Российский университет транспорта РУТ (МИИТ)
Высшая инженерная школа (ВИШ)

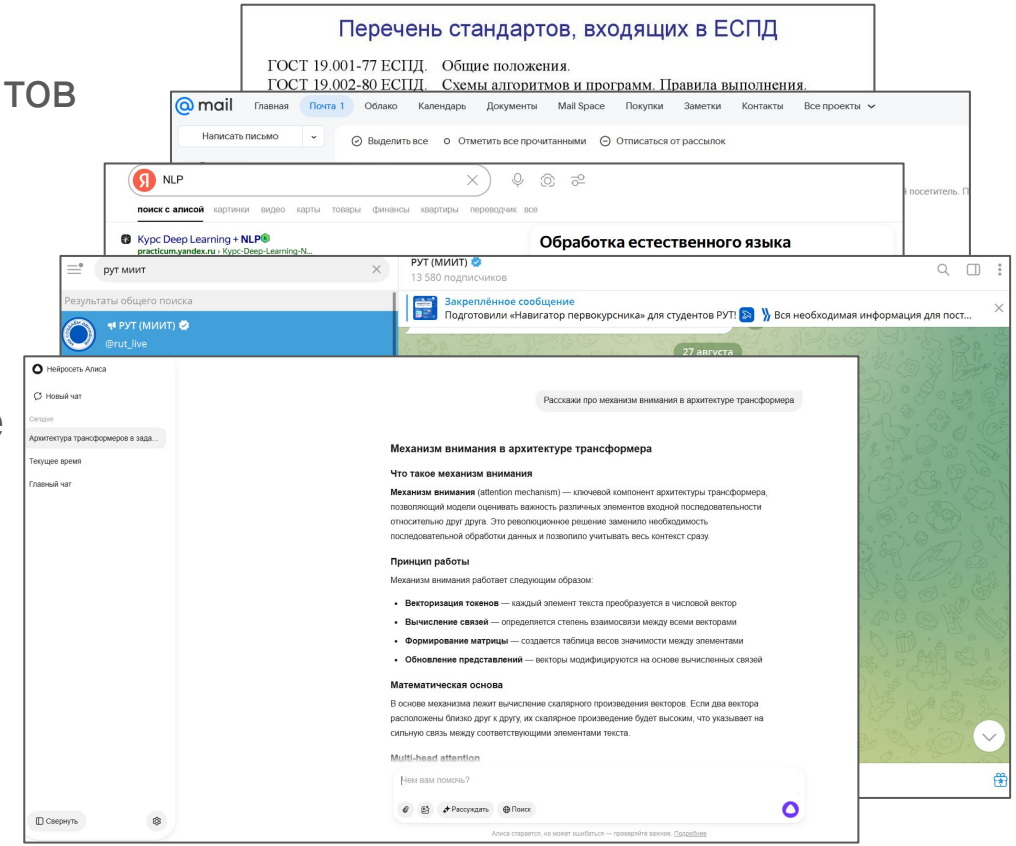
Анализ больших текстовых данных и текстовый поиск

Лектор: **Перчихин Олег Игоревич**
Преподаватели семинаров:
ШАД-311 - **Серенко Ирина Васильевна**
ШЦТ-311 - **Зиганшин Ренат Фанисович**

1 сентября 2025

Текстовые данные: где они встречаются?

- Архивы официальных документов
- Письма электронной почты
- Поисковые системы
- Социальные сети
- Базы данных
- Корпусы текстов в лингвистике
- Электронные библиотеки
- И др.



