Введение в АД

Весенний семестр 2023

Авторы

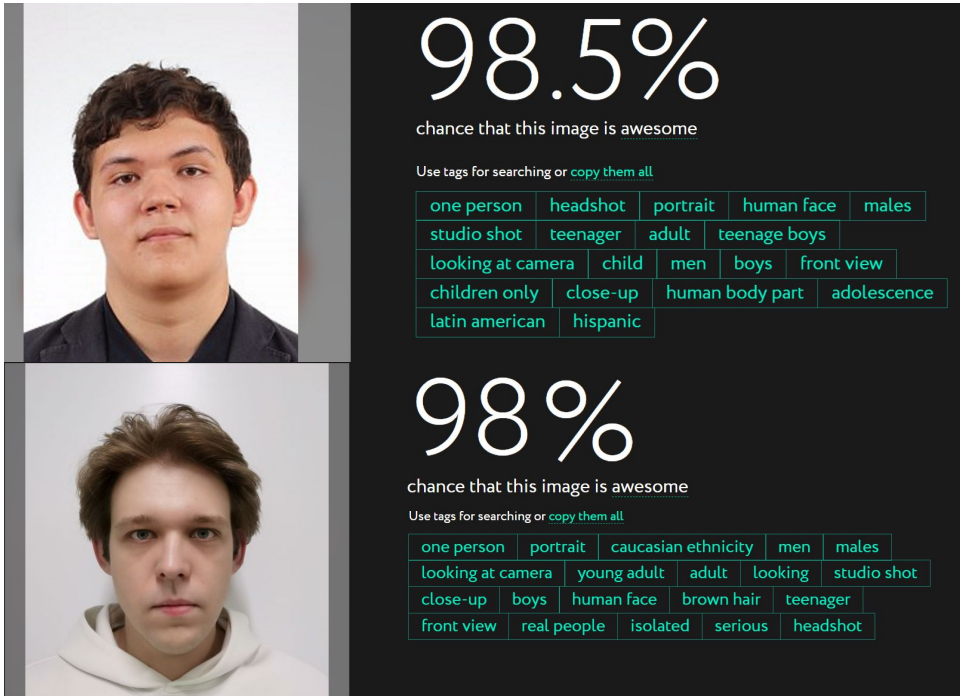
Гольдштейн Клим

- ФЭФМ’22
- м.н.с Лаборатории многомасштабного моделирования в физике мягкой материи
- ML engineer в ЦИТМ Экспонента

Шутов Григорий

- ФЭФМ->Сколтех’22,
- Research Engineer at RAIC, Skoltech
- 1st year PhD at CDSE, Skoltech

Обращаемся друг к другу на “ты”



как сделать так же

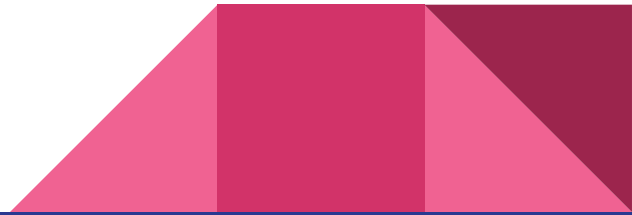
Что будем изучать

- Introduction to ML
- Supervised learning
- Unsupervised learning
- Deep Learning
- Bayesian methods/Reinforcement learning/Computational Neuroscience (на выбор, если успеем)

Что поделаем руками

Планируется 5 семинаров:

- Различаем кошек и собак за 40 строк кода 28.02
- Строим решающие леса 14.03
- Предсказываем выживаемость на титанике 28.03
- Строим языковую модель 11.04
- Определяем физтех-школу по фотографии 25.04



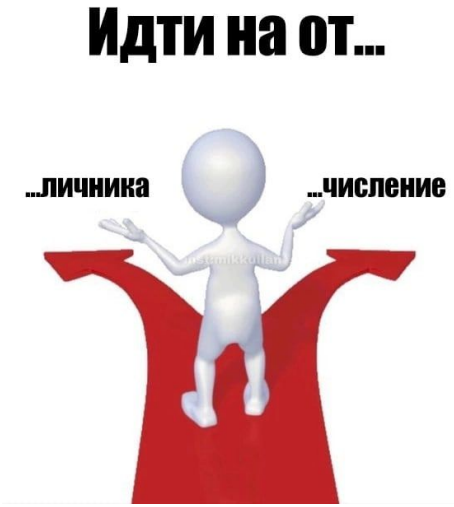
Про зачет

Варианты получить зачет:

1. Сделать проект на выбор в паре (пример – генерация и кластеризация двумерных материалов с дефектами)
2. Рассказать научную статью + сдать теорминимум на зачете

Варианты не получить зачет:

1. Прийти только на зачет с внезапным проектом/без статьи/не зная теорминимум
2. Не взять отрывной до начала зачетной недели (если берете как факультатив)



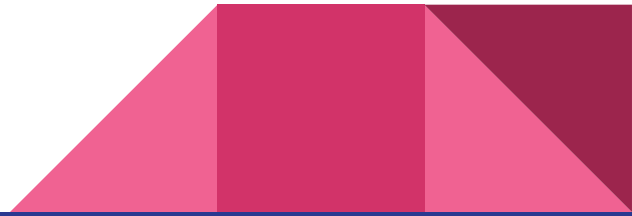
Prerequisites

1. Maths: 2 и 4 семестр матана, линал, теорвер, матстаты (WARNING: возможно будет много математики)
2. Python: numpy-pandas, algorithms
3. Linux&Github (не обязательно, но очень желательно)

Если с чем то из этого проблемы, материалы для подготовки вышлем в чат

References

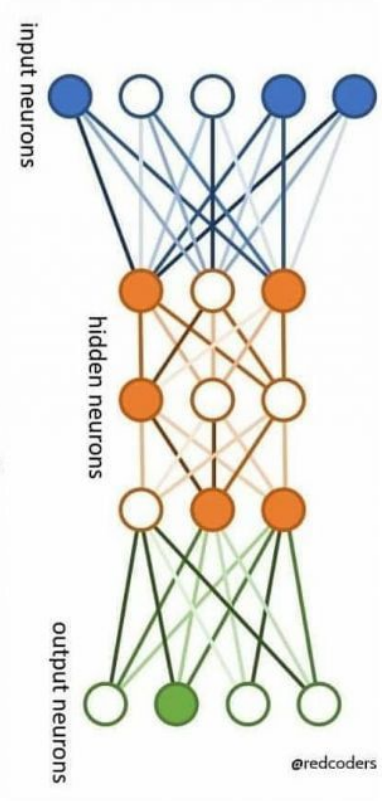
1) Машинное обучение ФПМИ (основной референс, но будут отличия)
2) Stanford CS229 (для true englishman)
3) Математические основы ML Воронцов (для любителей математики)
4) Машинное обучение ФКН (для любителей плохого звука)



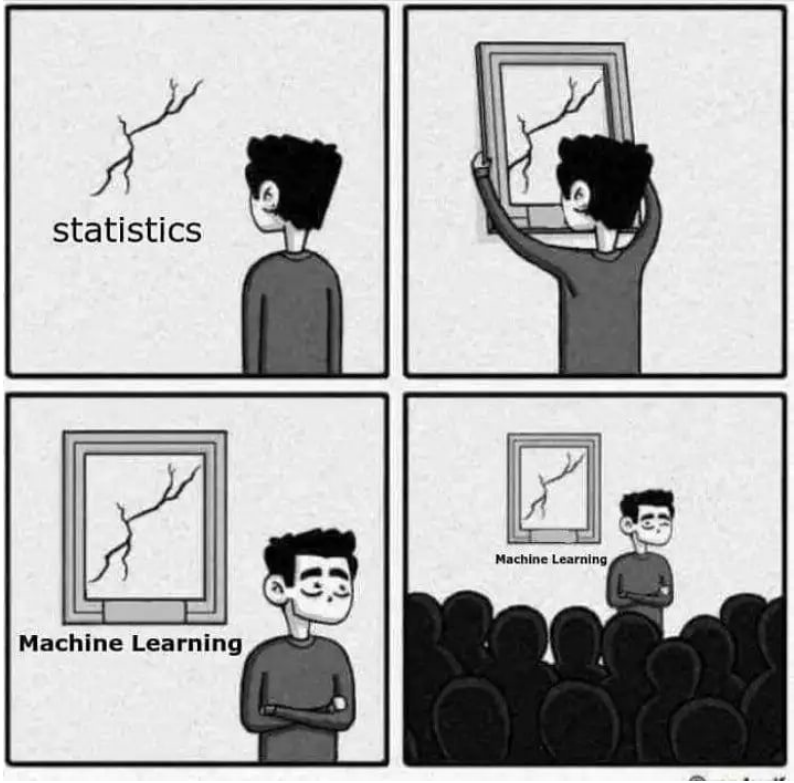
THIS IS A NEURAL NETWORK.

