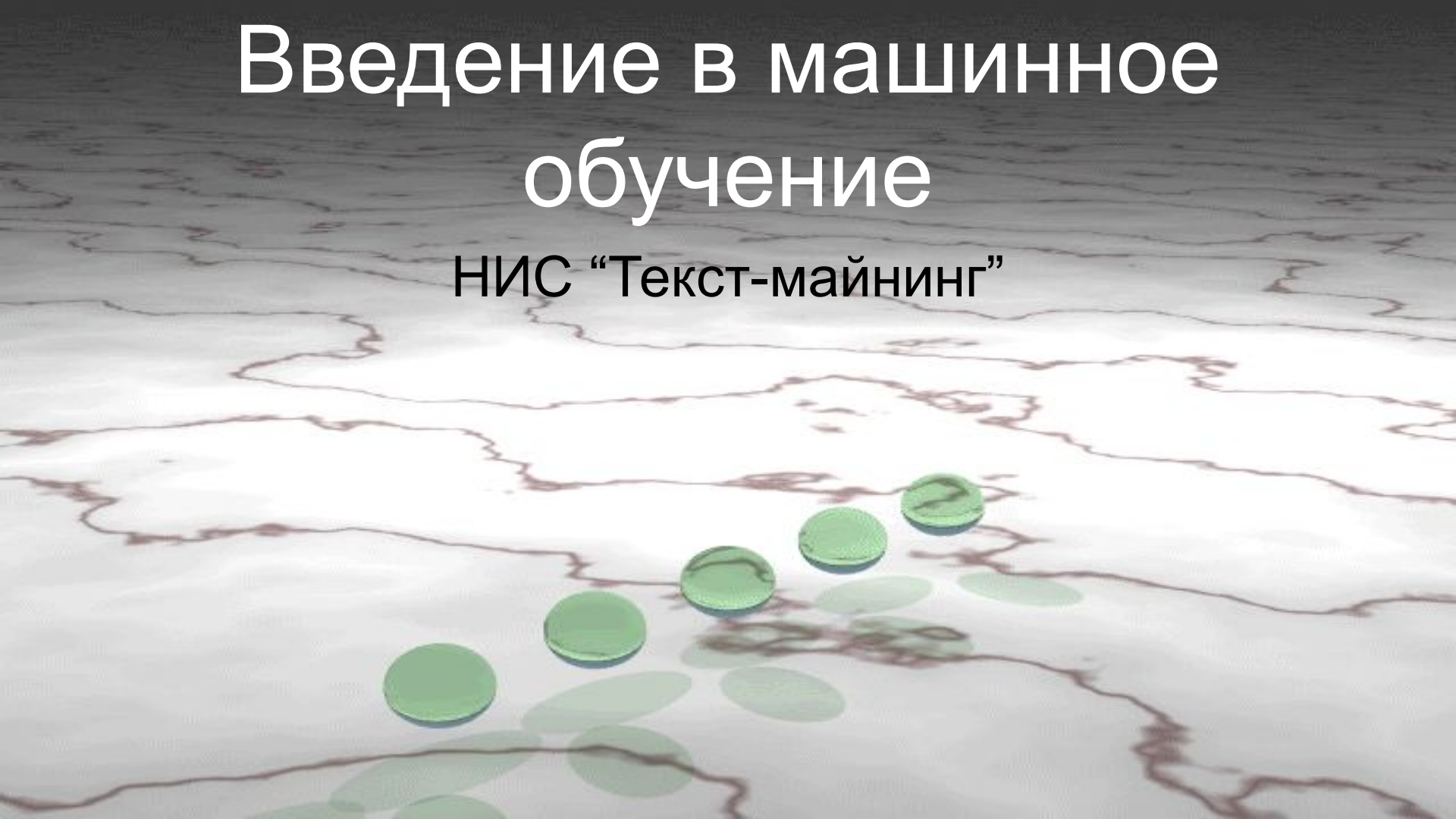
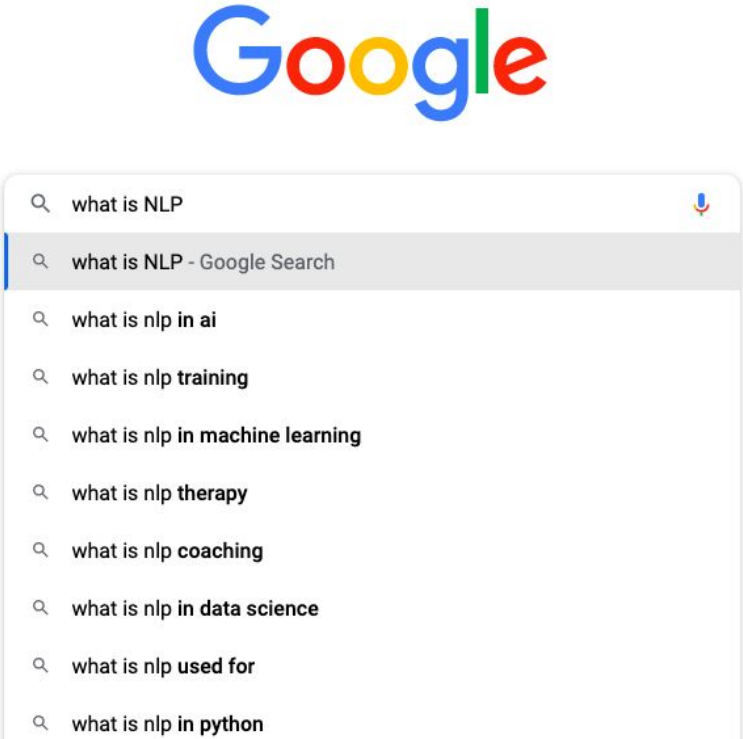
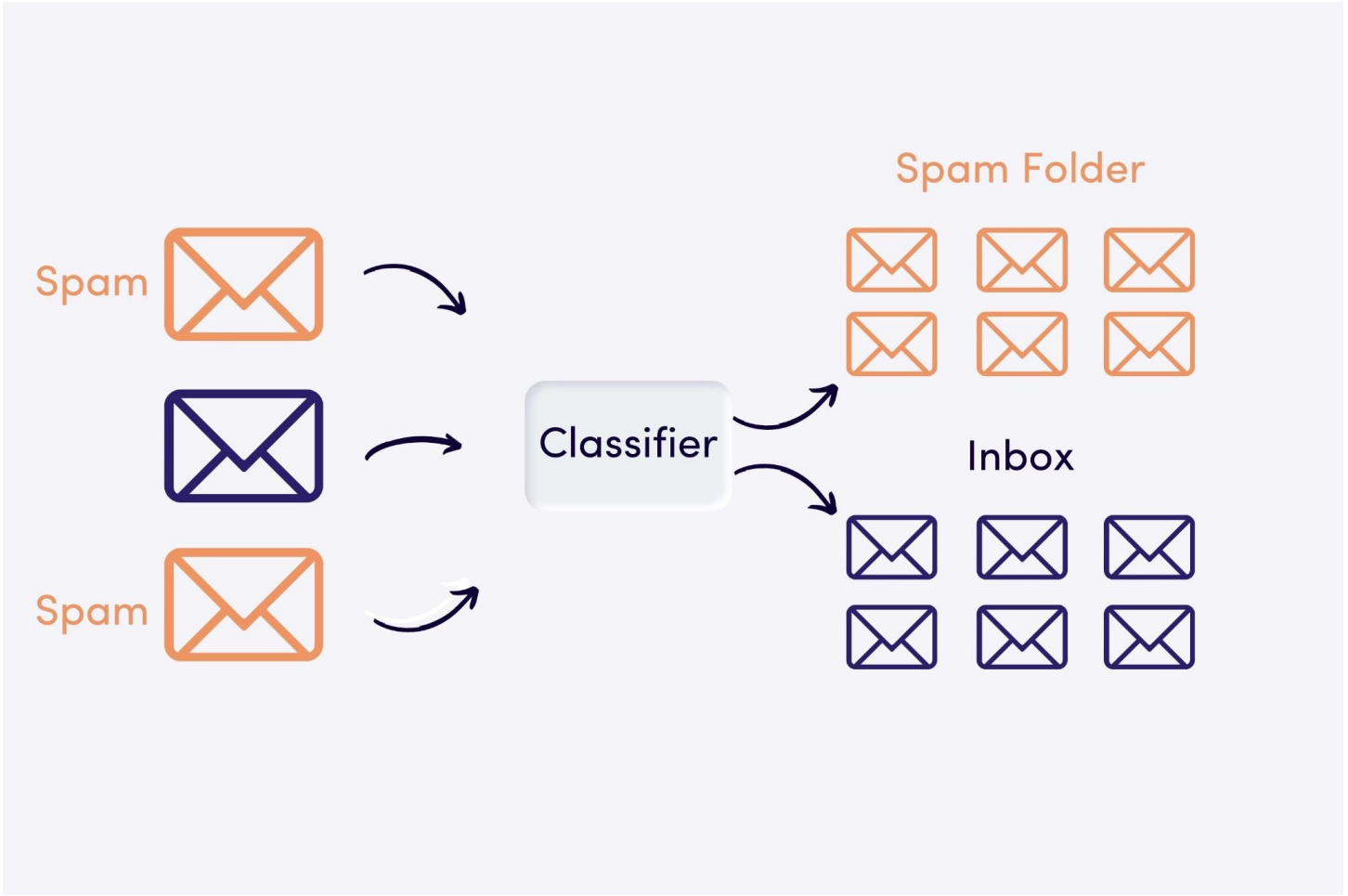
Введение в машинное обучение