Основные задачи обработки текстовых данных

- **Информационный/текстовый поиск** (**Information retrieval**) - задача поиска в большой коллекции документов тех, что удовлетворяют запросам пользователя (поисковые системы, базы данных)
- **Интеллектуальный анализ текстов** (**Text mining**) - задача выделения из неструктурированных текстовых данных полезной информации (классификация, кластеризация, аннотирование документов, поиск ключевых понятий, связей между ними)
- Извлечение информации и построение **графов знаний** (**Knowledge graphs**) - задача автоматического построения структурированных семантических сетей из неструктурированной информации, например, из коллекции текстов
- **Языковое моделирование** (**Language Modeling**) и **генерация текстов**

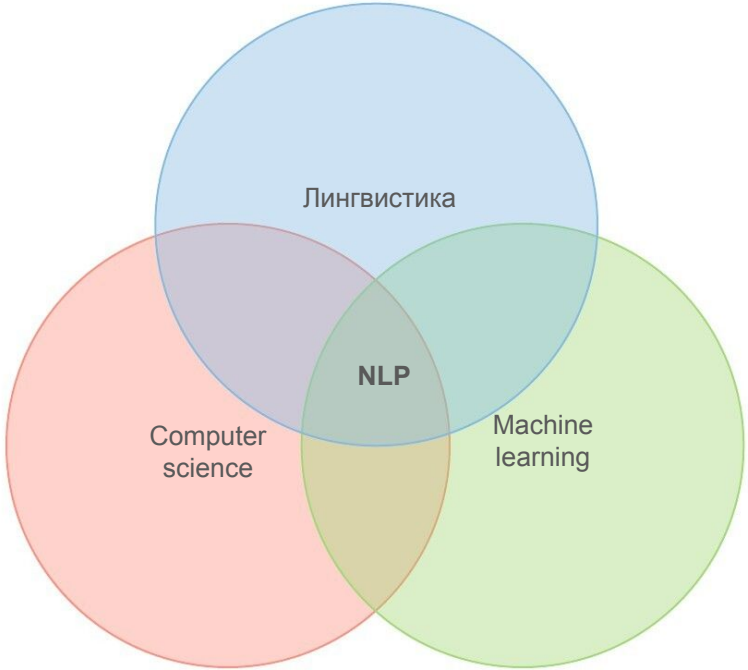
Natural Language Processing

NLP — обработка естественного языка

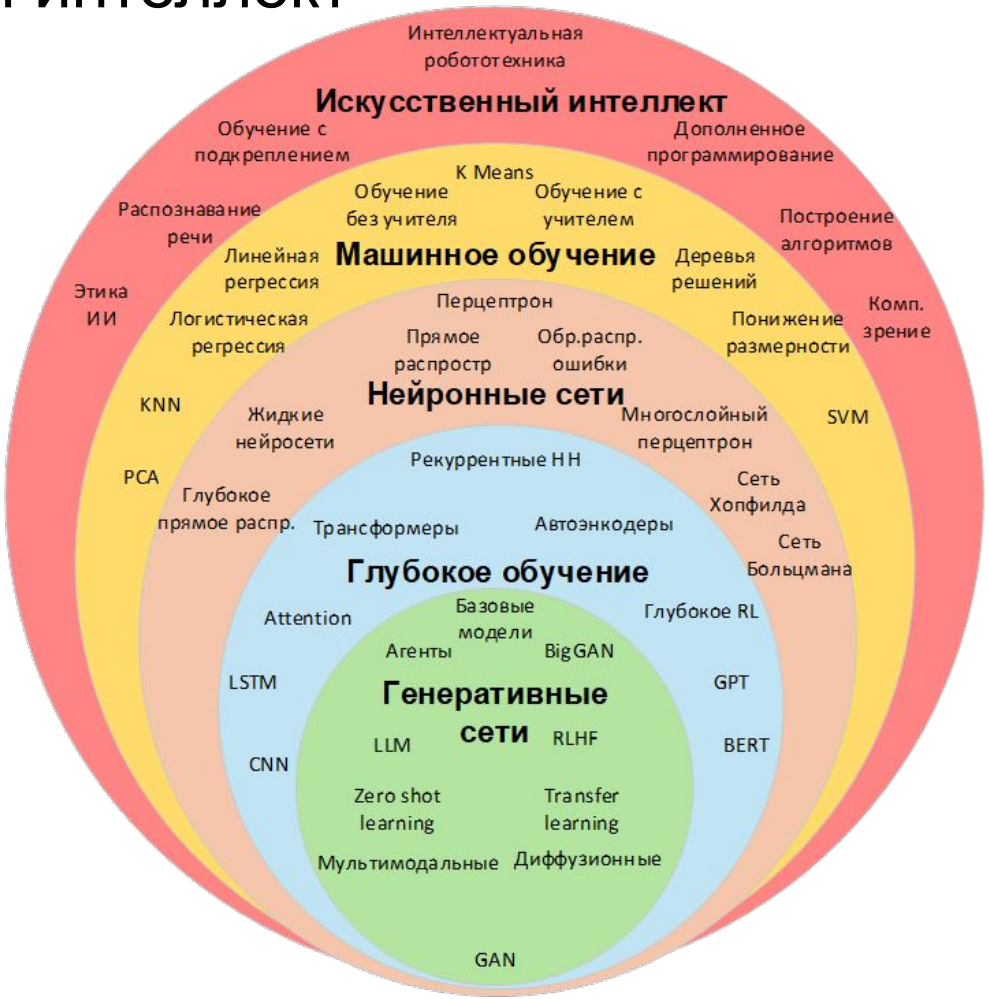
NLP позволяет применять алгоритмы машинного обучения для текста и речи