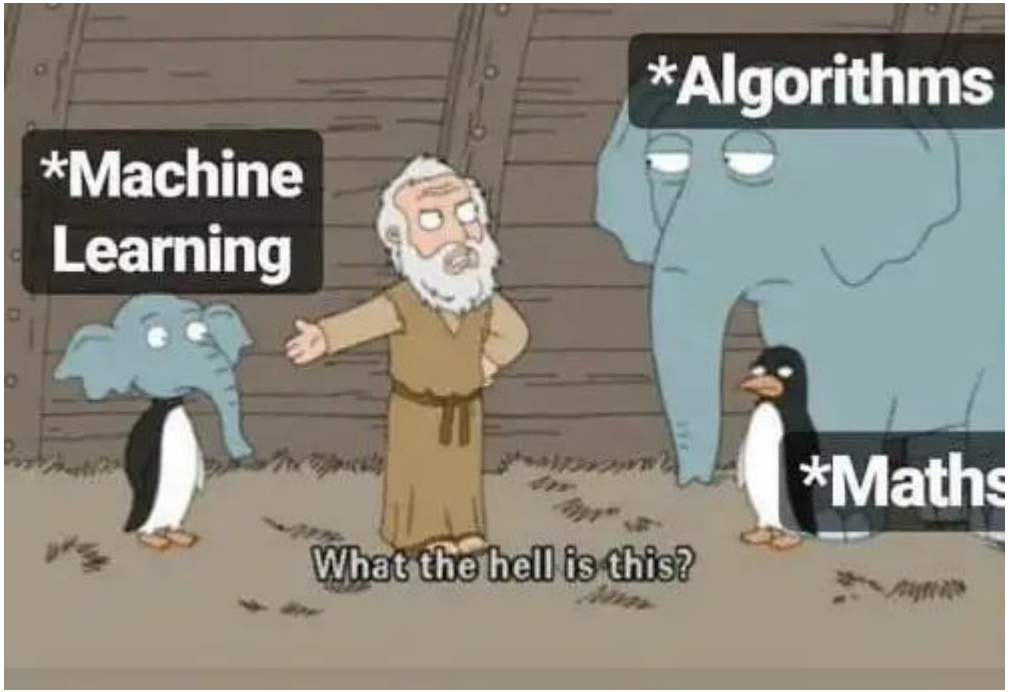
IT MAKES MISTAKES.
IT LEARNS FROM THEM.

BE LIKE A NEURAL NETWORK.