НИС “Текст-майнинг”



Машинное обучение и тексты




Spam Folder
Spam
Classifier
Inbox
Spam

COCA-COLA

Coca-Cola Soda

- Overall
- KPIs
- Pros & cons
- Relevance
- Predictive
- Wonder

MORE

Notifications

Back

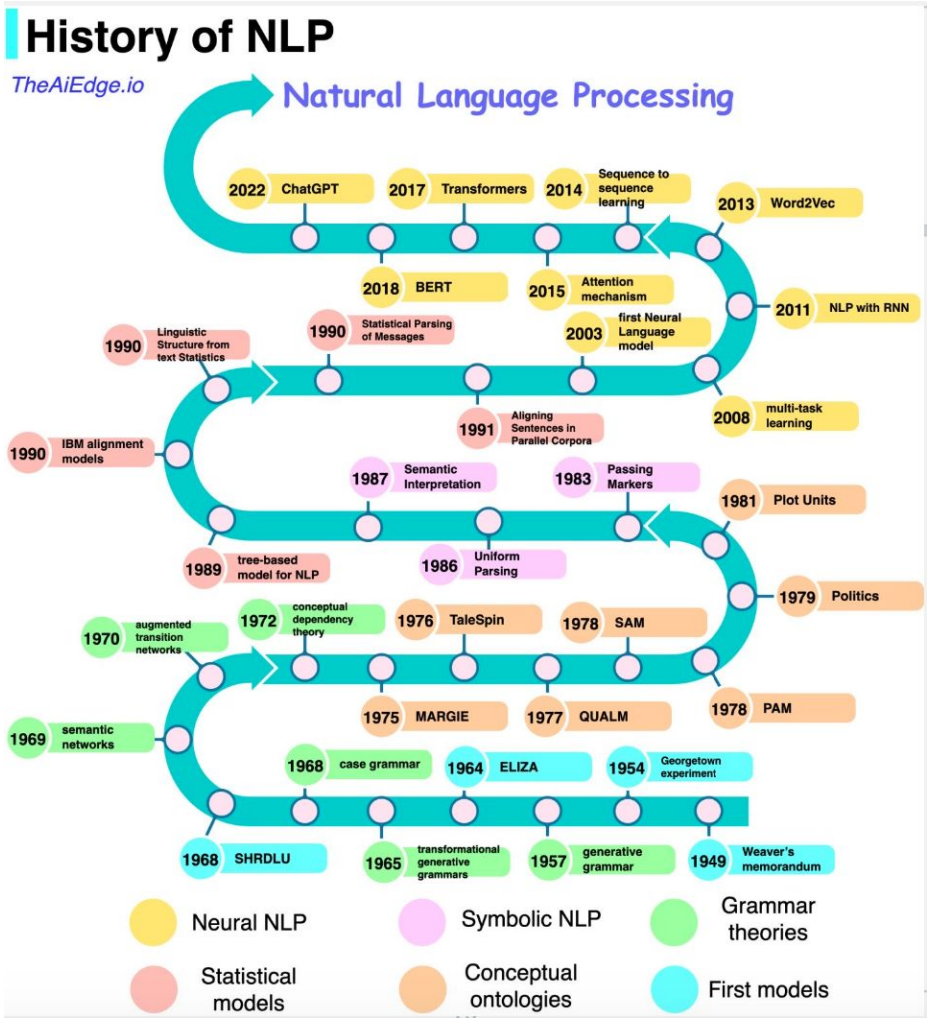
Wonder

Wonder

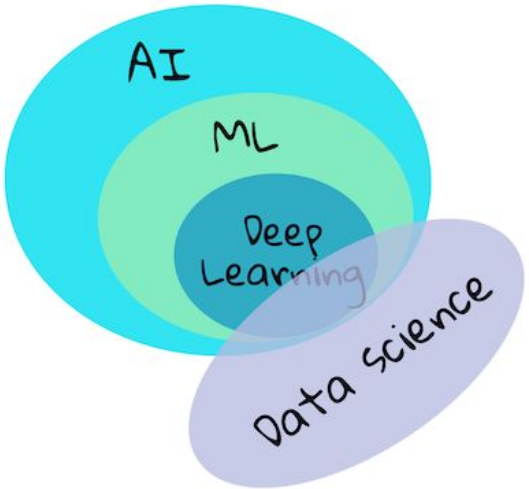
Insights powered by Wonderflow AI. Ask anything you would like to know, we have the answers.

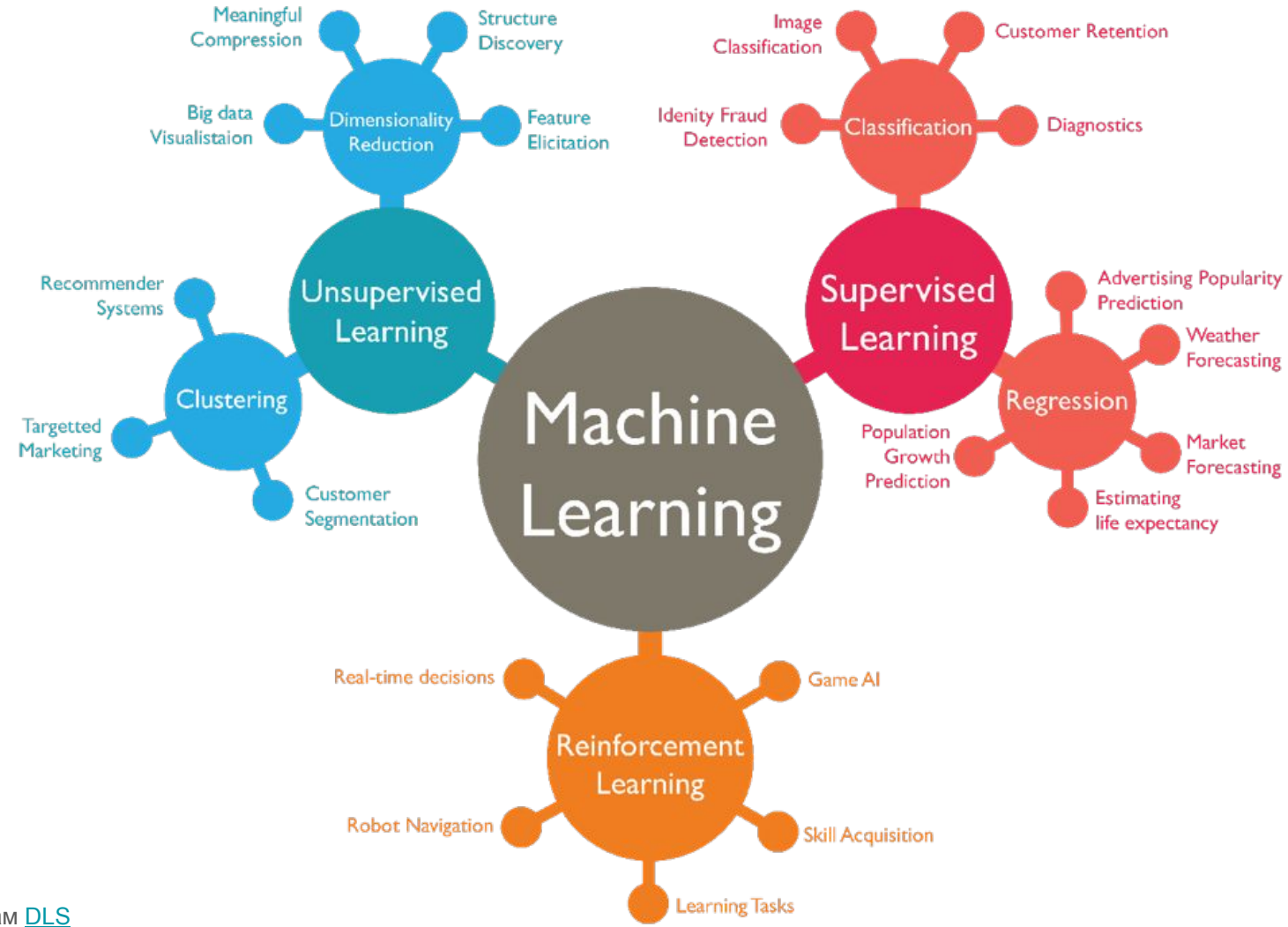
Performance	Advanced	Prescriptive
How is this product performing?	What macro-topics are more relevant for consumers?	What problems should we address first?
What are the main areas of improvements?	What problems, if fixed, generate the largest increase in star rating?	What should we do to make this product the best in its category?
What are the characteristics that customers love?	How is this product performing against its competitors?	What topics should we advertise to generate maximum return from consumers?

