

Основы машинного обучения

Поляк Марк Дмитриевич

2025



Введение

Лекция 1. Раздел 1.

Мотивация



Top 10 fastest growing jobs

1.	AI and Machine Learning Specialists
2.	Sustainability Specialists
3.	Business Intelligence Analysts
4.	Information Security Analysts
5.	Fintech Engineers
6.	Data Analysts and Scientists
7.	Robotics Engineers
8.	Big Data Specialists
9.	Agricultural Equipment Operators
10.	Digital Transformation Specialists

Source

World Economic Forum, Future of Jobs Report 2023.

Top 10 fastest declining jobs

1.	Bank Tellers and Related Clerks
2.	Postal Service Clerks
3.	Cashiers and ticket Clerks
4.	Data Entry Clerks
5.	Administrative and Executive Secretaries
6.	Material-Recording and Stock-Keeping Clerks
7.	Accounting, Bookkeeping and Payroll Clerks
8.	Legislators and Officials
9.	Statistical, Finance and Insurance Clerks
10.	Door-To-Door Sales Workers, News and Street Vendors, and Related Workers

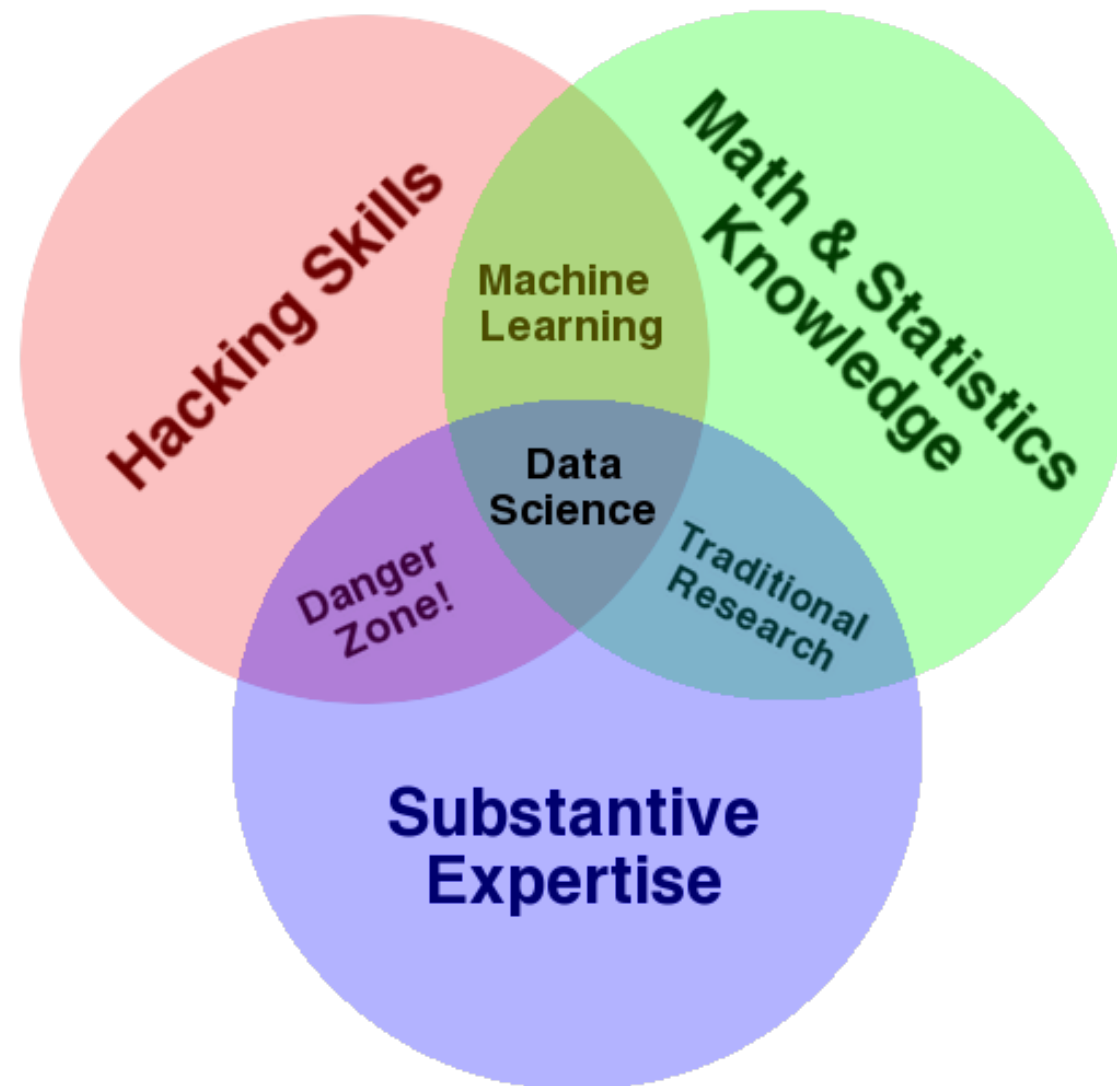
Note

The jobs which survey respondents expect to grow most quickly from 2023 to 2027 as a fraction of present employment figures

Специальности и вакансии

- **Machine Learning Engineer/Developer/Researcher**
- **Data Scientist**
- **Data Analyst**
- Data Engineer
- Data Architect
- MLOps
- Business Analyst
- Systems Analyst
- DevOps
- SecOps
- DevSecOps
- Software Developer
- ...

Кто такой Data Scientist?



Машинное обучение: определение

Машинное обучение, МО – это наука (и искусство) программирования вычислительных машин таким образом, чтобы они могли **учиться на основе данных**.

Более общее определение:

Машинное обучение — это область знаний, которая изучает способы **обучения** вычислительных машин без их явного программирования.

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.

- *Arthur Samuel, 1959*

Определение для инженера:

Говорят, что компьютерная программа **обучается** на опыте E в отношении некоторой задачи T и некоторой меры производительности P , если её производительность в задаче T , измеряемая с помощью P , улучшается с накоплением опыта E .

A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .