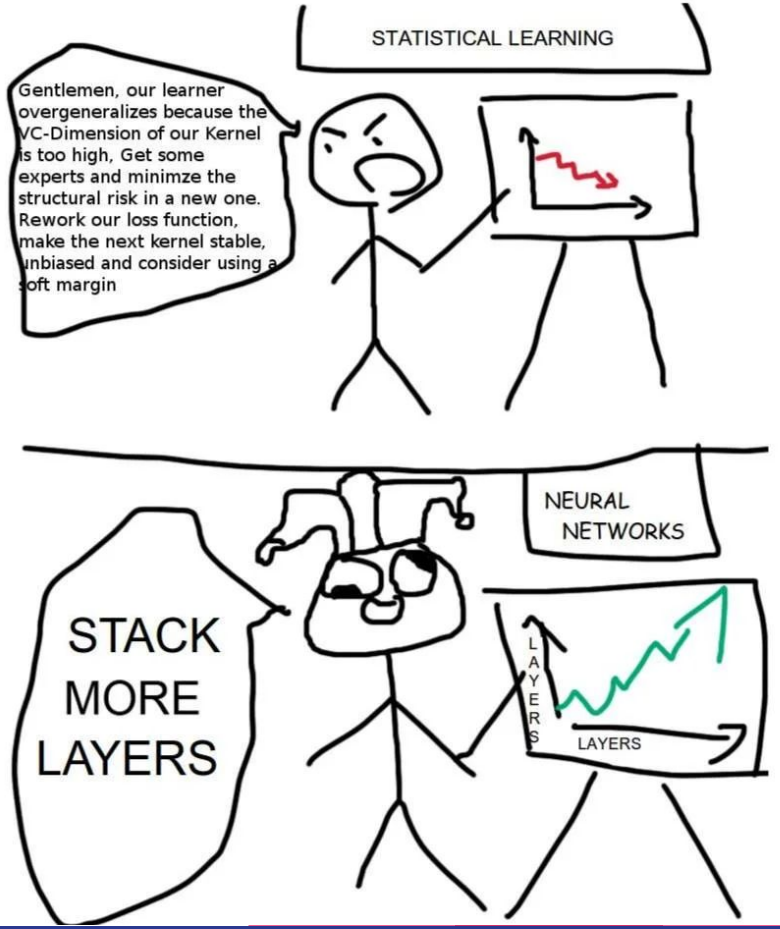
Что такое ML?


statistics
Machine Learning
Machine Learning


*Machine Learning
*Algorithms
*Maths
What the hell is this?

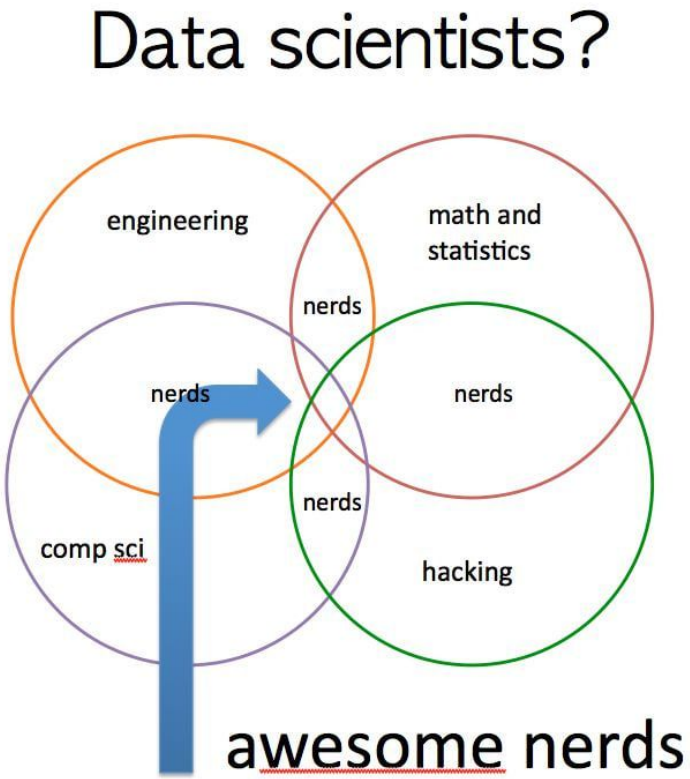
Зачем этим заниматься?

Во-первых, это весело



Во-вторых, это интересно

В-третьих, за это платят деньги
(как в компаниях, так и в науке)



Истоки

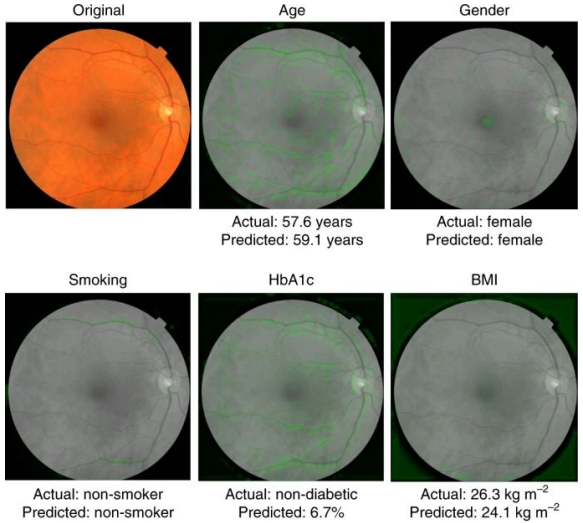
Математические истоки – теорема Колмогорова-Арнольда [1961]

