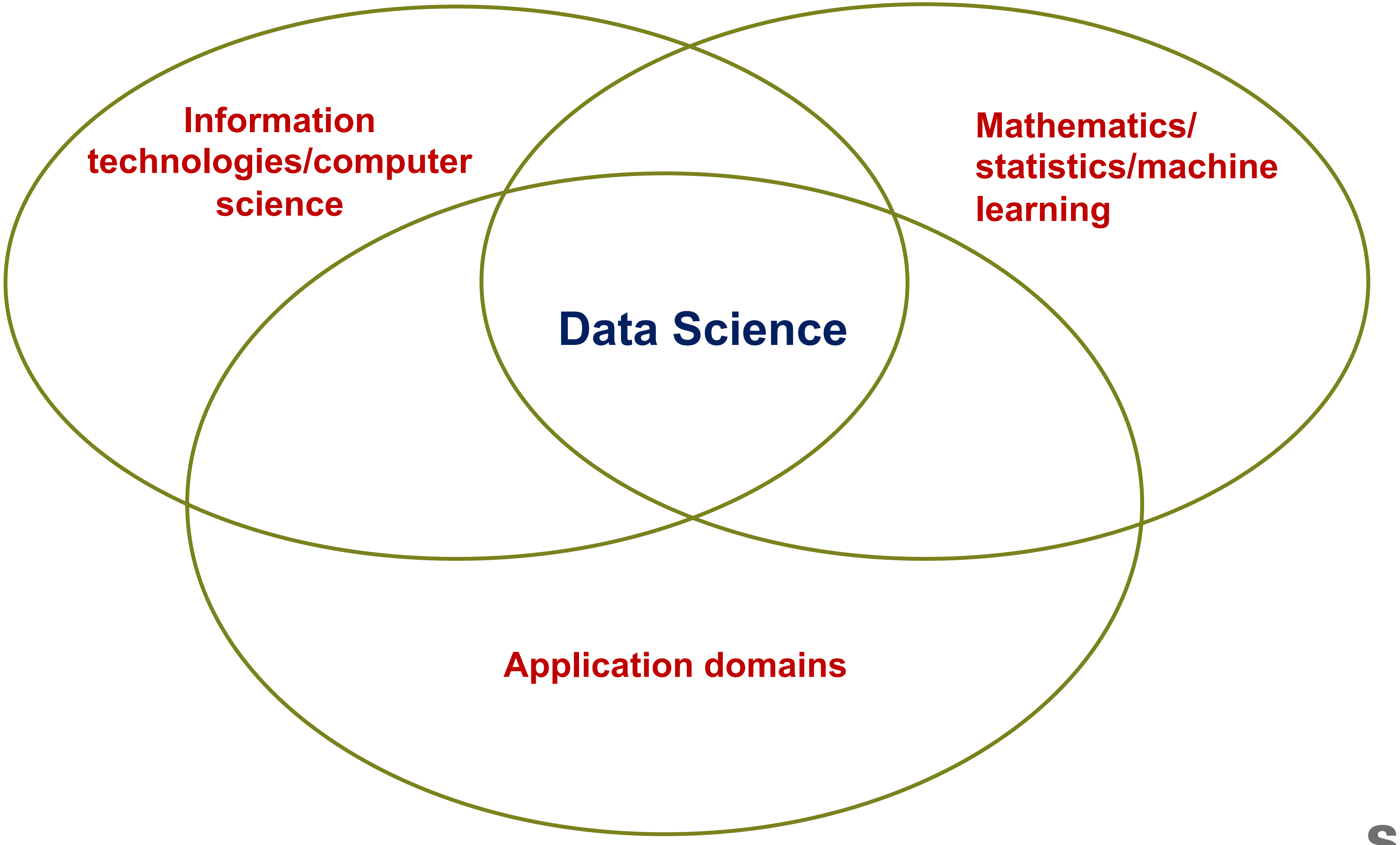
Data Science: overview

- **Data Science**: methods for data processing and analysis are used to extract tendencies, analyze and forecast behavior of observed engineering, socio-economical and biological systems