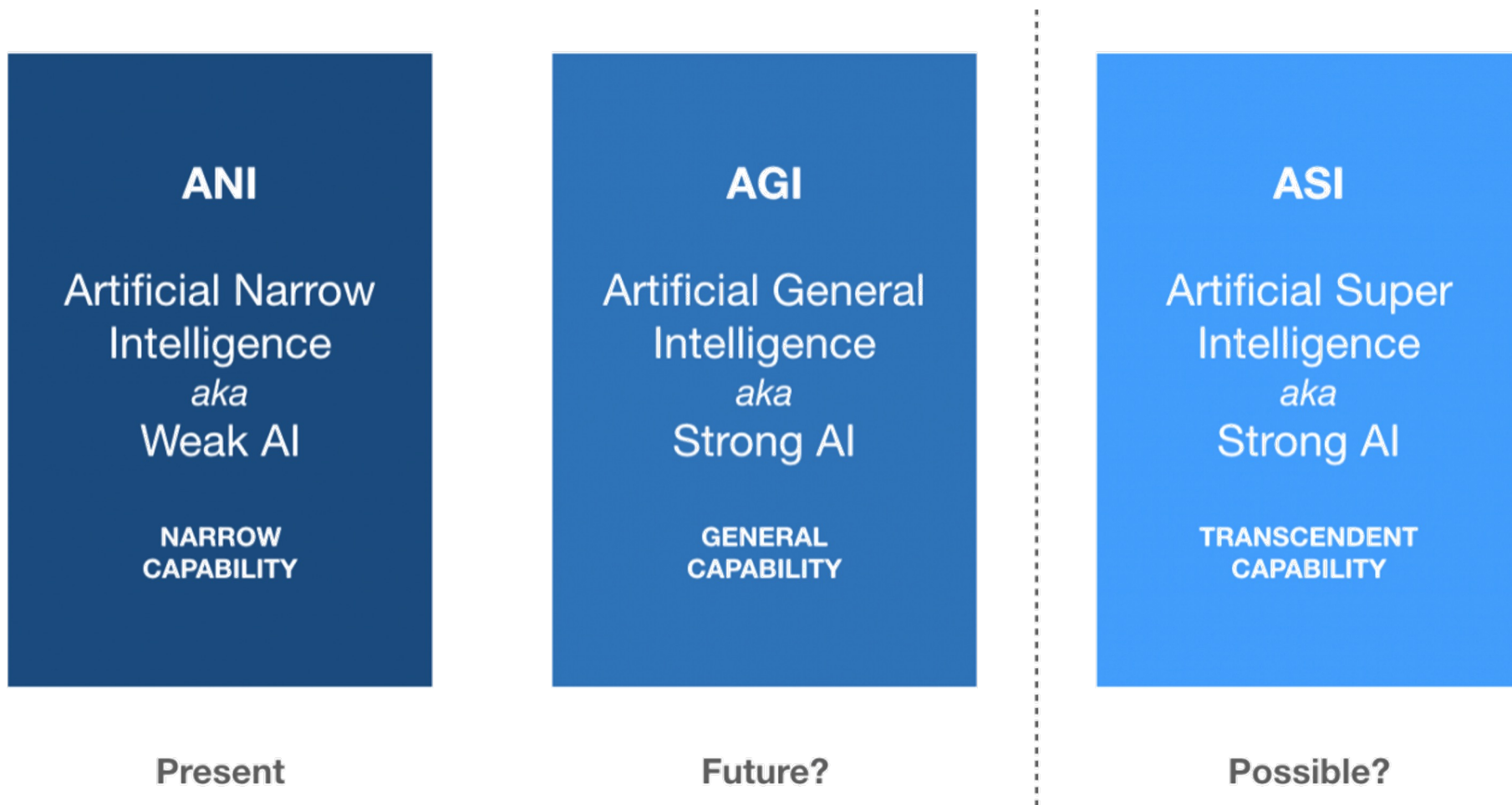
- *Tom Mitchell, 1997*

Искусственный интеллект

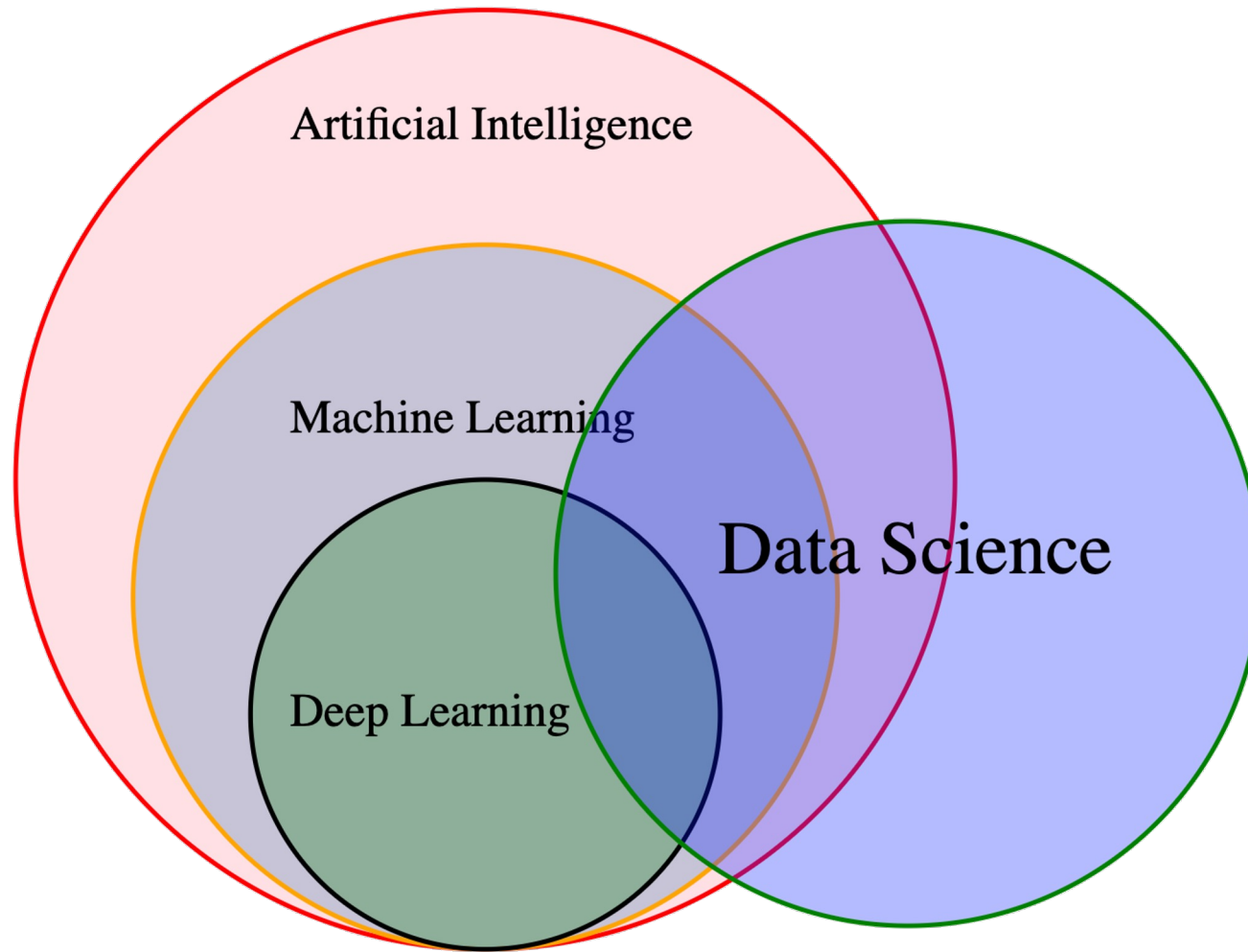
- Искусственный интеллект (ИИ; Artificial Intelligence, AI) — это интеллект, демонстрируемый машинами, в отличие от естественного интеллекта, проявляемого животными, включая людей
- Ведущие учебники по ИИ определяют эту область как изучение «интеллектуальных агентов»: любой системы, которая воспринимает свою среду и предпринимает действия, максимизирующие её шансы на достижение целей

- [Wikipedia](#)