$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right).$$

“Прикладные” истоки – автоматизация работы с данными (для чего был построен первый компьютер? [1944])

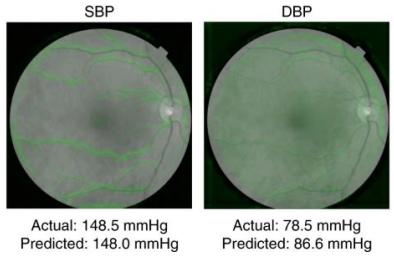
Почему такое развитие получилось только сейчас?

Магия ML



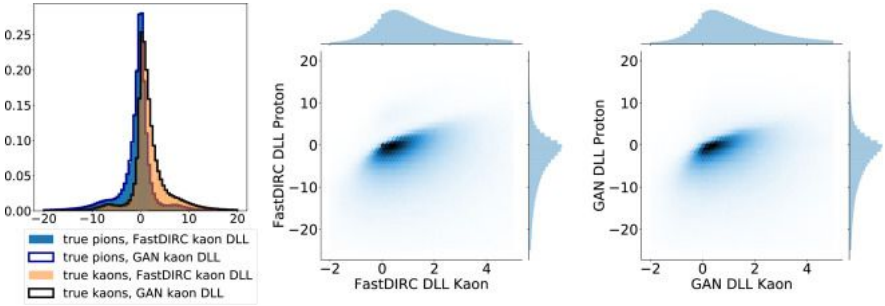
DSP-1181

Потенциальное лекарство, найденное с помощью AI



Набор моделей,

предсказывающих диабет, ИМТ, глазное давление, риск ССД по фотографии сетчатки глаза

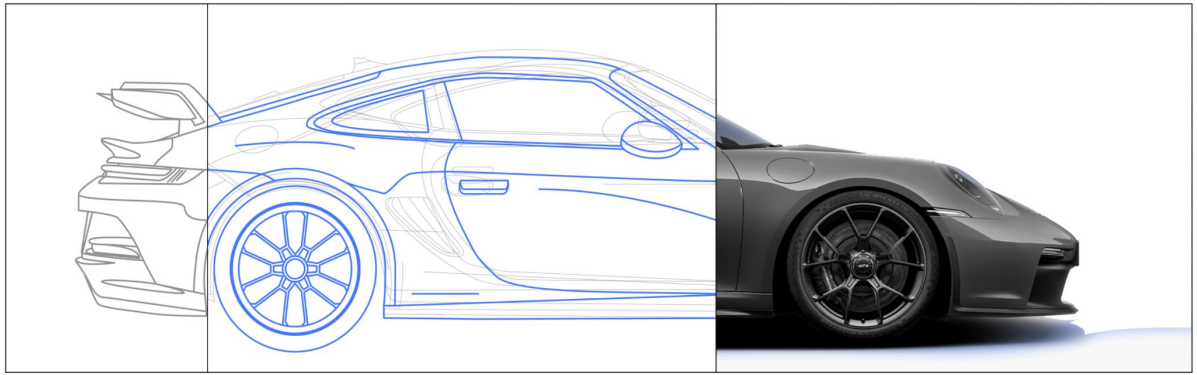


Генеративная сеть,

для симуляции Черенковского детектора
Используется на БАК

Инженерное проектирование

осуществляемое с помощью глубоких нейронных сетей



Эволюционный алгоритм,

предсказывающий кристаллические структуры, молекулы, 2D кристаллы, поверхности, молекулярные кристаллы и MOFы