Ask me anything


History of NLP
TheAiEdge.io
Natural Language Processing
2022 ChatGPT
2017 Transformers
2014 Sequence to sequence learning
2013 Word2Vec
2018 BERT
2015 Attention mechanism
2011 NLP with RNN
1990 Linguistic Structure from text Statistics
1990 Statistical Parsing of Messages
2003 first Neural Language model
2008 multi-task learning
1991 Aligning Sentences in Parallel Corpora
1990 IBM alignment models
1987 Semantic Interpretation
1983 Passing Markers
1981 Plot Units
1989 tree-based model for NLP
1986 Uniform Parsing
1979 Politics
1970 augmented transition networks
1972 conceptual dependency theory
1976 TaleSpin
1978 SAM
1975 MARGIE
1977 QUALM
1978 PAM
1969 semantic networks
1968 case grammar
1964 ELIZA
1954 Georgetown experiment
1968 SHRDLU
1965 transformational generative grammars
1957 generative grammar
1949 Weaver's memorandum
Neural NLP
Symbolic NLP
Grammar theories
Statistical models
Conceptual ontologies
First models

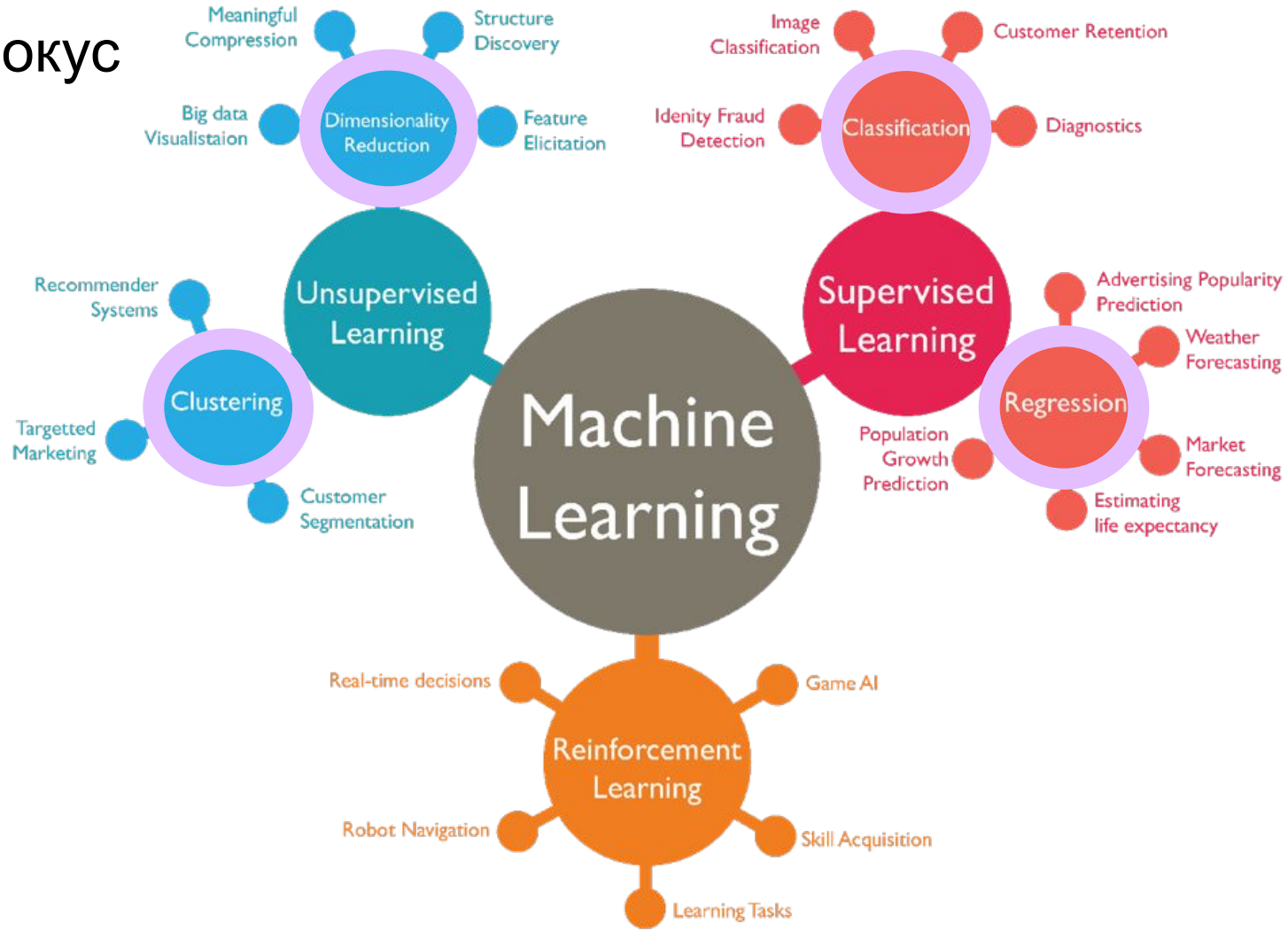
Области машинного обучения





По материалам DLS

Наш фокус


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

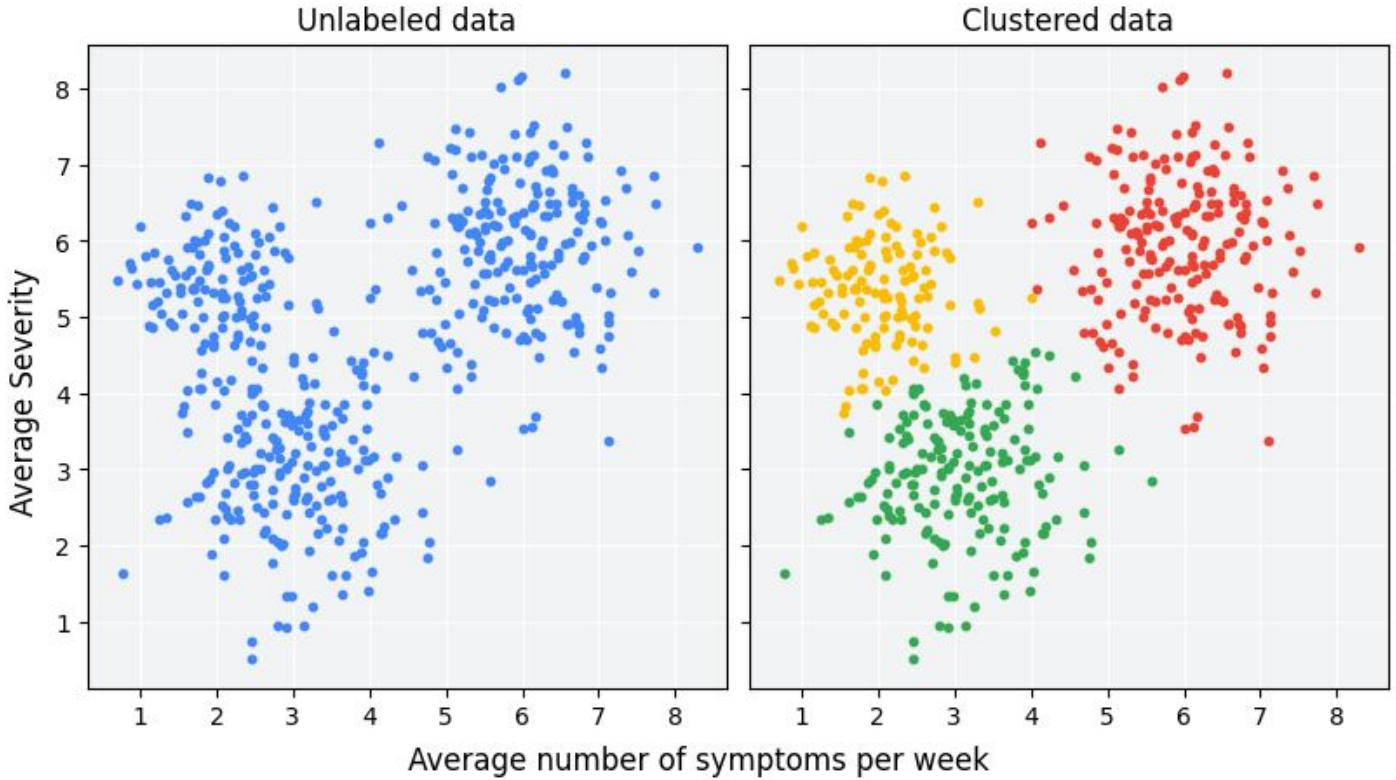
Ассоциация

«Найди какие вещи я часто ношу вместе»

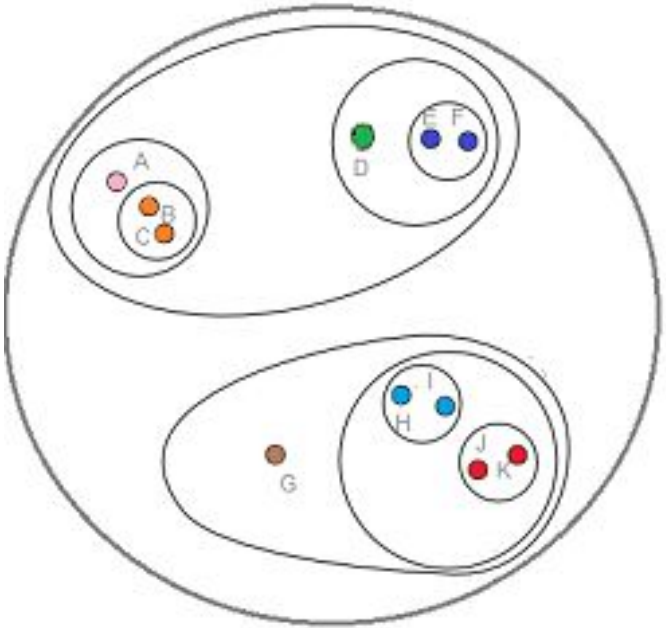
Уменьшение Размерности (обобщение)

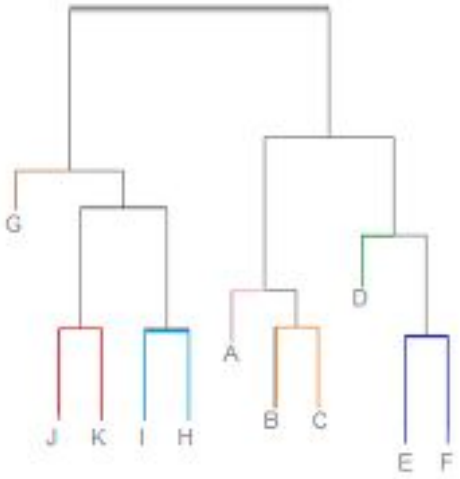
«Собери из вещей лучшие наряды»

Кластеризация

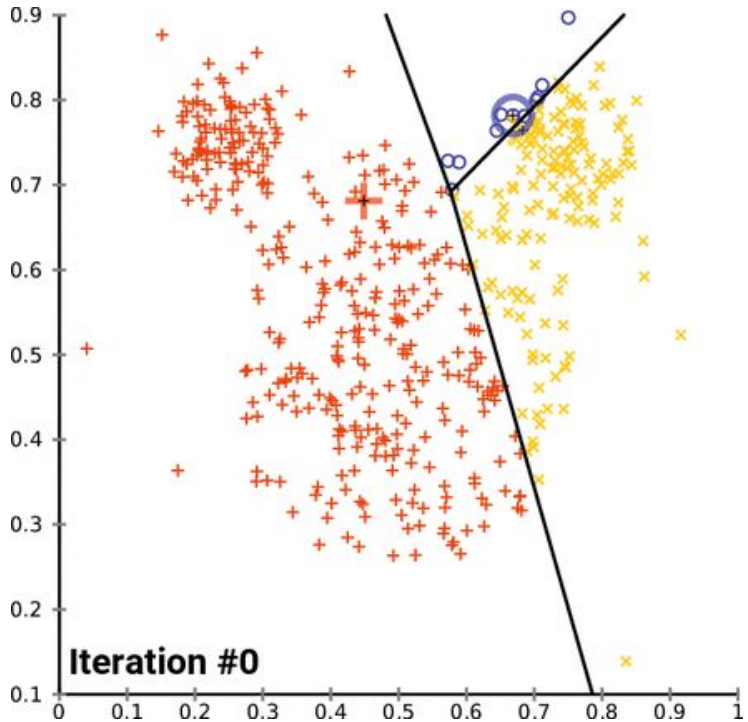


Алгоритмы кластеризации

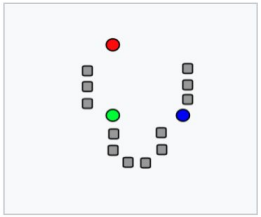




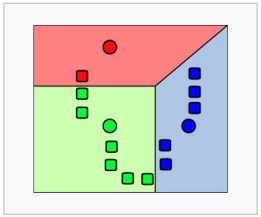
K-means



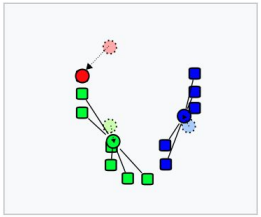
Demonstration of the standard algorithm



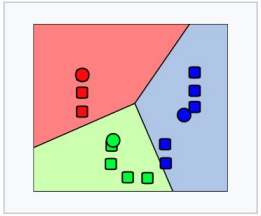
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



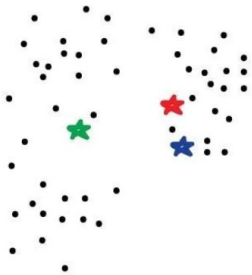
3. The centroid of each of the *k* clusters becomes the new mean.



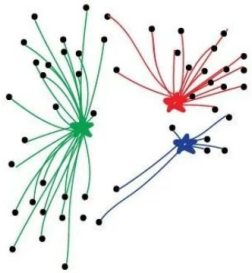
4. Steps 2 and 3 are repeated until convergence has been reached.