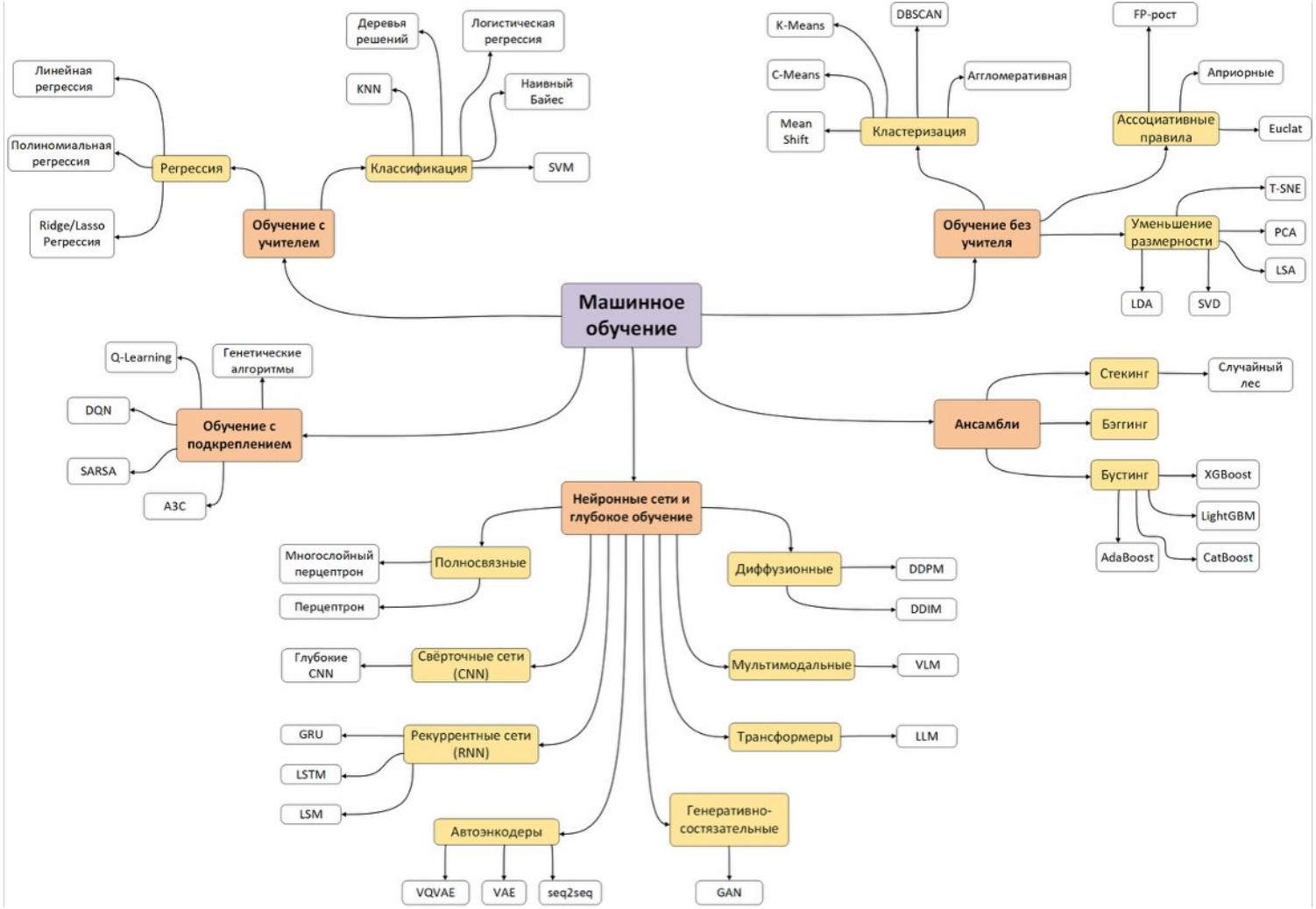
Цель: понимать, интерпретировать и генерировать человеческий язык