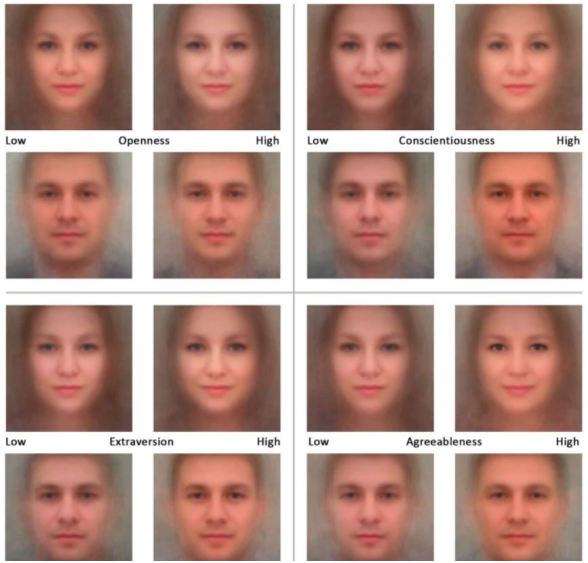
Нейросеть, которая упрощает решение уравнений квантовой механики

One day Joe Bear was hungry. He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe walked to the oak tree. He ate the beehive. The End.

ChatGPT, которая могла бы написать текст для этой презентации

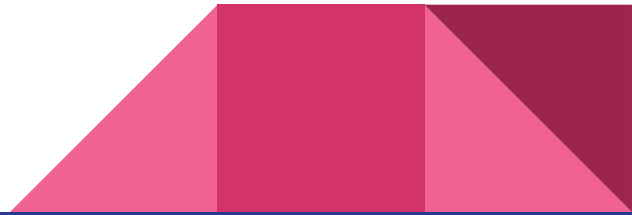
Предсказание Big Five personality traits по фотографии

Figure 1



Thesaurus (необходимые понятия)

- Признаки x, таргет y (features, target)
- модель f (model)
- предсказание $\hat{y} = f(x)$ (prediction)
- функция ошибки $L(y, \hat{y})$ (Loss function)
- выборка S (sample)
- гиперпараметры (e.g. число кластеров)
- правдоподобие $L(\theta, x)$ (Likelihood)



Problem statement

Supervised learning:

- Training set $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where
  - $(\mathbf{x} \in \mathbb{R}^p, y \in \mathbb{R})$ for regression
  - $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{+1, -1\}$ for binary classification
- Model $f(\mathbf{x})$ predicts some value for every object
- Loss function $Q(\mathbf{x}, y, f)$ that should be minimized

- Regression
- Classification
- Prognosis

Unsupervised learning:

- Clusterization
- Anomaly detection
- Dimensionality reduction

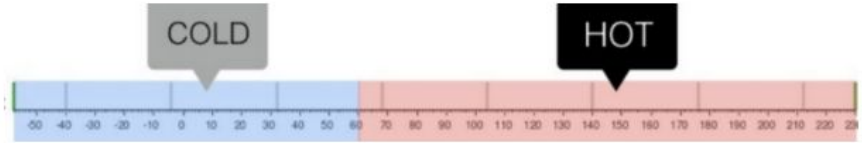
Regression

What will be the temperature tomorrow?