Ставим три ларька с шаурмой оптимальным образом

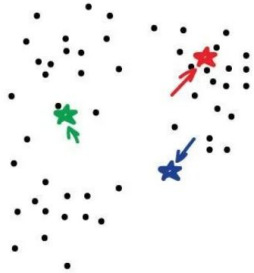
(иллюстрируя метод K-средних)



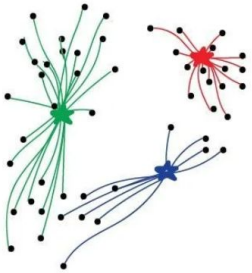
1. Ставим ларьки с шаурмой в случайных местах



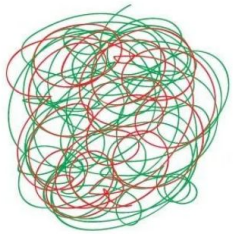
2. Смотрим в какой кому ближе идти



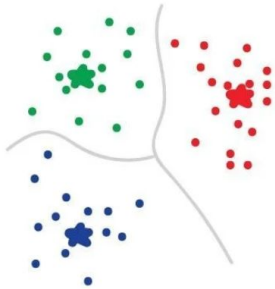
3. Двигаем ларьки ближе к центрам их популярности



4. Снова смотрим и двигаем



5. Повторяем много раз



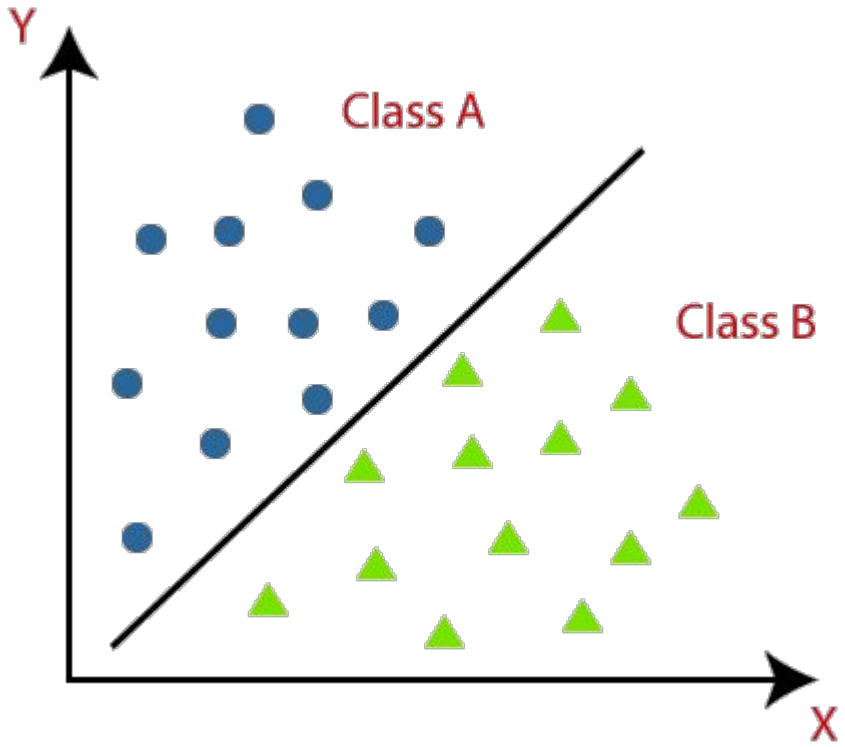
6. Готово, вы великолепны!

Примеры задач кластеризации

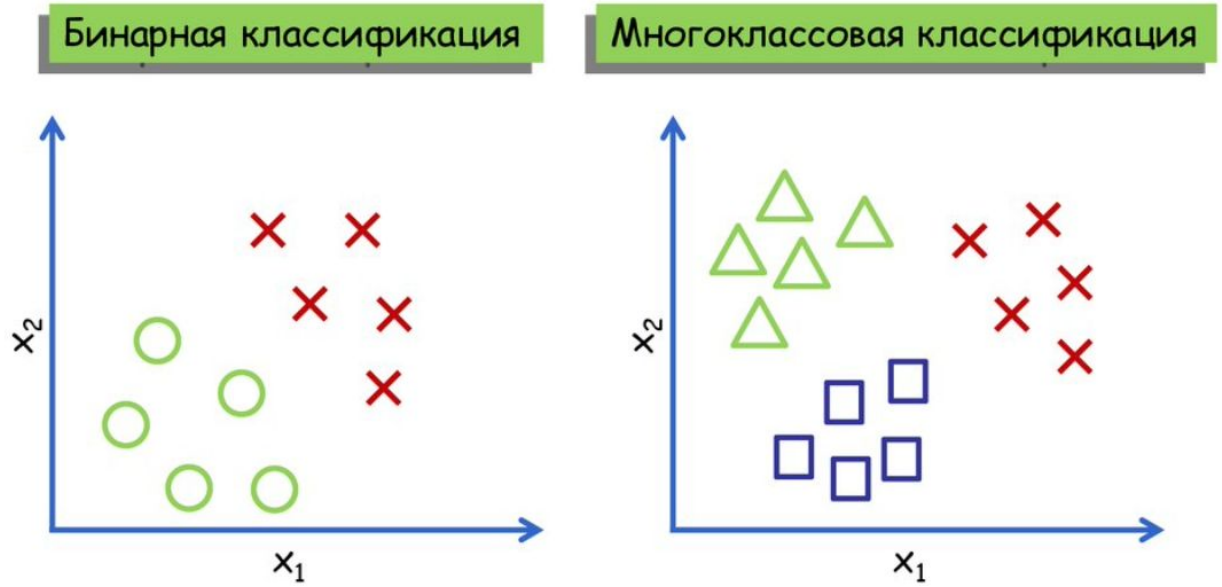
- NLP:
  - Анализ отзывов
  - Тематическое моделирование
  - Определение тональности и настроений

- Другие области:
  - Кластеризация генов или белков на основе их функций или экспрессии в геномных данных для выявления групп сходных биологических функций
  - Группировка клиентов по потребительским привычкам
  - Кластеризация пользователей по общим интересам

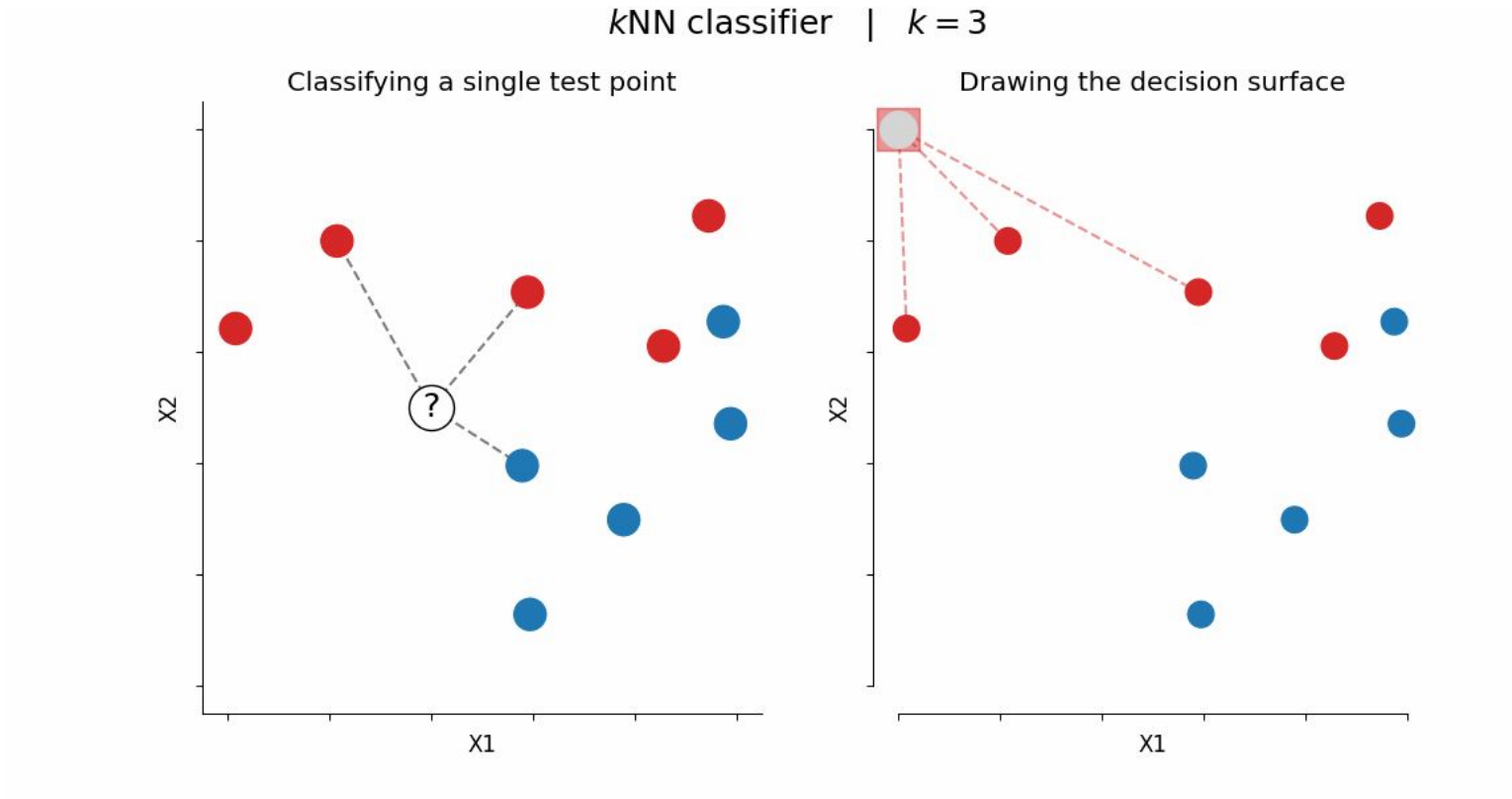
Классификация - заранее знаем классы!



