Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра “Прикладная математика”

Использование *Transfer Learning* *для выявления вредных мутаций в белках млекопитающих*

Студент: Плеханова Е.С.
Научный руководитель: Самсонова М.Г.

Transfer learning. Формальное определение

Пусть $\mathcal{D}$ — данные, $\mathcal{T}$ — задача. Индексом ***S*** будем помечать исходные данные, индексом ***T*** — целевые данные.

$\mathcal{D} = \{\mathcal{X}, P(X)\}$, где $\mathcal{X}$ — пространство признаков, $P(X)$ — их распределения.

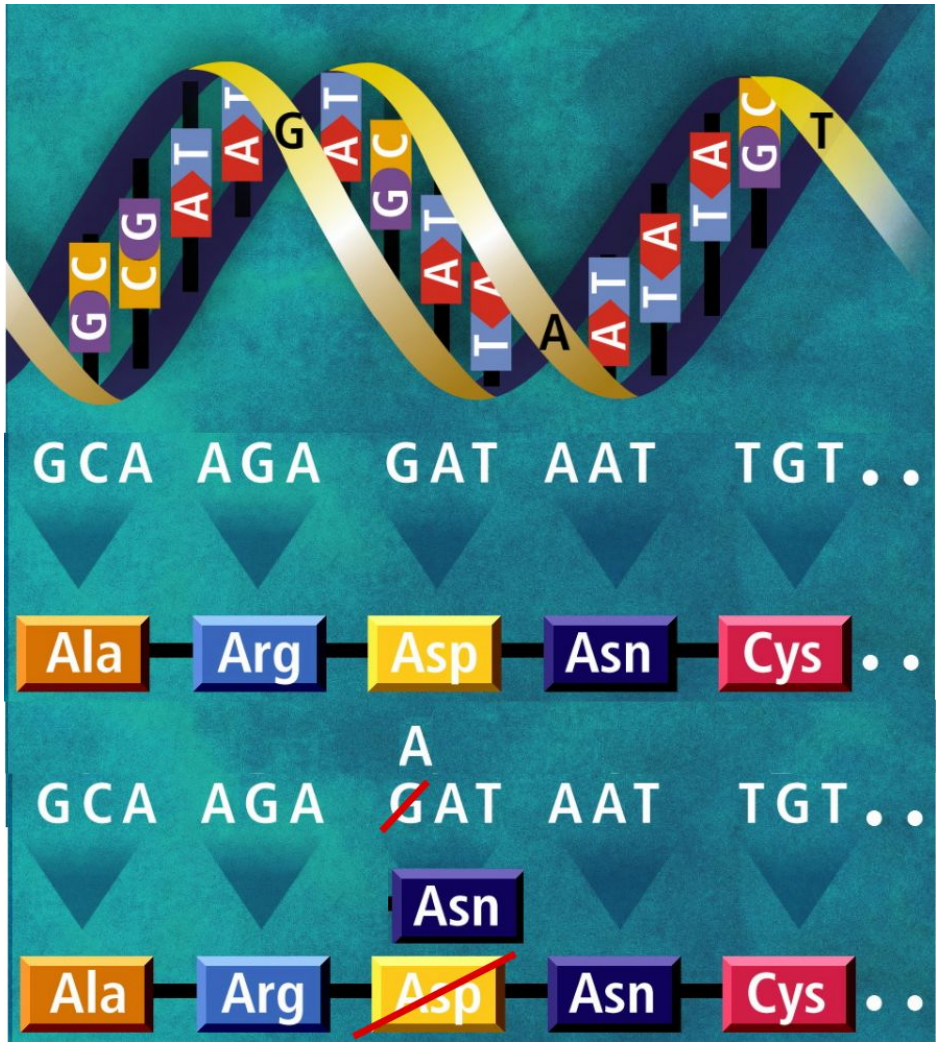
$\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, где $\mathcal{Y}$ — пространство ответов, $f(\cdot)$ — функции предсказания.

Transfer learning направлен на улучшение функции предсказания $f_T(\cdot)$ на данных $\mathcal{D}_T$, используя знания в $\mathcal{D}_S$ и $\mathcal{T}_S$,

причем выполнено хотя бы одно:

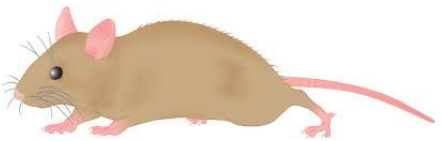
$$\begin{bmatrix} \mathcal{X}_S \neq \mathcal{X}_T \\ \mathcal{Y}_S \neq \mathcal{Y}_T \\ \mathrm{P}(\mathrm{X_S}) \neq \mathrm{P}(\mathrm{X_T}) \\ \mathrm{P}(\mathrm{Y_S}|\mathrm{X_S}) \neq \mathrm{P}(\mathrm{Y_T}|\mathrm{X_T}) \end{bmatrix}$$

Вредные мутации



— мутации, приводящие к изменению в белке и ассоциированные с заболеванием.