Уровни развития искусственного интеллекта



ИИ и машинное обучение



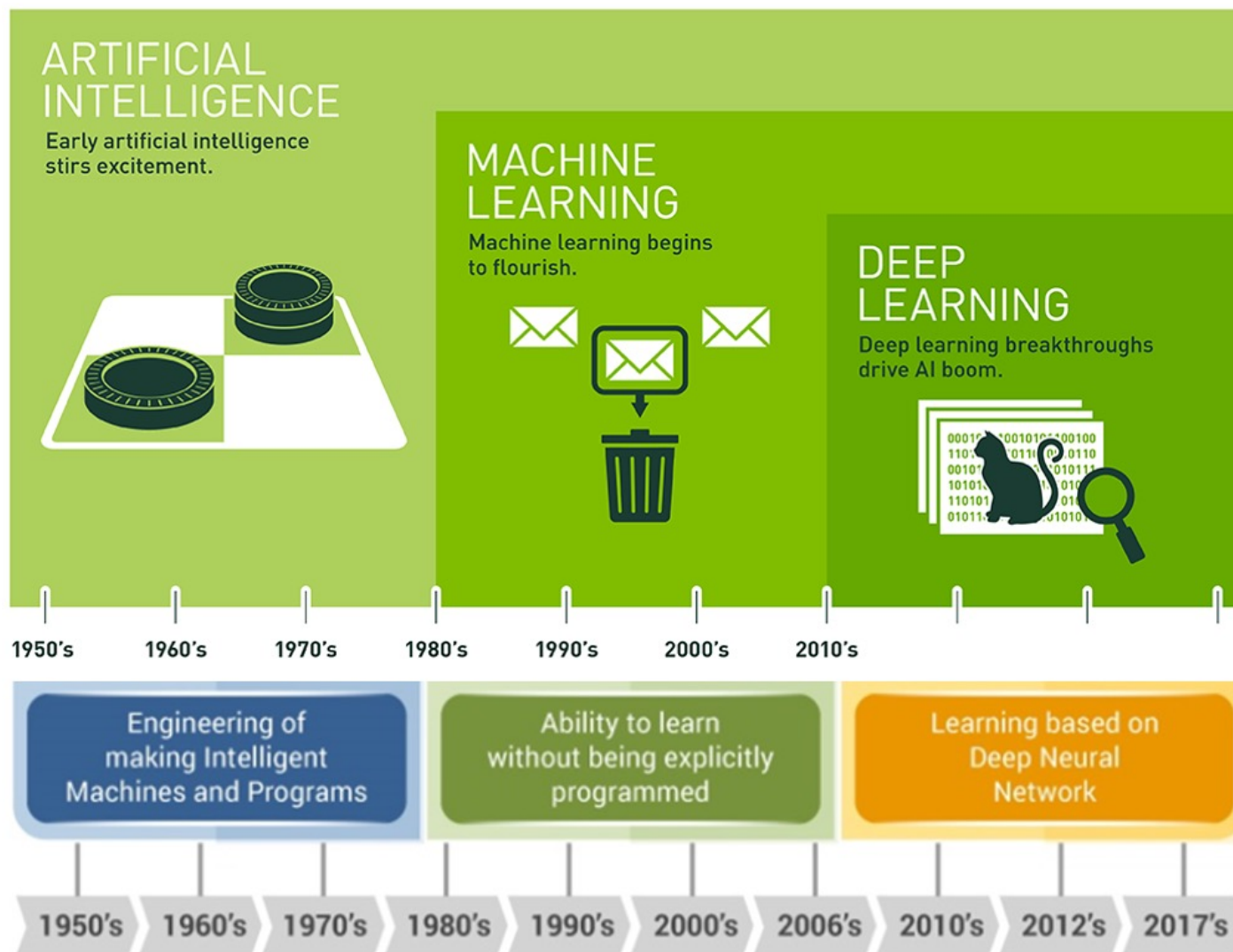
Чем (не) является машинное обучение

- Машинное обучение – это:
 - Математика + Статистика + Программирование
 - Алгоритмический подход к обработке (больших) данных
 - Слабый искусственный интеллект
- Машинное обучение **не является:**
 - Полноценным искусственным интеллектом (т.н. Общим ИИ / Сильным ИИ)