Fahrenheit

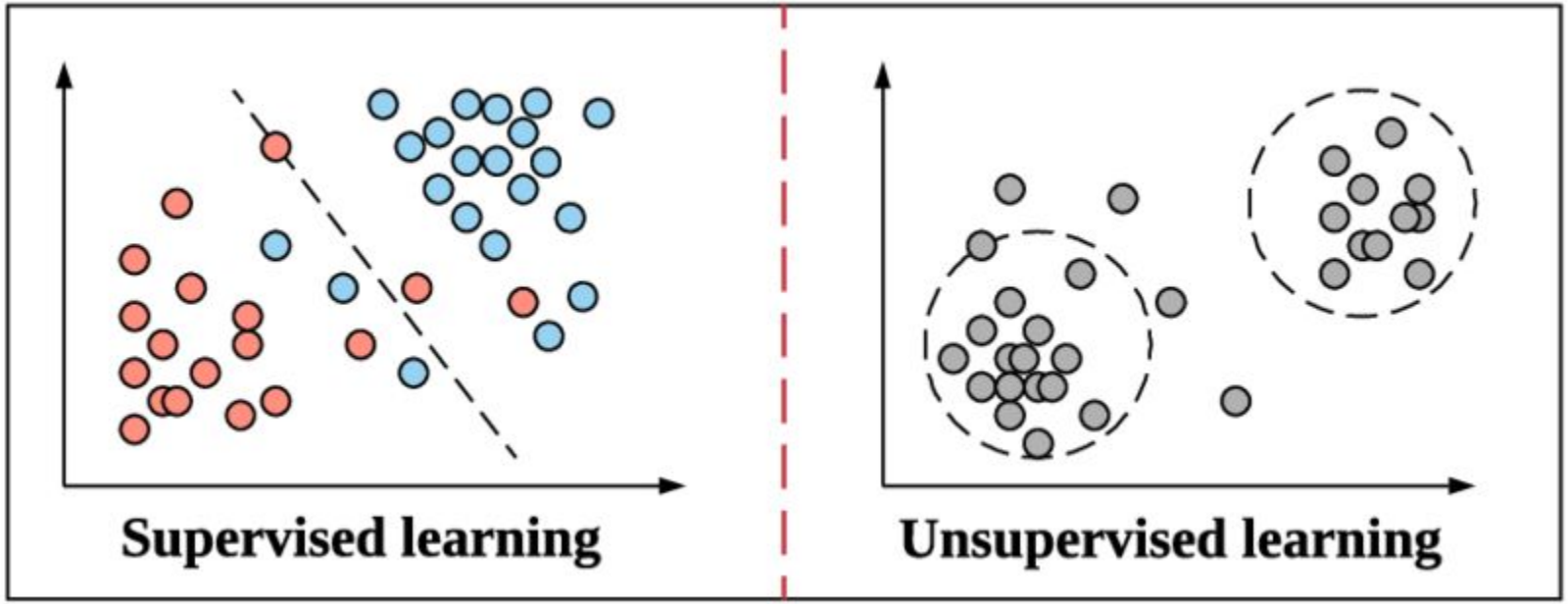
Classification

Will it be hot or cold tomorrow?

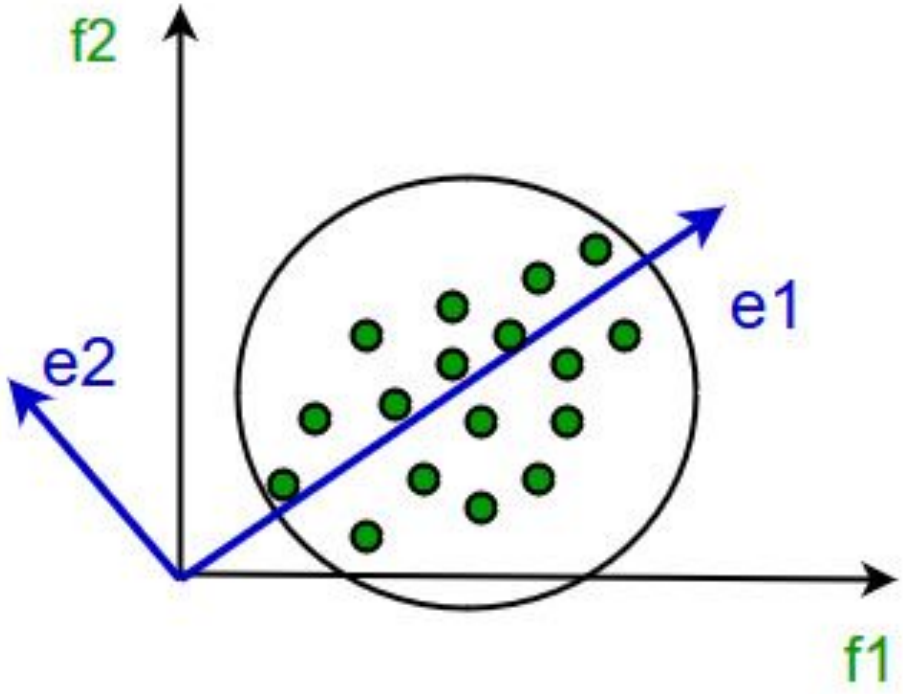


Fahrenheit

Не имеем изначальных классов, группируем точки по “похожести”



Имеем данные по принадлежности к красному или синему классу – предсказываем, что если новая точка слева от прямой, то она красная, если слева, то синяя



В этом случае, мы можем заменить представление точек в виде пары координат положением их проекции на e1, таким образом мы сократим размерность и сохраним наибольшее количество информации