Современные подходы в анализе текстовых данных и текстовом поиске используют алгоритмы NLP



Искусственный интеллект

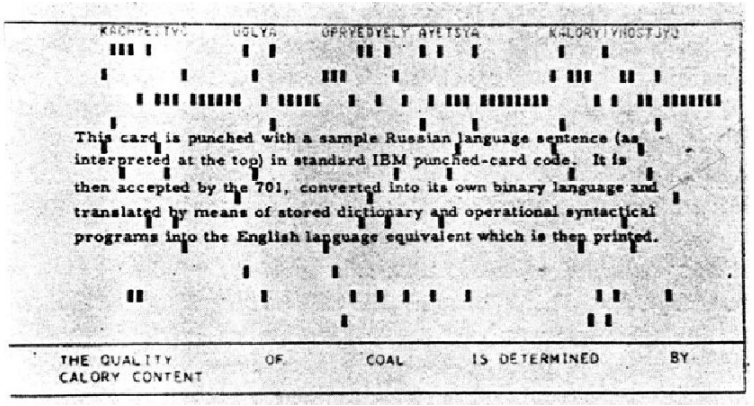



Машинное обучение
Обучение с учителем
Регрессия
Линейная регрессия
Полиномиальная регрессия
Ridge/Lasso Регрессия
Классификация
Деревья решений
Логистическая регрессия
KNN
Наивный Байес
SVM
Обучение без учителя
Кластеризация
K-Means
C-Means
Mean Shift
DBSCAN
Агломеративная
Ассоциативные правила
FP-рост
Априорные
Euclat
Уменьшение размерности
T-SNE
PCA
LSA
LDA
SVD
Обучение с подкреплением
Q-Learning
Генетические алгоритмы
DQN
SARSA
A3C
Ансамбли
Стекинг
Случайный лес
Бэггинг
Бустинг
XGBoost
LightGBM
CatBoost
AdaBoost
Нейронные сети и глубокое обучение
Полносвязные
Многослойный перцептрон
Перцептрон
Свёрточные сети (CNN)
Глубокие CNN
Рекуррентные сети (RNN)
GRU
LSTM
LSM
Автоэнкодеры
VQVAE
VAE
seq2seq
Генеративно-состязательные
GAN
Диффузионные
DDPM
DDIM
Мультимодальные
VLM
Трансформеры
LLM

Сфера применения NLP

- Поиск информации и рекомендательные системы
- Переводчики (Google Translate, DeepL)
- Голосовые помощники (Siri, Alexa, Алиса)
- Large Language Models (ChatGPT, DeepSeek)
- Анализ отзывов и соцсетей (тональности)
- Анализ документов (моделирование тем)
- Медицинские системы (обработка историй болезней)
- Чат-боты

История NLP



1950-1970: Простые правила и словари, «rule-based NLP»

- Тест Тьюринга, «Джорджтаунский эксперимент»