 - Осведомленным о предметной области (домене)
 - Панацеей от всех проблем человечества

Машинное обучение дополняет человеческий интеллект, но не может предоставить машине интеллект на уровне человека [1]

[1] <https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says>

История развития ИИ, МО и глубокого обучения



Области применения машинного обучения



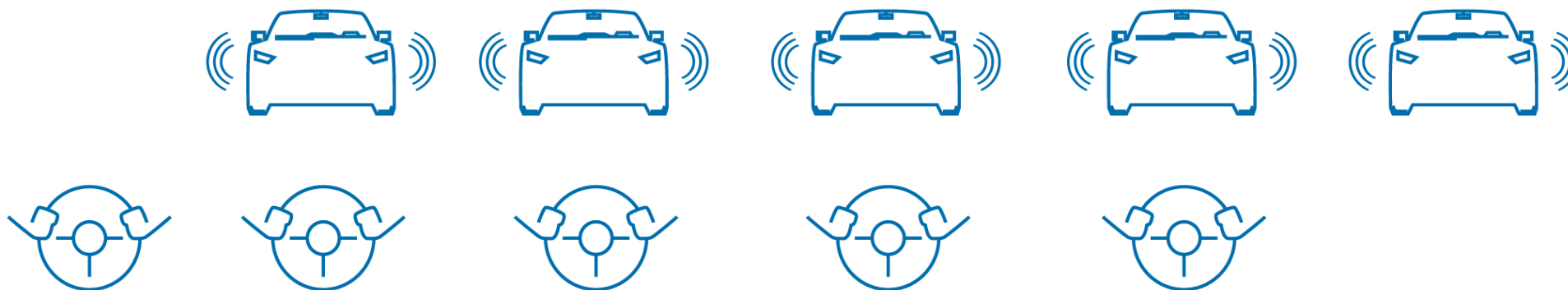
... и множество других!