Остальные — нейтральные.



Постановка задачи

Цель работы: применение методов Transfer Learning к задаче классификации вредных мутаций у мышей и собак.

(Исходные данные - человек, целевые - мышь, собака)

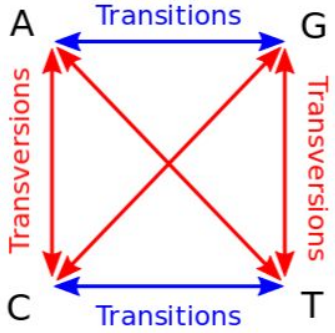
Задачи:

- Подготовить **наборы мутаций** и извлечь признаки для классификации
- Обучить **классификаторы** на данных по человеку, выбрать наилучшие классификаторы.
- Применить методы Transfer Learning к полученным **наборам мутаций** с использованием **классификаторов**.

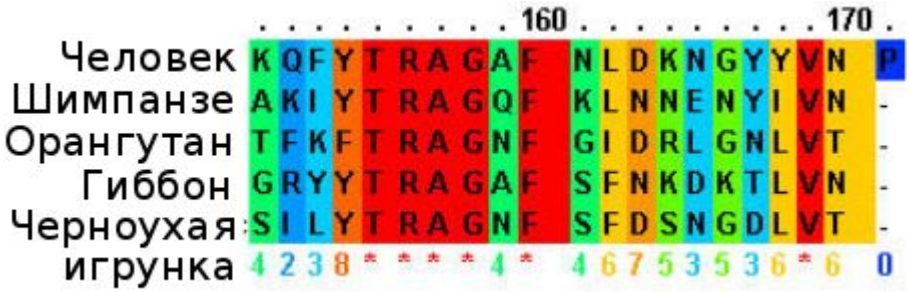
Набор признаков

С помощью программы Polyphen наборы признаков получены из:

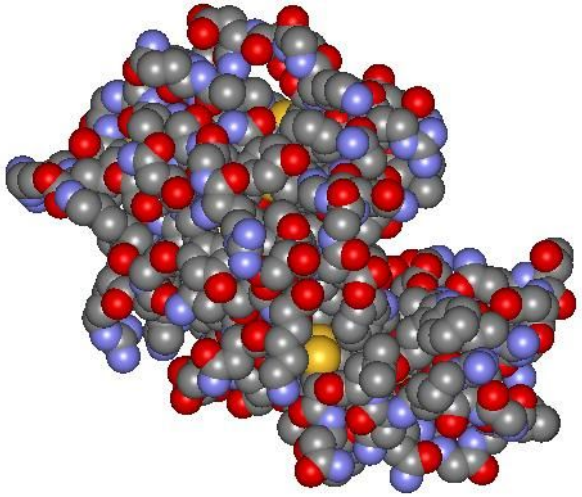
- Контекста последовательности



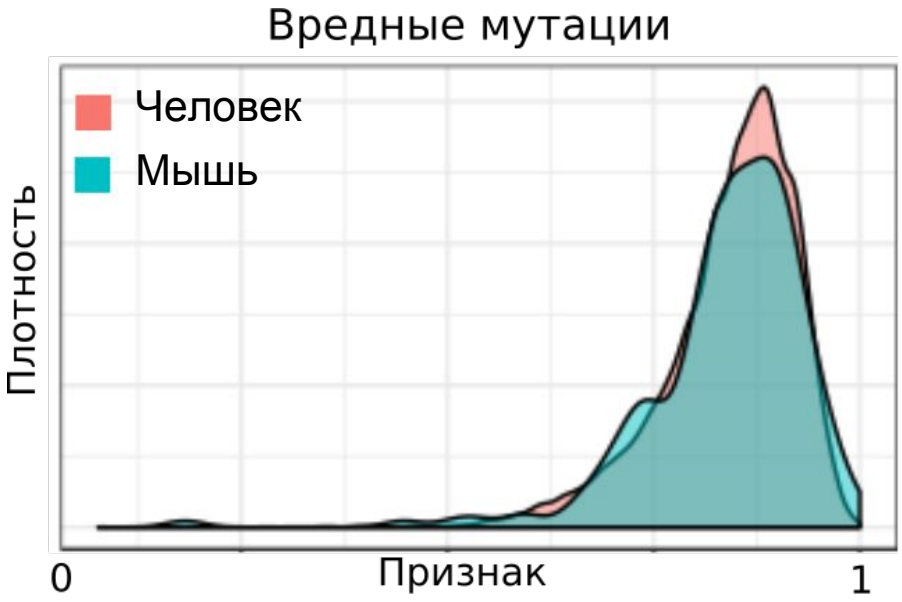
- Множественного выравнивания

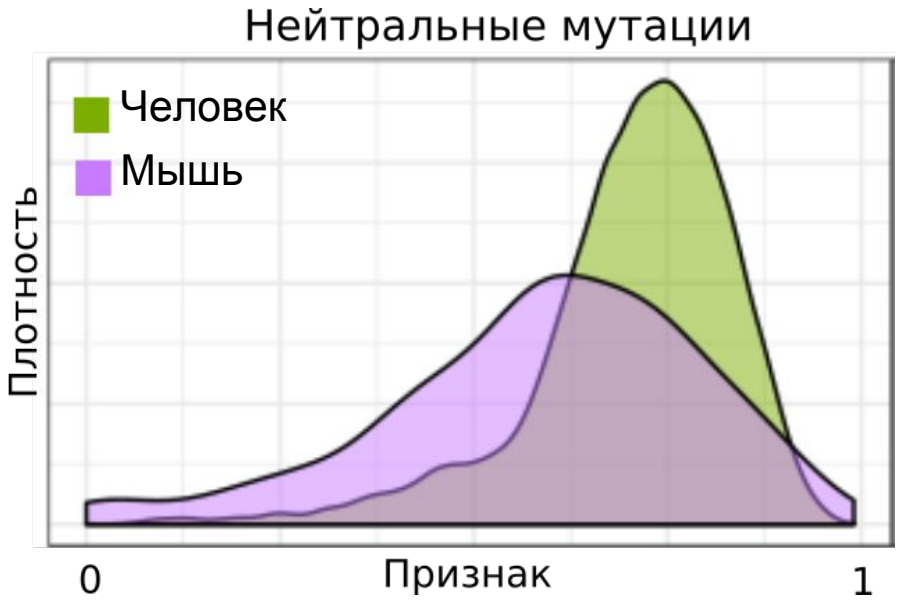


- 3D структуры белка:



Распределения исходных и целевых данных отличаются





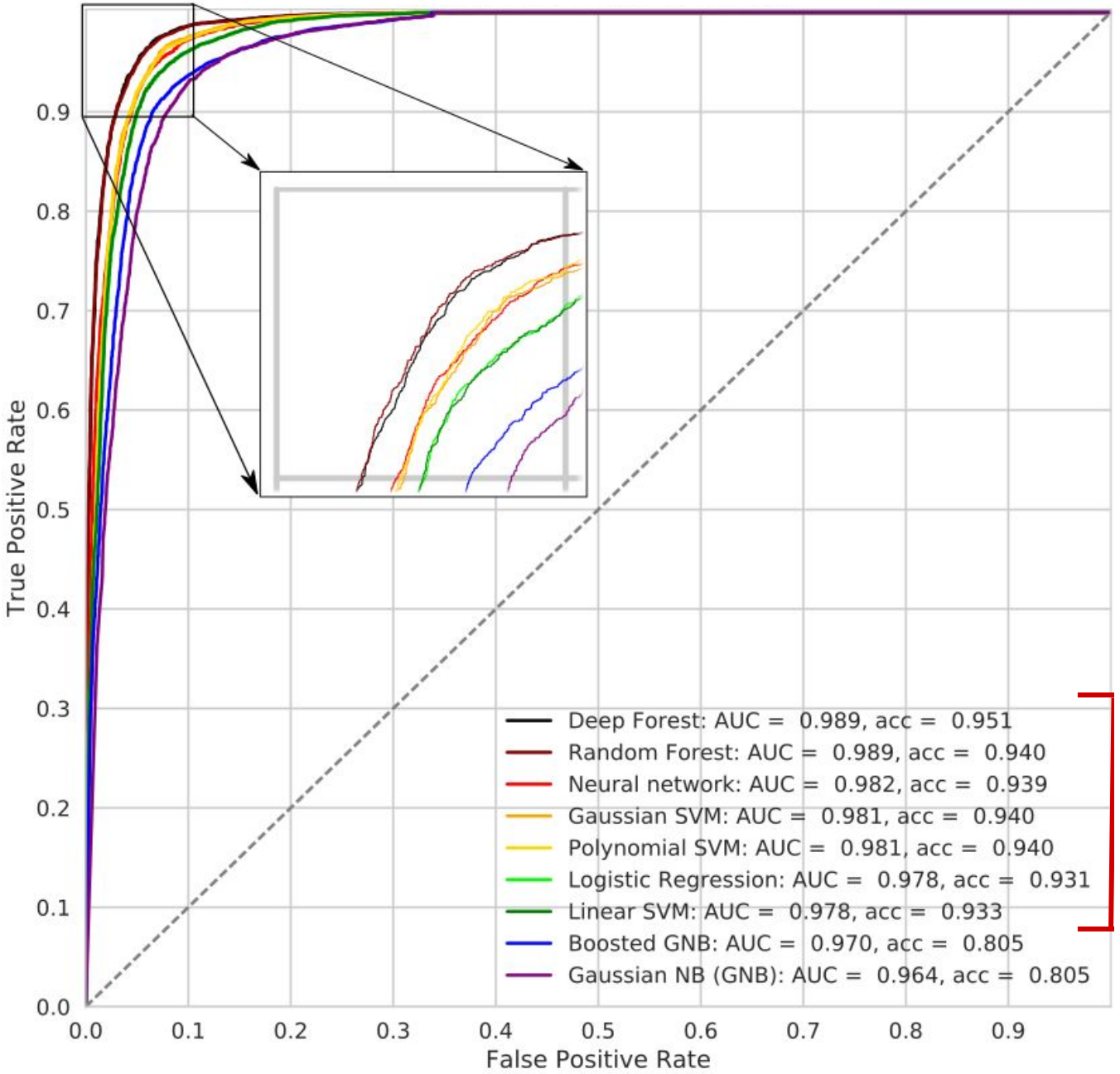
Используемые классификаторы

- Naive Bayes (Наивный байесовский классификатор)
- Boosted Naive Bayes (Бустинг наивных байесовских классификаторов)
- Logistic regression (Логистическая регрессия)
- SVM: Linear, Gaussian, Polynomial (Машина опорных векторов с различными ядерными функциями)
- Random forest (Случайный лес)
- Neural Network (Нейронная сеть)
- Deep Forest (Глубокий лес)

<u>Наборы данных по человеку:</u>

- **HumDiv.** Вредные мутации — ассоциированные с болезнью
- **HumVar**. Вредные — ассоциированные с изменением функции белка.

Roc-кривые на данных по человеку (HumDiv)

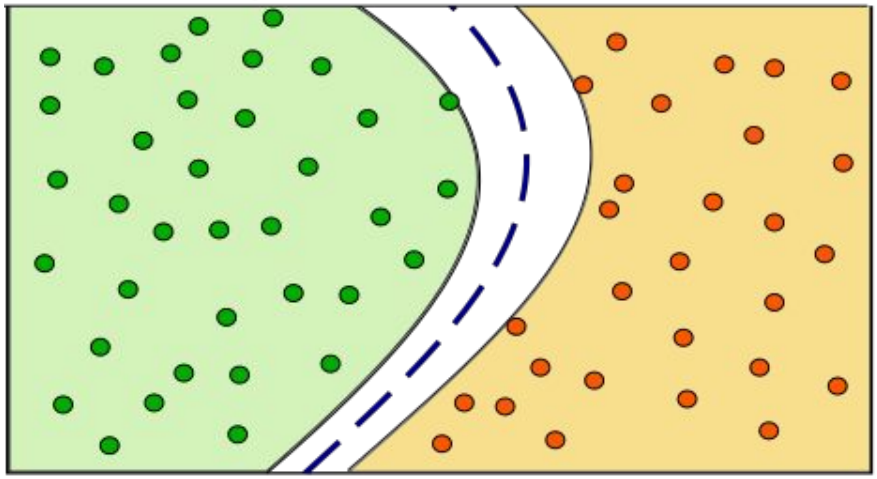


Методы Transfer learning

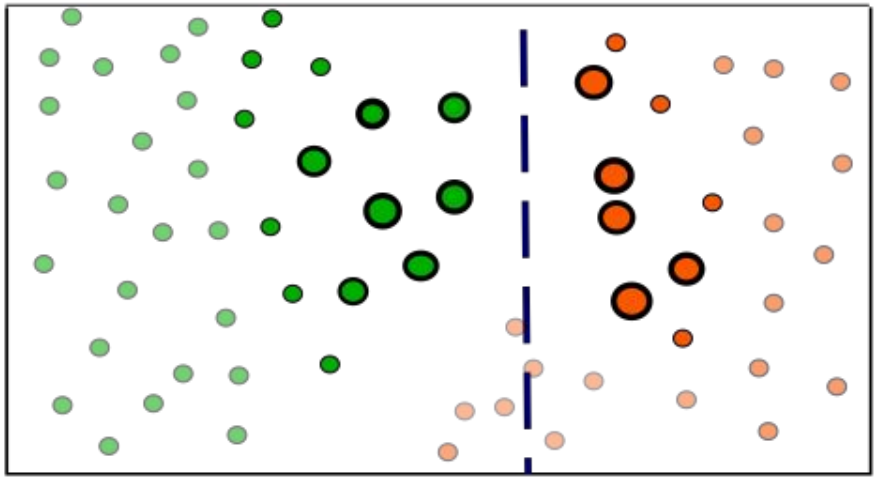
1. Перенос образцов
2. Новое признаковое пространство