1970-1990: Синтаксические парсеры, ручные грамматики

1990-2000: Статистический подход. N-gram модели

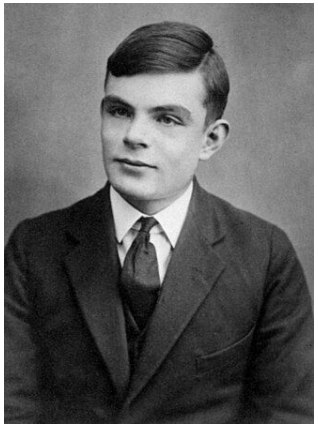
- Вероятность вместо правил — лучше масштабируемость

2000-2010: Нейросетевой подход. Появление word embeddings

2010-2020: Революция глубокого обучения. LSTM, Transformer

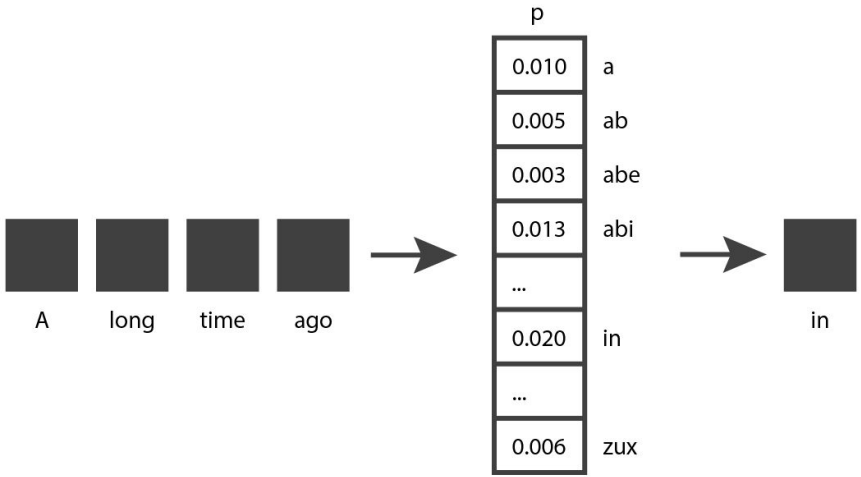
Настоящее время: Массовое внедрение: от поиска и перевода до ассистентов и генеративного контента

Тренды будущего: Универсальные мультимодальные модели, автономность, этические аспекты



Языковые модели можно условно разделить на

- Статистические (классические) языковые модели. N-граммы
- Рекуррентные нейросетевые языковые модели. RNN, LSTM
- Трансформеры (подвид LLM). BERT, GPT
- *Мультимодальные нейросетевые модели. CLIP, Vision-to-Text LLMs, Multimodal RAG, Stable Diffusion, DALL-E, OpenAI's Sora, …



Процесс генерации текста
Языковая модель предсказывает следующее слово в последовательности. Слово «in» имеет наибольшую вероятность среди возможных продолжений

Экспертные системы

Экспертная система - программный комплекс, который оперирует знаниями в определенной предметной области с целью выработки рекомендаций или решения проблем.

ЭС может полностью взять на себя функции эксперта или играть роль ассистента для человека, принимающего решение.

Технология ЭС - это одно из направлений искусственного интеллекта.