Применение МО: беспилотные автомобили



Применение МО: беспилотные автомобили

FIVE LEVELS OF VEHICLE AUTONOMY



LEVEL 0

No Automation:

the driver is in complete control of the vehicle at all times.

LEVEL 1

Driver Assistance:

the vehicle can assist the driver or take control of either the vehicle's speed, through cruise control, or its lane position, through lane guidance.

LEVEL 2

Occasional Self-Driving:

the vehicle can take control of both the vehicle's speed and lane position in some situations, for example on limited-access freeways.

LEVEL 3

Limited Self-Driving:

the vehicle is in full control in some situations, monitors the road and traffic, and will inform the driver when he or she must take control.

LEVEL 4

Full Self-Driving Under Certain Conditions:

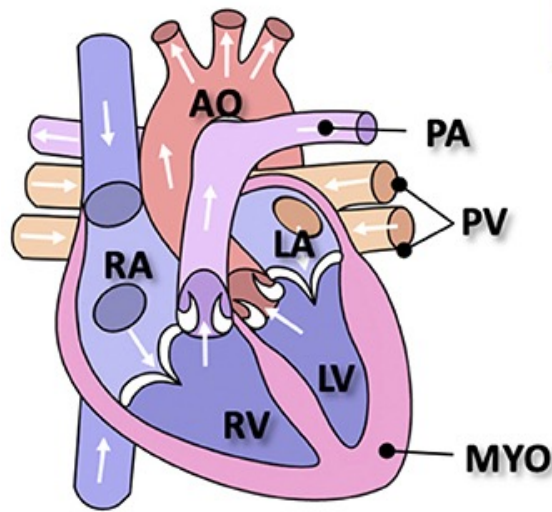
the vehicle is in full control for the entire trip in these conditions, such as urban ride-sharing.