1. Перенос образцов

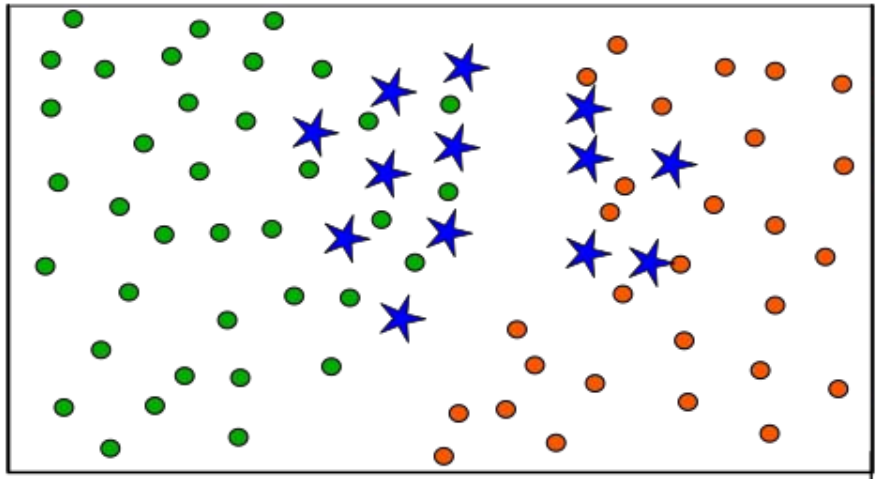
1. Классификация исходных данных



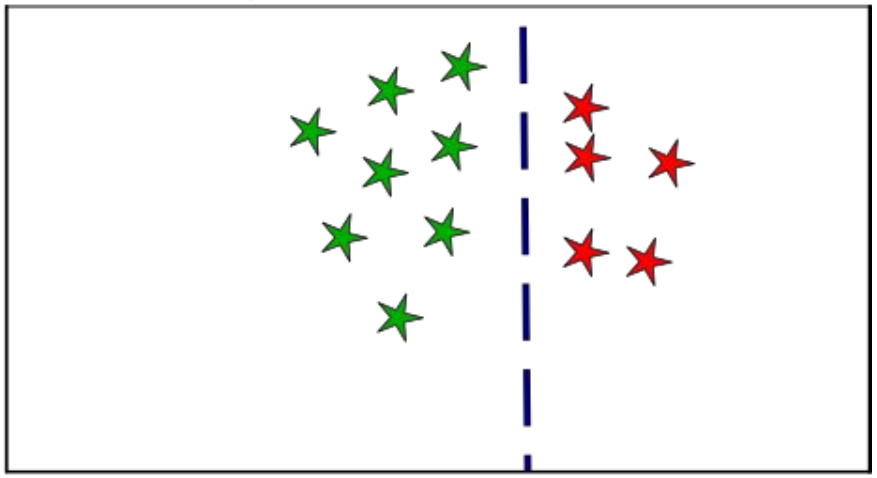
3. Чем ближе данные, тем больше веса



2. Исходные и целевые данные



4. Классификация целевых данных



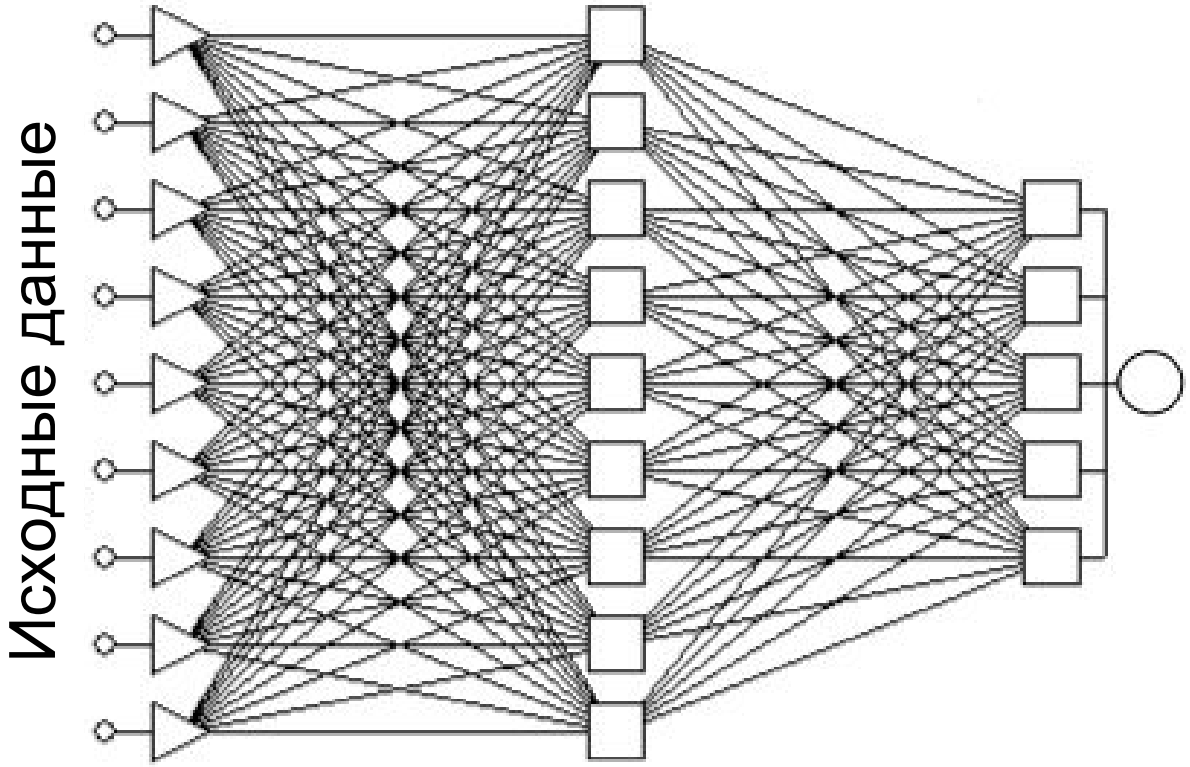
1. Перенос образцов. Результаты

Значение точности классификации с использованием Transfer learning (TL) и без.

Классификатор	**Собака**				**Мышь**			
	Обучение на **HumDiv**		Обучение на **HumVar**		Обучение на **HumDiv**		Обучение на **HumVar**	