Бинарная и многоклассовая



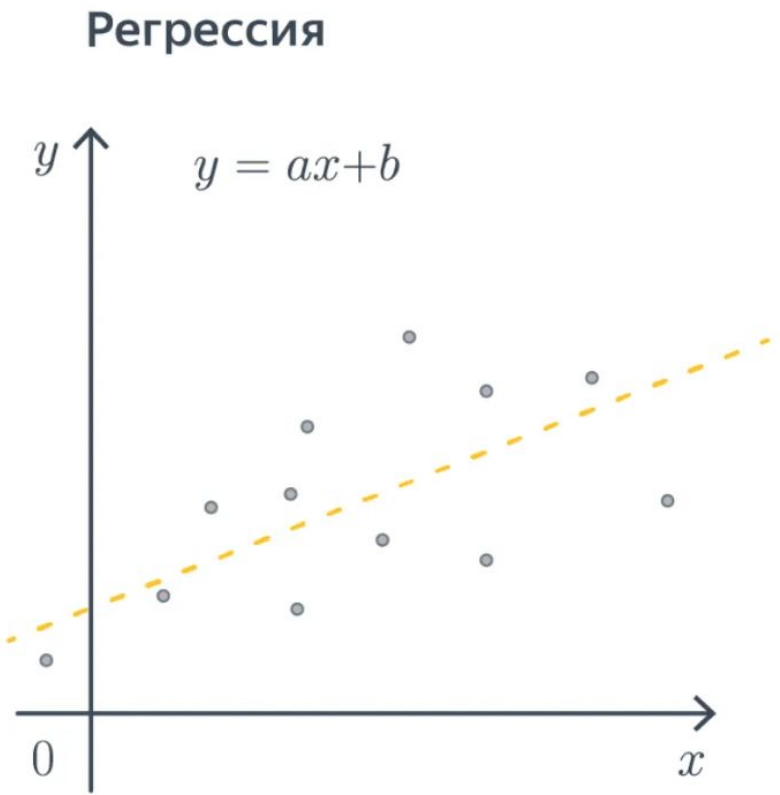
KNN


kNN classifier | k = 3
Classifying a single test point
Drawing the decision surface
X2
X1
?
X2
X1

Примеры задач классификации

- NLP:
  - Выявление фейковых и реальных новостей (бинарная классификация)
  - Токсичность
  - Спам
- Другие области:
  - Кредиты
  - Наличие заболевания

Немного про регрессию



x — (единственный) признак

y — таргет

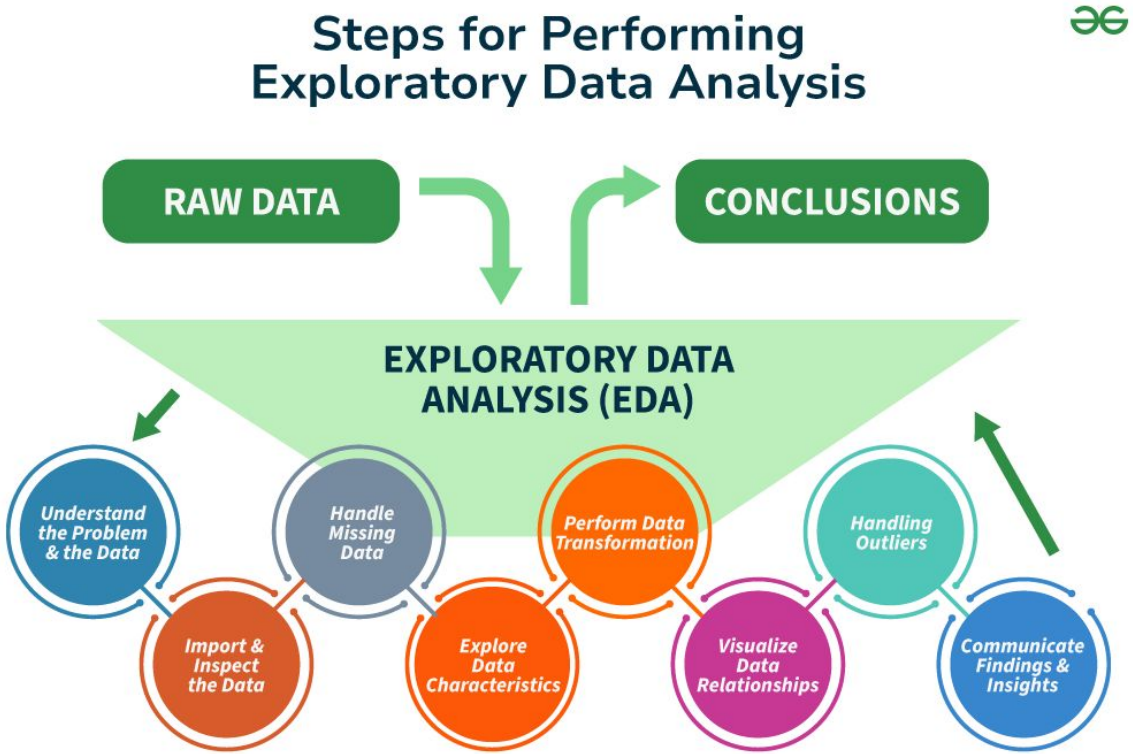
Как работать с данными?

- Выполняем предобработку, которой мы научились + векторизация
- EDA (exploratory data analysis)
  - На этом этапе важно:
    - Обработать NaN (заменить средним / медианой / dropna)
    - Обработать выбросы (слишком высокие или слишком низкие значения)
    - Проверить дубликаты (как правило их удаляют)
    - Постройте визуализации, которые помогут лучше понять ваши данные
- Разбиваем данные на train и test
- Обучение
- Валидация
- Наслаждаемся жизнью модель делает все за нас))) (если бы…)