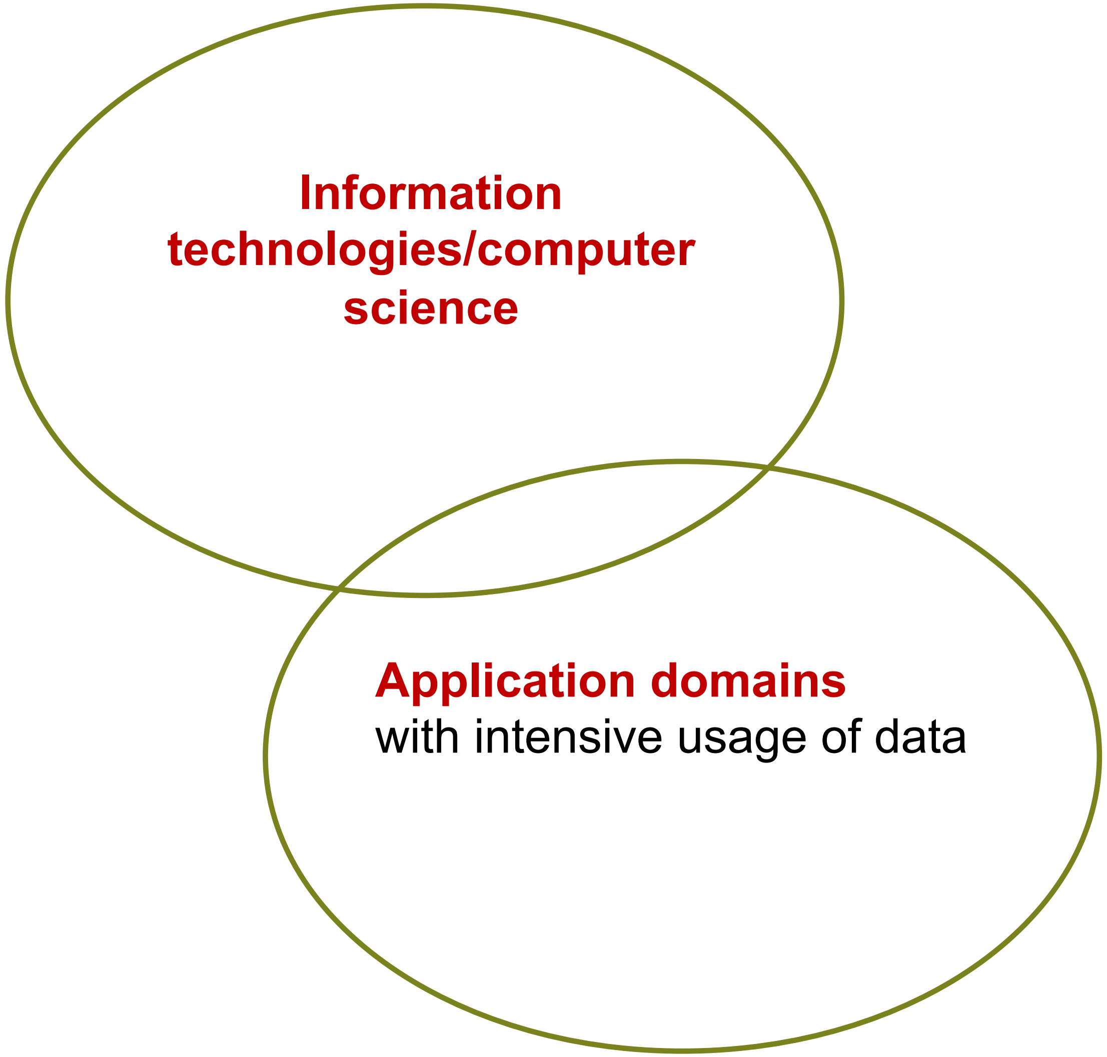
- **Various methods**: from mathematics and statistics; visualization, pattern recognition and machine learning, computer science, data mining, etc.

- **Technological basis**: data warehouses, high performance computing and distributed systems (including cloud/fog computing)

Data Science: structure


Information technologies/computer science
Mathematics/ statistics/machine learning
Data Science
Application domains