	+TL	—	**+TL**	—	**+TL**	—	**+TL**	—
Random Forest	0.855	0.638	**0.889**	0.884	**0.846**	0.682	0.841	0.682
Polynomial SVM	0.874	0.657	0.715	0.454	0.764	0.764	0.812	0.528
Gaussian SVM	0.686	0.662	0.753	0.618	0.655	0.539	0.833	0.560
Logistic Regression	**0.908**	0.667	0.855	0.701	0.777	0.576	**0.875**	0.565
Linear SVM	0.672	0.672	0.715	0.715	0.597	0.597	0.852	0.568

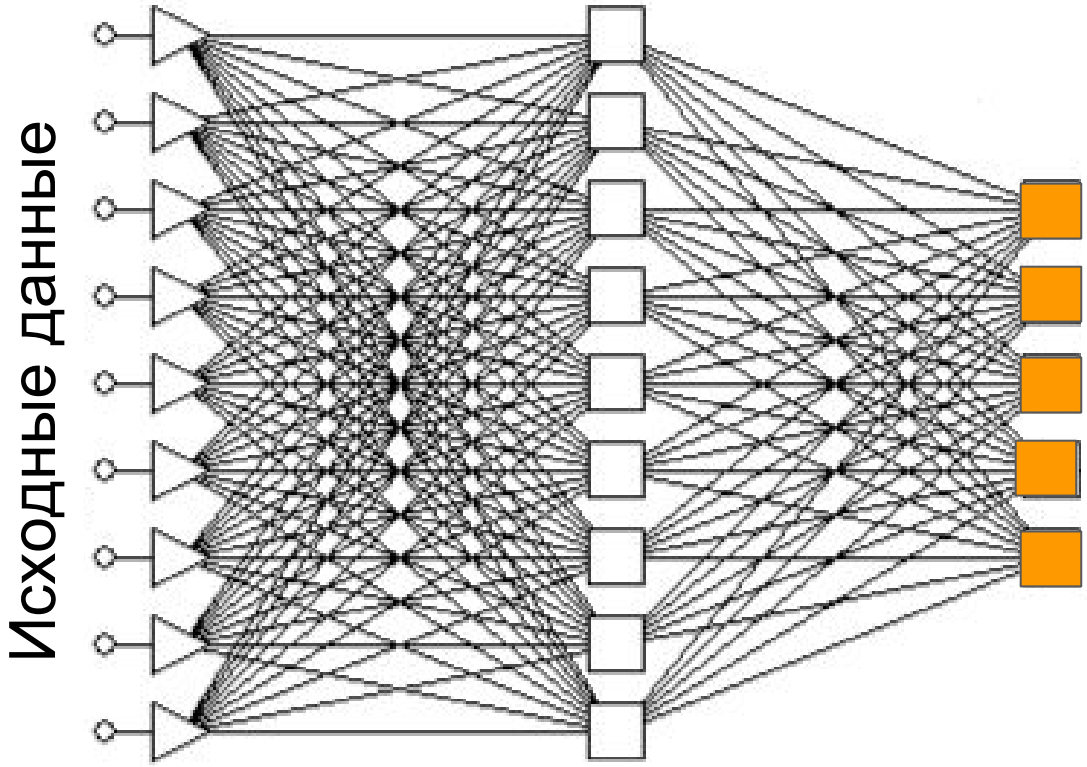
Результат классификации программой SIFT
на данных по собаке: acc = 0.852; на данных по мыши acc = 0.849