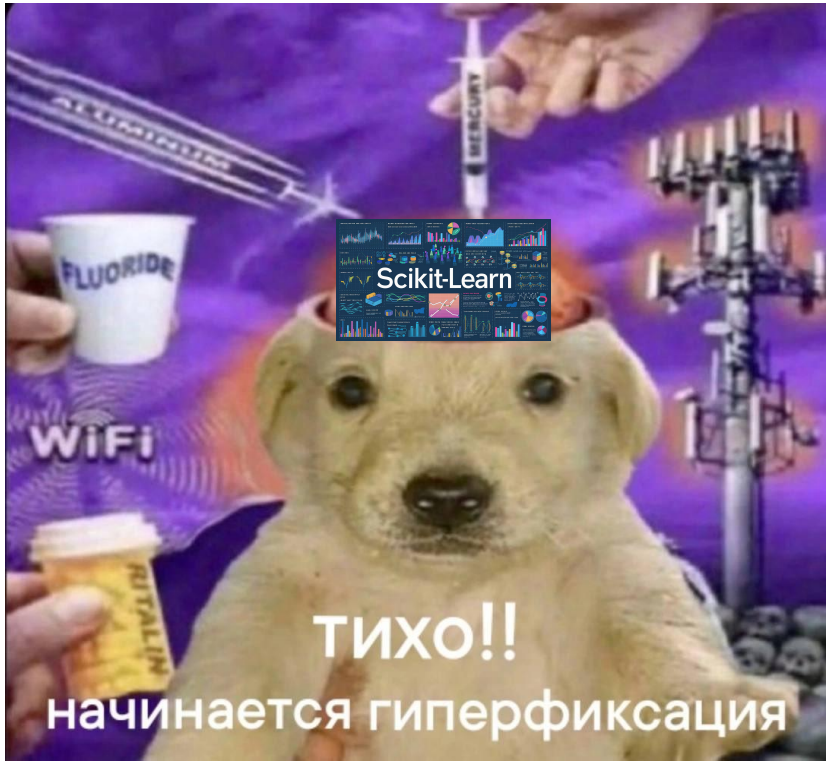
Векторизация

https://lena-voita.github.io/nlp_course/word_embeddings.html

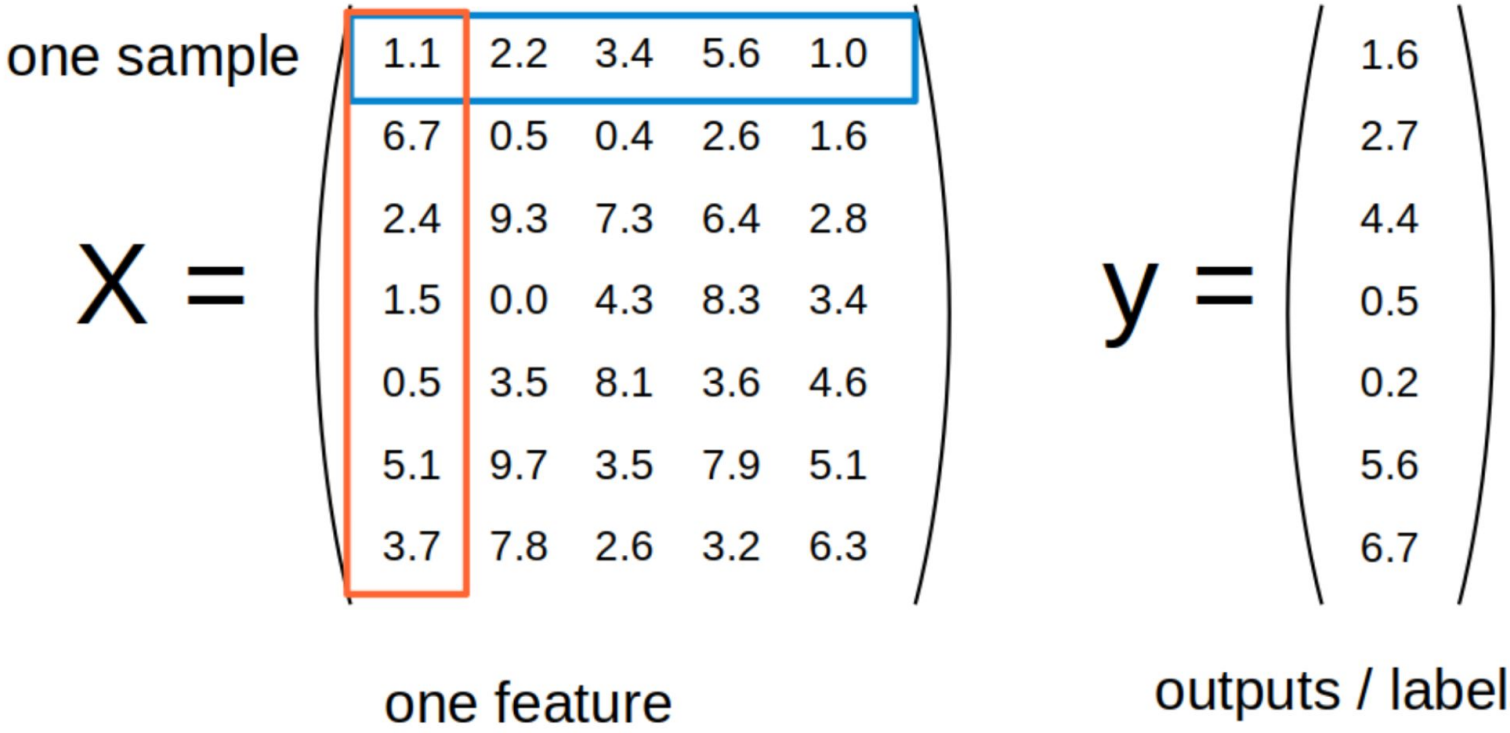
Подробнее о EDA



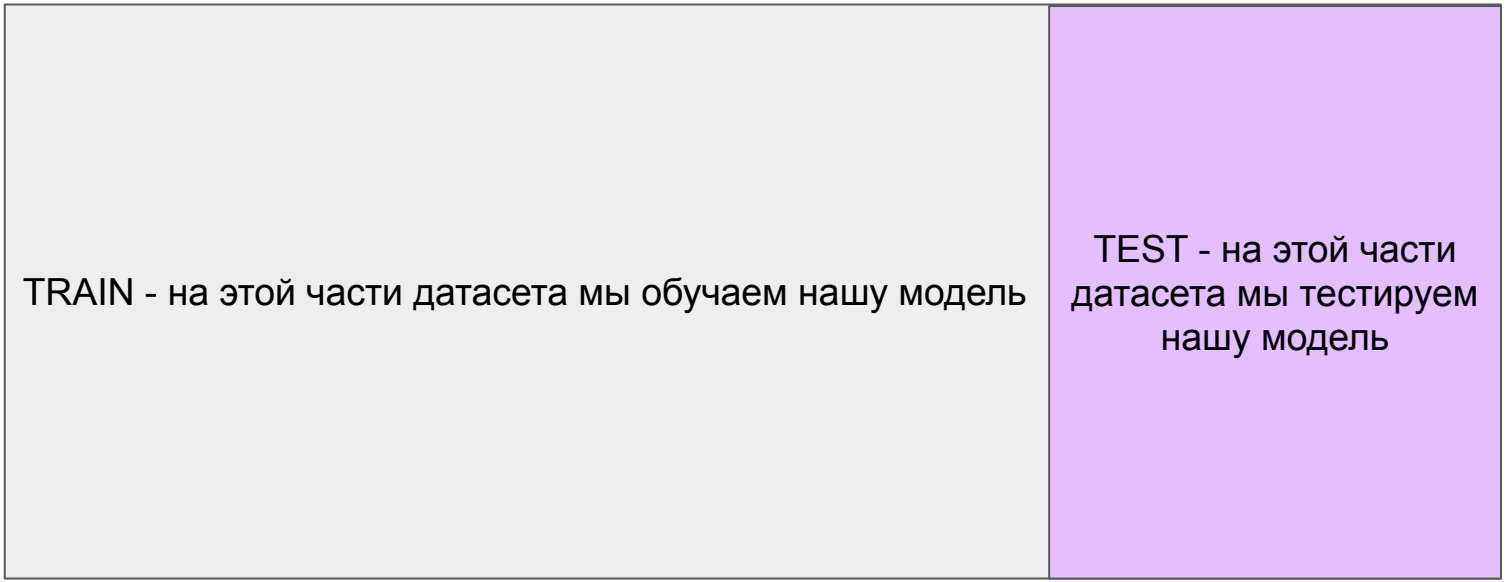
Начинаем обучение!