IT/CS vs. Appl. domain



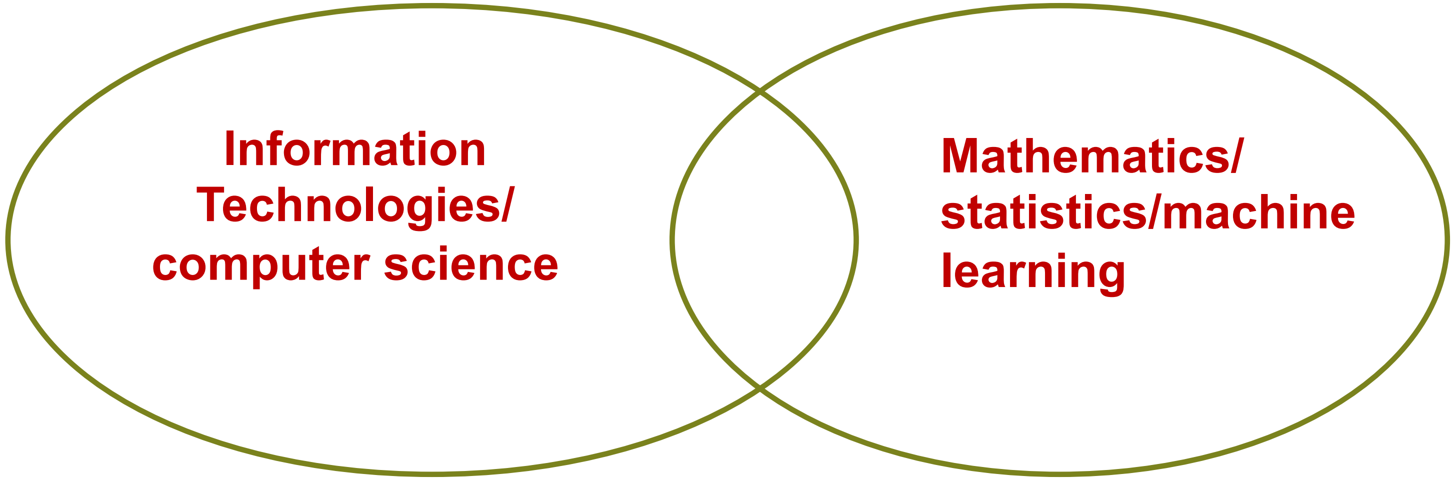
Information technologies/computer science

Application domains with intensive usage of data

IT: methods, algorithms, procedures to analyze and process data (clustering, classification, approximation, forecasting, …) including Software to solve problems from various applications domains, such that:

- Anomaly detection and its interpretation, prediction of failures, fraud detection, churn prediction, etc.
- Selection of dominant attributes,
- Identification and analysis of relationships (finding dependencies, identification of affiliations between different objects/events), forecasting, etc.

IT/CS vs. Math/Stat/ML

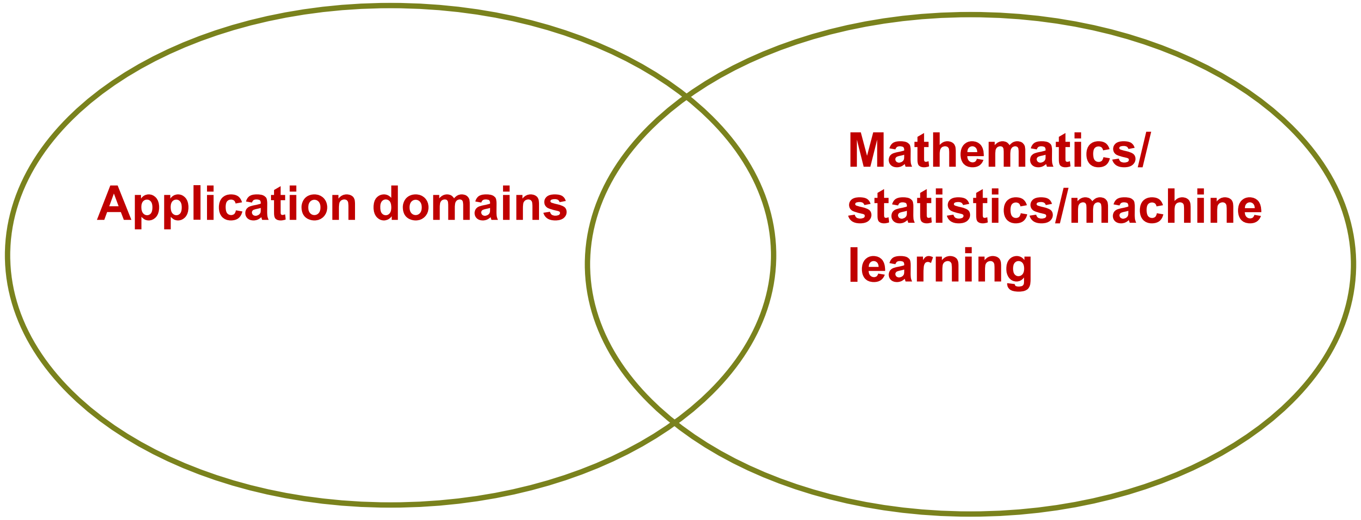

Information Technologies/ computer science
Mathematics/ statistics/machine learning