2. Новое признаковое пространство



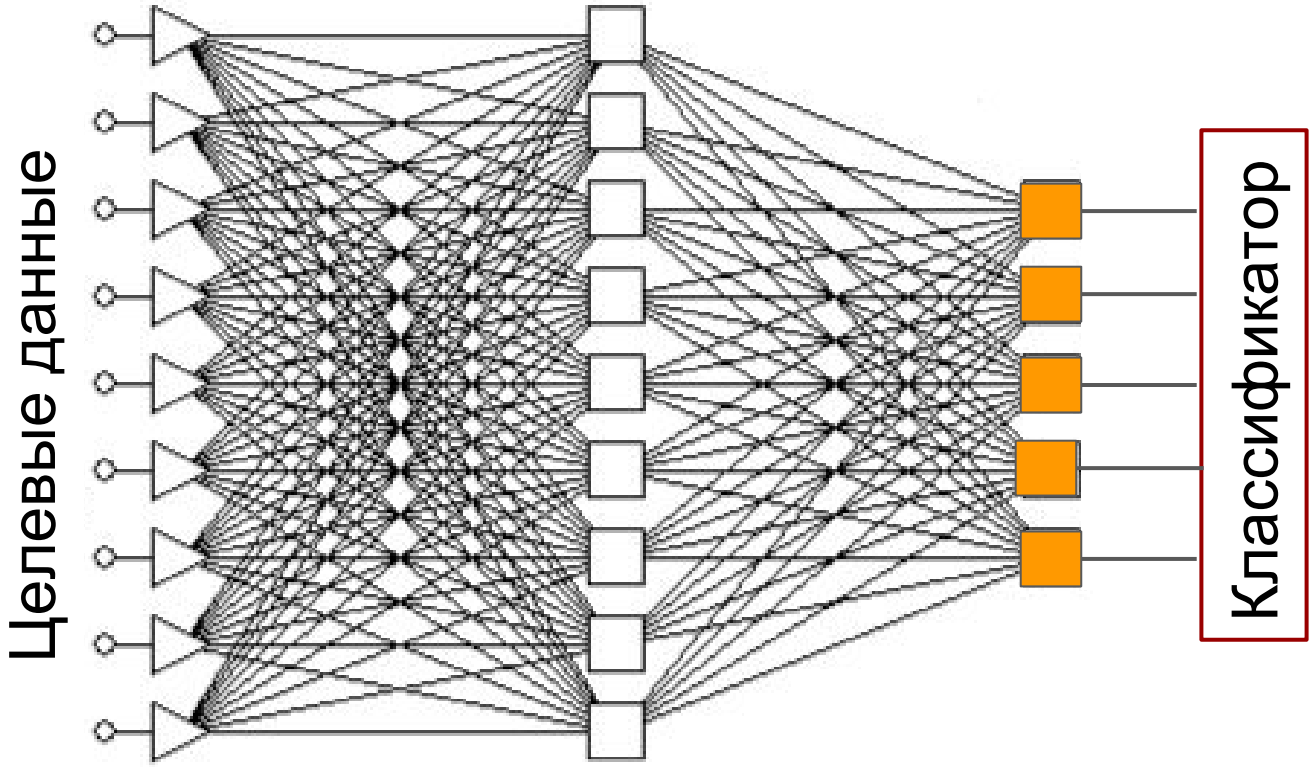
1. Происходит обучение и подбор параметров нейронной сети на исходных данных.

2. Новое признаковое пространство



2. Нейронная сеть отсекается на предпоследнем слое.

Transfer learning. Новое признаковое пространство



3. Целевые данные пропускаются через построенную нейронную сеть и дальнейшая классификация производится уже на преобразованных данных.

2. Новое признаковое пространство. Результаты

Классификатор	Без предобработки		С предобработкой		p-value
	acc	Дов. интервал	acc	Дов. интервал	
Deep Forest	**0.879**	(0.873, 0.884)	0.881	(0.877, 0.887)	0.3201
Random Forest	0.875	(0.869, 0.882)	**0.884**	(0.879, 0.889)	0.0028
Neural Network	0.840	(0.830, 0.850)	0.866	(0.854, 0.873)	0.0017
Gaussian SVM	0.866	(0.858, 0.875)	0.878	(0.872, 0.884)	0.0023
Polynomial SVM	0.866	(0.858, 0.874)	0.866	(0.859, 0.871)	0.6051
Logistic Regression	0.867	(0.859, 0.874)	0.876	(0.869, 0.883)	0.0038
Linear SVM	0.871	(0.863, 0.878)	0.873	(0.865, 0.880)	0.6766

*acc - accuracy (точность)