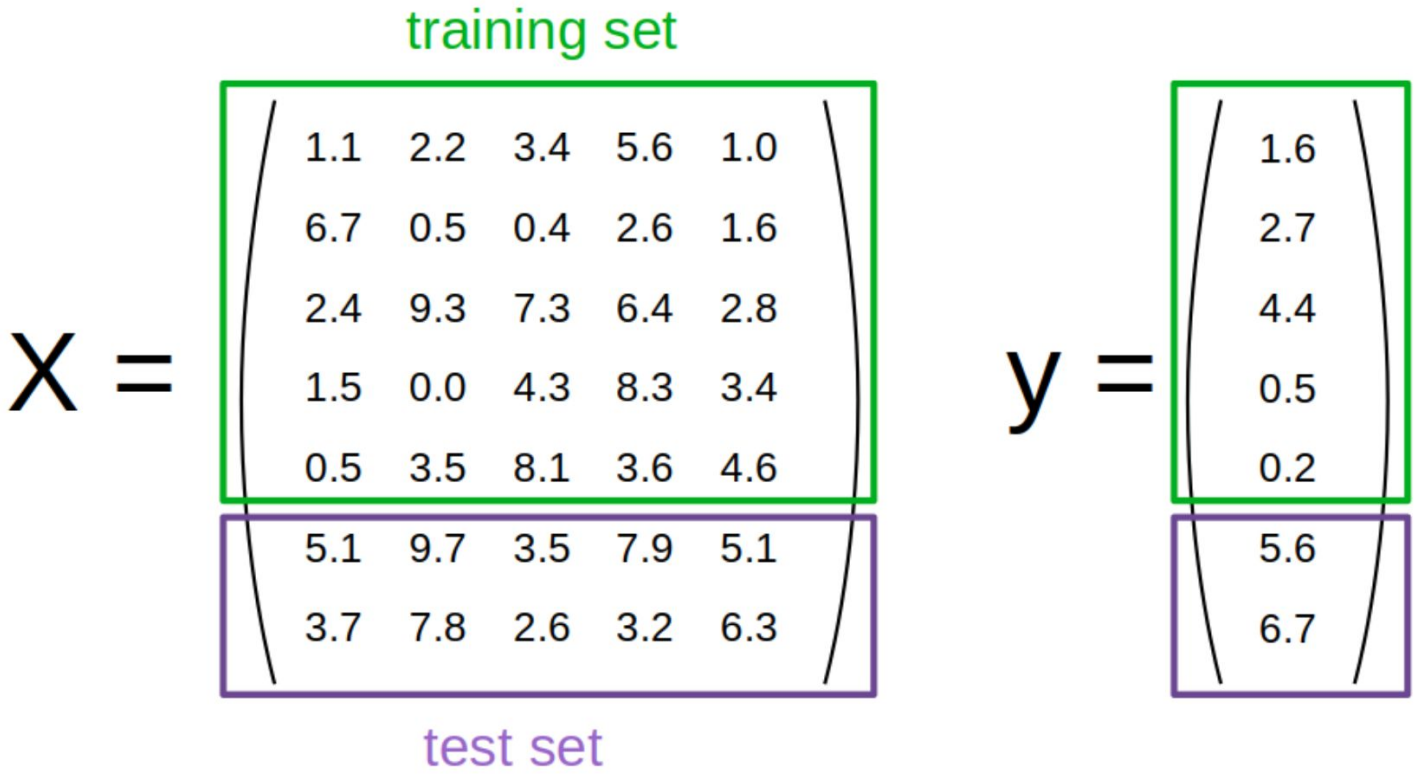
Representing Data



Train и Test - для обучения с учителем



Training and Test Data

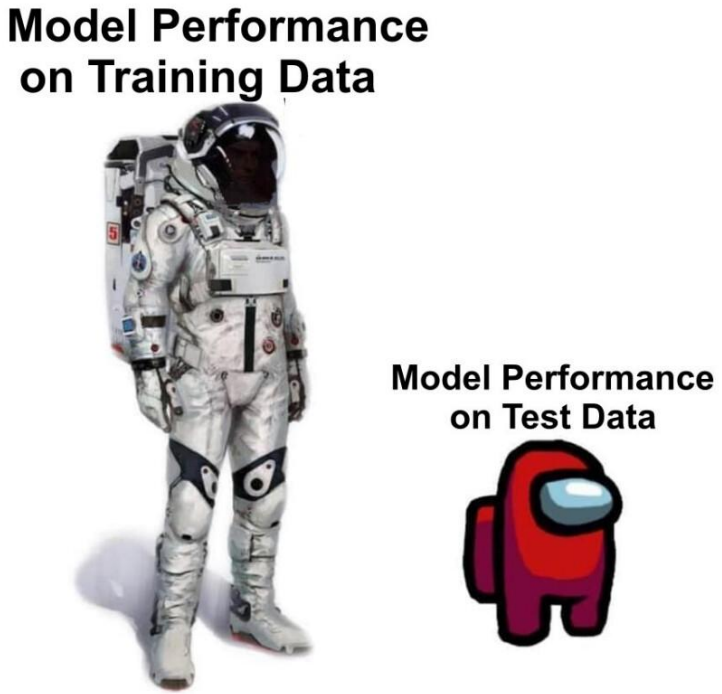


Обучаем модель

Fit - обучаем

Predict - делаем предсказания (первым делом как раз на тестовой выборке и смотрим метрики)

Тестируем нашу модель



классификация

Confusion Matrix

Actual / Predict	0	1
0	TN	FN
1	FP	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

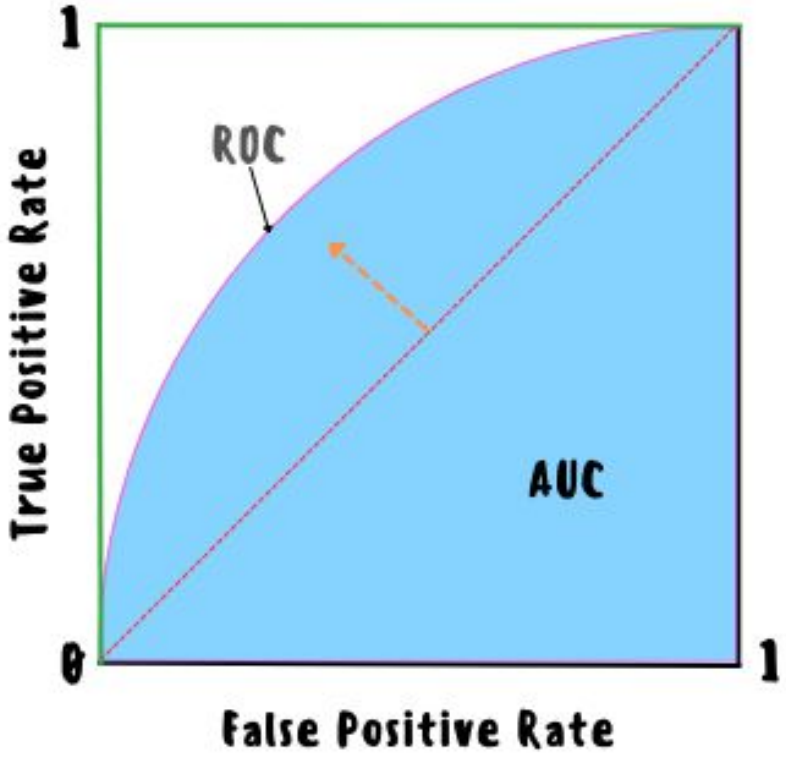
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy дает общую картину того, насколько можно полагаться на прогноз модели. Этот показатель не учитывает разницу между классами и типами ошибок. Поэтому он недостаточно хорош для несбалансированных наборов данных.

Precision показывает сколько реальных объектов класса среди всех тех, что классификатор отнес к этому классу.

Recall измеряет способность модели обнаруживать выборки, относящиеся к классу Positive.

F1-score - это среднее гармоническое значение между precision и recall.



Perfect prediction

Normal prediction

Random prediction

Better quality direction

Немного про LLM

GPT-4o: The flagship model across audio, vision, and text by OpenAI	Grok-2: Grok-2 by xAI	Claude 3.5: Claude by Anthropic
Llama 3.1: Open foundation and chat models by Meta	Gemini: Gemini by Google	Mixtral of experts: A Mixture-of-Experts model by Mistral AI
GPT-4-Turbo: GPT-4-Turbo by OpenAI	Jamba 1.5: Jamba by AI21 Labs	Gemma 2: Gemma 2 by Google
Claude: Claude by Anthropic	DeepSeek Coder v2: An advanced code model by DeepSeek	Nemotron-4 340B: Cutting-edge Open model by Nvidia
Llama 3: Open foundation and chat models by Meta	Athene-70B: A large language model by NexusFlow	Qwen Max: The Frontier Qwen Model by Alibaba
GPT-3.5: GPT-3.5-Turbo by OpenAI	Yi-Large: State-of-the-art model by 01 AI	Yi-Chat: A large language model by 01 AI
Phi-3: A capable and cost-effective small language models (SLMs) by Microsoft	Reka Core: Frontier Multimodal Language Model by Reka	Reka Flash: Multimodal model by Reka
Command-R-Plus: Command R+ by Cohere	Command R: Command R by Cohere	Qwen 1.5: The First 100B+ Model of the Qwen1.5 Series
GLM-4: Next-Gen Foundation Model by Zhipu AI	DBRX Instruct: DBRX by Databricks Mosaic AI	InternLM: A multi-language large-scale language model (LLM), developed by SHLAB.

HuggingFace

https://huggingface.co/models