Mathematics: strict solutions of **formal** problem statements

- Methods, algorithms, procedures for data analysis, which we can either applied straightforwardly (after software realization), or use when developing some heuristics for data analysis

- Evaluate accuracy of developed or existing methods (algorithms, procedures) in order to determine the limits of applicability of the applied algorithms and/or to identify the bottlenecks of these algorithms, etc.

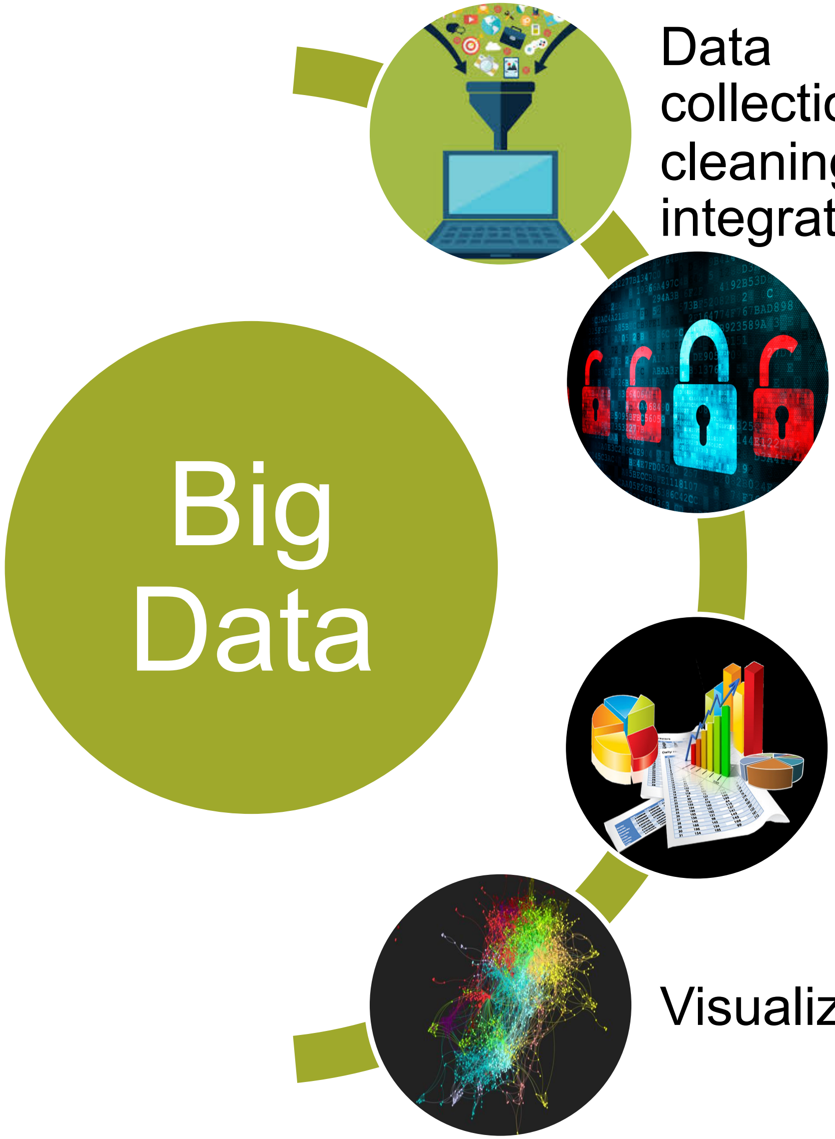
Appl. domain vs. Math/Stat/ML



Mathematics:

- strict math. problem statement (mathematical model), adequate to the problems of the subject area; take into account features (structure, properties, …) of input data
- finding solutions that allow efficient computational implementation, and have meaningful interpretation within the application domain

Buzzwords: what is Big Data?