LEVEL 5

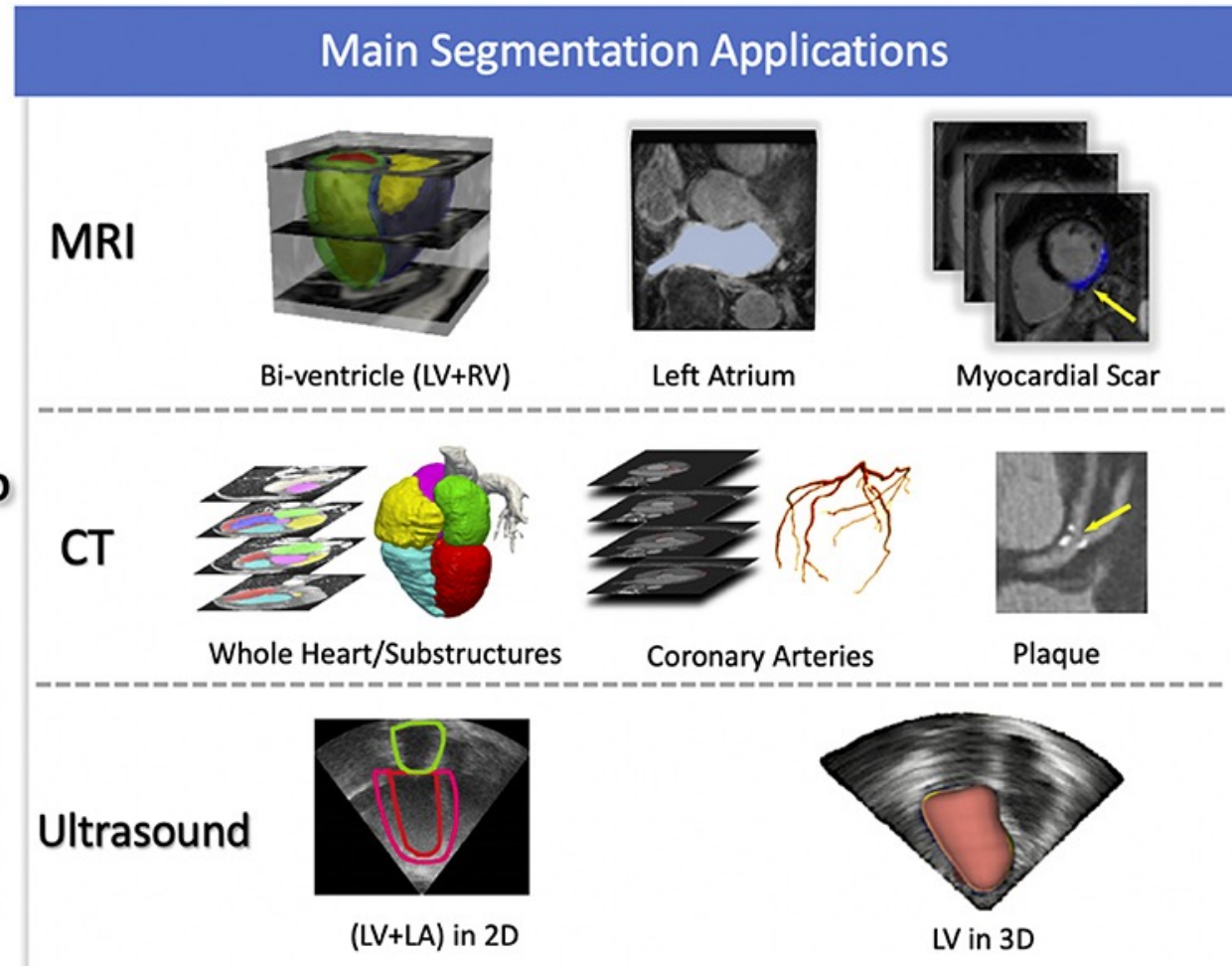
Full Self-Driving Under All Conditions:

the vehicle can operate without a human driver or occupants.

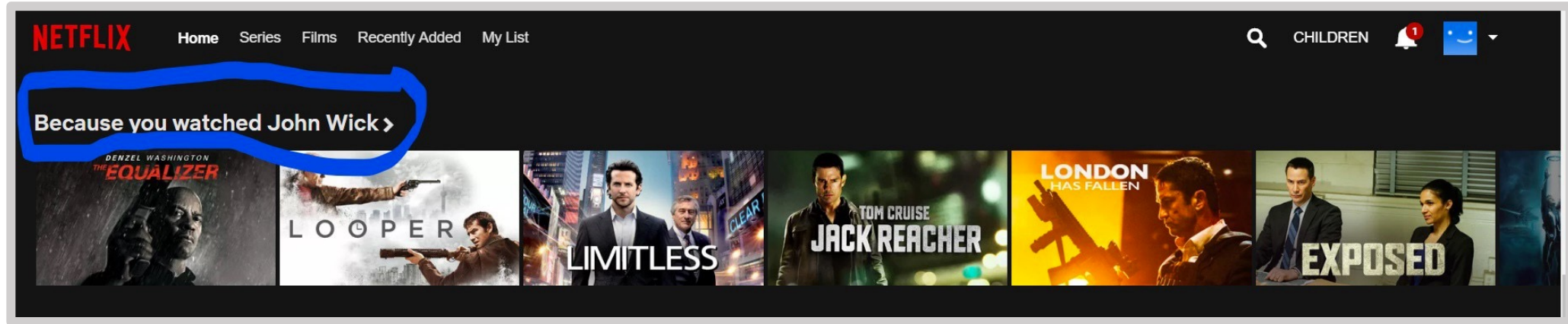
Применение МО: сегментация медицинских изображений