Характеристики экспертных систем

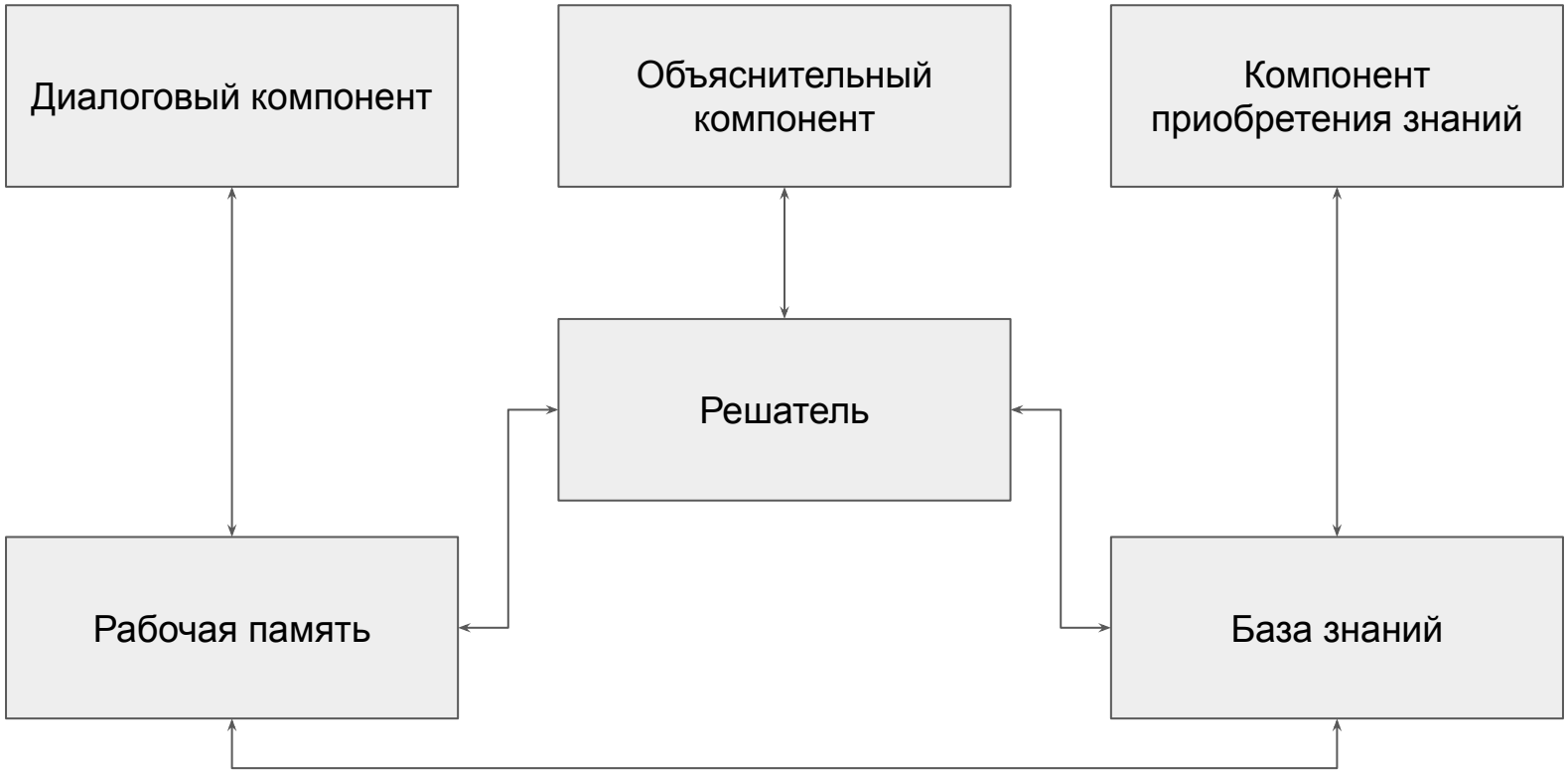
- Помимо выполнения вычислительных операций ЭС формирует определенные соображения и выводы, основываясь на тех знаниях, которыми она располагает
- Знания в ЭС представлены на специальном языке и хранятся отдельно от программного кода, который и формирует выводы и соображения
- Этот компонент программы принято называть базой знаний

Структура экспертных систем

Типичная статическая ЭС состоит из следующих основных компонентов:

1. Решатель (интерпретатор)
2. Рабочая память, называемая также базой данных
3. База знаний
4. Компоненты приобретения знаний
5. Объяснительный компонент
6. Диалоговый компонент

Структура экспертных систем



Структура курса

1. Лекции (Понедельник 16:55-18:15):

   теоретические материалы

2. Семинары:

   практические занятия на Python, обсуждения курсовых работ

   ШАД-311 (Пятница 11:40-13:00)

   ШЦТ-311 (Понедельник 18:30-19:50)

3. Репозиторий курса на GitHub: все материалы курса будут публиковаться там после занятий

Что будет оцениваться

1. Выполнение заданий на семинарах
2. Курсовые работы
3. Итоговый экзамен