Результат классификации программой SIFT: acc = 0.849

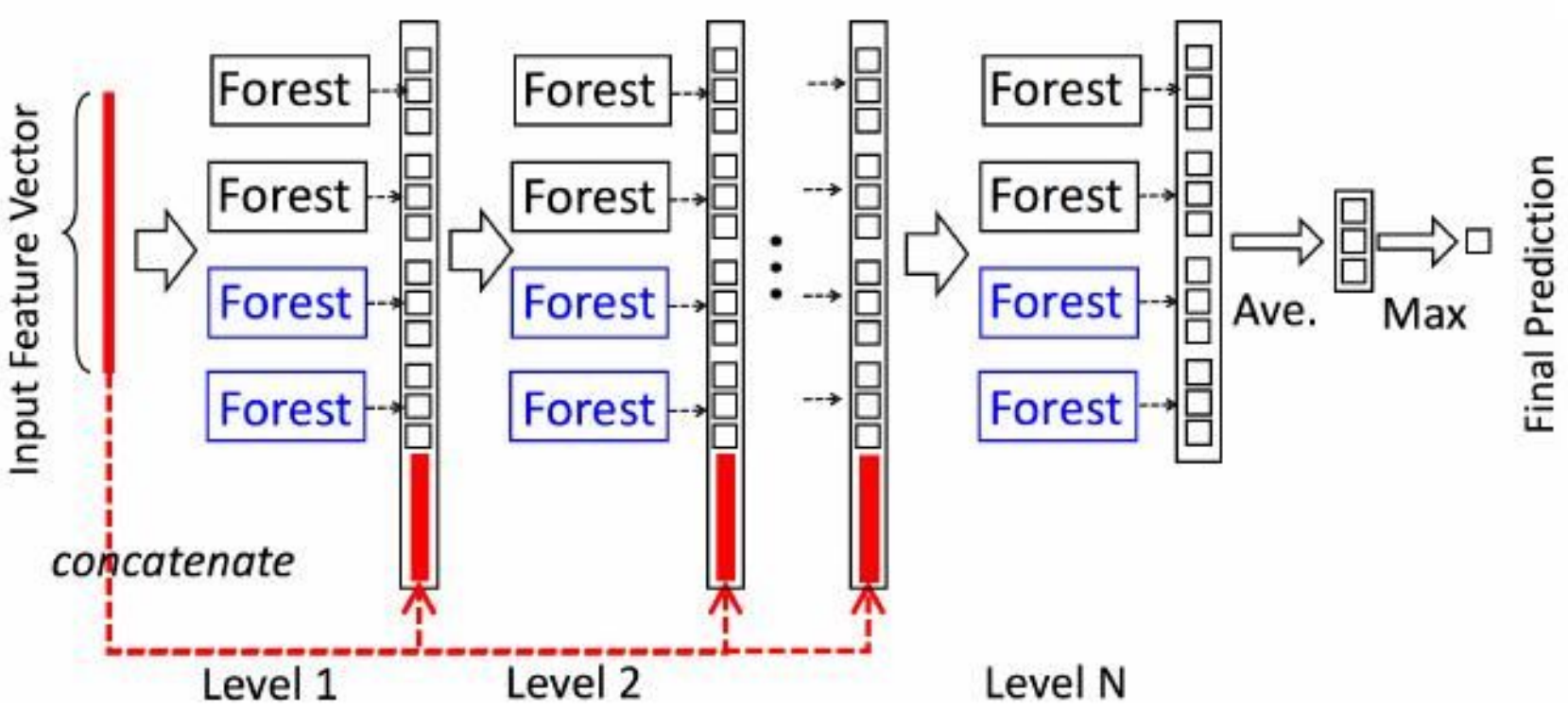
Результаты

- Подготовлены наборы вредных и нейтральных мутаций у мышей и собак.
- На наборах данных по человеку обучены 9 различных классификаторов, семь из которых оставлены для дальнейшего анализа.
- Рассмотрены два типа методов Transfer learning и их применение на полученных наборах данных по млекопитающим.
- Некоторые классификаторы после применения техник Transfer learning показали результат, лучший чем программа SIFT, широко использующаяся для решения данной задачи.

Вывод: использование методов Transfer learning является перспективным для решения задачи классификации вредных мутаций.

Спасибо за внимание!

Deep Forest



Z.-H. Zhou and J. Feng, "Deep Forest: Towards An Alternative to Deep Neural Networks," Feb. 2017.

Источники наборов данных

Организм	**Вредные мутации**	**Нейтральные мутации**
Человек (HumDiv dataset)	UniprotKB database (связь с болезнью)	Различия между гомологичными видами
Человек (HumVar dataset)	UniprotKB database (изменение функции белка)	Мутации внутри вида, частые в популяции
Собака	OMIA database	Различия между гомологичными видами
Мышь	MGI database	Мутации внутри вида, частые в популяции

Roc-кривые на данных по человеку (HumVar)