- AO: Aorta
- PA: Pulmonary Arteries
- PV: Pulmonary Veins
- RA: Right Atrium
- RV: Right Ventricle
- LA: Left Atrium
- LV: Left Ventricle
- MYO: Myocardium



Применение МО: рекомендательные системы



A screenshot of the Amazon product page for the book 'Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition' by Aurélien Géron. The page shows the book's title, author, a 4.5-star rating from 2,475 ratings, and a 'Best Seller' badge. Below the main product, there's a section titled 'Customers who viewed this item also viewed' showing a carousel of related books, including 'Practical Statistics for Data Scientists', 'Python for Data Analysis', 'Deep Learning with Python', 'Python Data Science Handbook', 'Deep Learning (Adaptive Computation and Machine Learning series)', 'The Hundred-Page Machine Learning Book', and 'Introduction to Machine Learning with Python'. The page also includes navigation links like 'All', 'Today's Deals', 'Customer Service', 'Registry', 'Gift Cards', and 'Sell'.

A screenshot of an Avito marketplace listing for 'Марки с автомобилями' (Stamps with cars) for 100 rubles. The listing is located in Moscow under the category 'Хобби и отдых' (Hobbies and leisure). Below the main listing, there's a section titled 'Похожие объявления' (Similar listings) showing three other listings: 'Марки СССР флора цветы 1988-1989' for 70 rubles, 'Альбом марок СССР 1988-1989' for 1,500 rubles, and 'Марки СССР. 1978. Соп. 4868-73. Авиация' for 50 rubles. Each listing includes a thumbnail image of the stamps and a brief description of the collection.

Рекомендуемая литература

- Geron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly, 2019
- Wes McKinney. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly, 2017
- VanderPlas J. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly, 2017
- Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006

Базовые концепции

Лекция 1. Раздел 2.

Постановка задачи обучения на примерах

- X – множество *объектов*
- Y – множество *ответов* (предсказаний, оценок, прогнозов)
- $\varphi(x), \varphi: X \rightarrow Y$ – неизвестная зависимость (target function)

Дано:

- $\{x_1, \dots, x_\ell\} \subset X$ – обучающая выборка (training sample)
- $y_i = \varphi(x_i), i = 1, \dots, \ell$ – известные ответы

Найти:

- $g(x, \theta), g: X \times \Theta \rightarrow Y$ – алгоритм, функция принятия решений или параметрическая модель, приближающая φ на всей выборке X
- $\theta \in \Theta$ – вектор параметров модели, такой, что $g(x, \theta) \approx \varphi(x)$

Обучение на примерах

Весь курс машинного обучения посвящен поиску ответов на следующие вопросы:

- Как задаются (описываются) множества объектов и ответов?
- Насколько точно алгоритм g аппроксимирует целевую функцию φ ?
- Как можно оценить алгоритм g ?

Описание объектов. Векторы признаков

$f_j: X \rightarrow D_j, j = 1, \dots, n$ – признаки объектов (features)

Типы скалярных признаков:

- $D_j = \{0,1\}$ – бинарный признак f_j ;
- $|D_j| < \infty$ – номинальный признак f_j ;
- $|D_j| < \infty, D_j$ упорядочено – порядковый признак f_j ;
- $D_j = \mathbb{R}$ – количественный признак f_j : интервал или число.

Вектор $(f_1(x), \dots, f_n(x))$ – *признаковое описание* объекта x .

Матрица признаков: $F = \|f_j(x_i)\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$

Описание ответов. Типы задач МО

Задача классификации:

- $Y = \{-1, +1\}$ – бинарная классификация (два класса);
- $Y = \{1, \dots, M\}$ – классификация между M не пересекающимися классами;
- $Y = \{0,1\}^M$ – M классов, которые могут пересекаться.

Задача регрессии:

- $Y = \mathbb{R}$ or $Y = \mathbb{R}^m$.

Задача ранжирования:

- Y – конечное отсортированное множество.