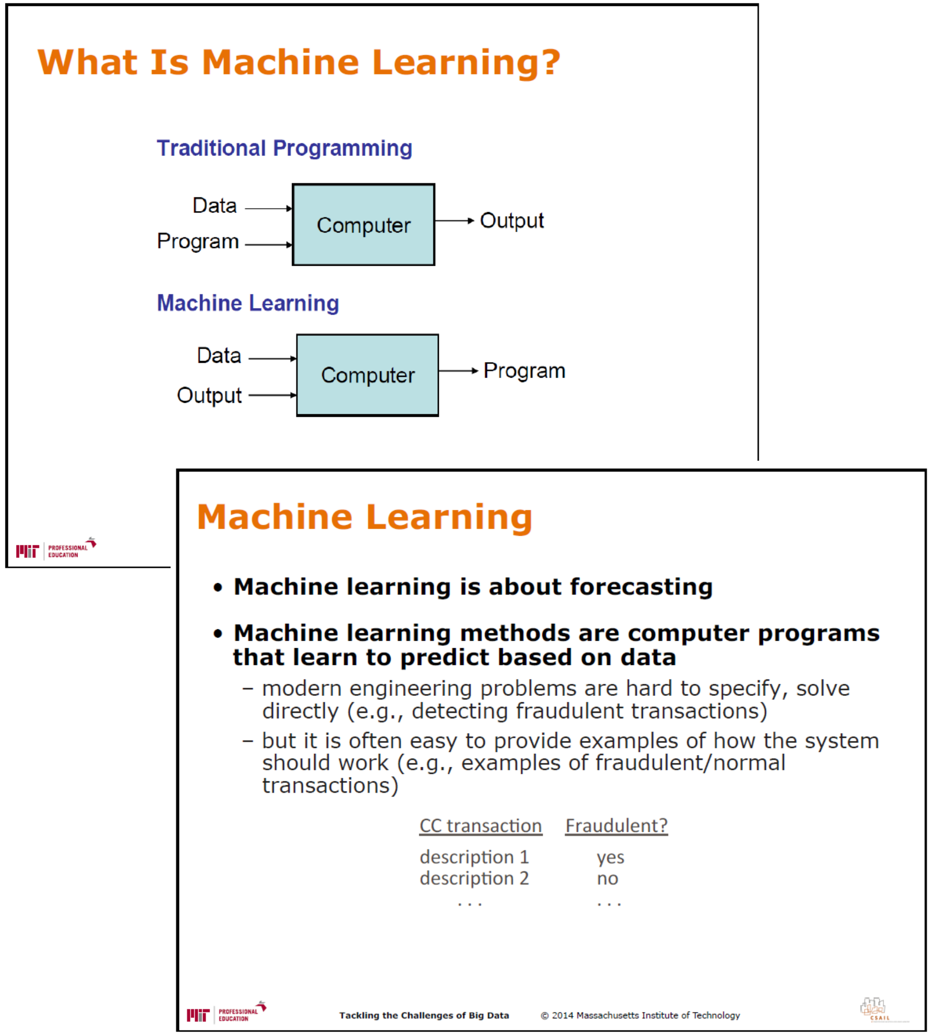
Big Data

Data collection, cleaning and integration

Platforms, security, data storing

Analytics and Machine Learning

Visualization



What Is Machine Learning?

Traditional Programming

Data → Computer → Output
Program → Computer

Machine Learning

Data → Computer → Program
Output → Computer

Machine Learning

- **Machine learning is about forecasting**
- **Machine learning methods are computer programs that learn to predict based on data**
  - modern engineering problems are hard to specify, solve directly (e.g., detecting fraudulent transactions)
  - but it is often easy to provide examples of how the system should work (e.g., examples of fraudulent/normal transactions)

CC transaction	Fraudulent?
description 1	yes
description 2	no
. . .	. . .

Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

Buzzwords: a Glossary of Artificial-Intelligence Terms

Artificial Intelligence –

➜ science and technology to create intelligent machines and computer programs

➜ ability of intelligent machines to perform creative and analytic functions

➜ the broadest term, applying to any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)

Machine Learning –

➜ a broad subfield of Artificial Intelligence,

➜ the mathematical discipline aimed at

- ✓ extracting patterns from data and based on mathematical statistics, numerical methods, optimization, probability theory, discrete analysis, geometry, etc.
- ✓ enabling machines to improve at tasks with experience. The category includes deep learning

Data Mining –

➜ an umbrella term for methods aimed at identifying knowledge and regularities in data, which are

- ✓ unknown a priori and are non-trivial
- ✓ practically important and can be interpreted
- ✓ necessary to make decisions

изображение с thefuturesagency.com

Skoltech
Skolkovo Institute of Science and Technology