Искусственный интеллект – мечта и/или кошмар юриста

Машинное обучение в LegalTech: задачи, возможности, ограничения

Воронцов Константин Вячеславович

● Московский Физико-Технический Институт ●
● Вычислительный Центр им. А.А.Дородницына ФИЦ ИУ РАН ●
● ШАД Яндекс ●

voron@forecsys.ru

Распознавание

навигация
- Заглавная страница
- Сообщество
- Новости
- Последние правки
- Случайная статья
- Справка
- Инструктаж
- Вопросы и ответы
- ToDo

- Энциклопедия анализа данных
- Популярные и обзорные статьи
- Публикации
- Полезные ссылки

поиск
Перейти Найти

Участник:Vokov

Воронцов Константин Вячеславович

профессор РАН, д.ф.-м.н.
Зав. отделом «Интеллектуальные системы» Вычислительного центра ФИЦ ИУ РАН.
Зав. лабораторией машинного интеллекта МФТИ.
Проф. каф. «Интеллектуальные системы» ФУПМ МФТИ.
Доц. каф. «Математические методы прогнозирования» ВМК МГУ.
Преподаватель Школы анализа данных Яндекс.
Зам. директора по науке ЗАО «Форексис», www.forecsys.ru.

Один из идеологов и Администраторов ресурса **MachineLearning.RU**.

Прочие подробности — на подстранице **Curriculum vitæ**.

Мне можно написать письмо.

- Профиль ORCID = 0000-0002-4244-4270
- Профиль SCOPUS ID = 6507982932
- Профиль WoS ResearcherID = G-7857-2014
- Профиль Google Scholar
- Профиль DBLP
- Профиль РИНЦ ID = 15081
- Профиль в системе ИСТИНА
- Профиль MathNet.ru

Содержание [убрать]

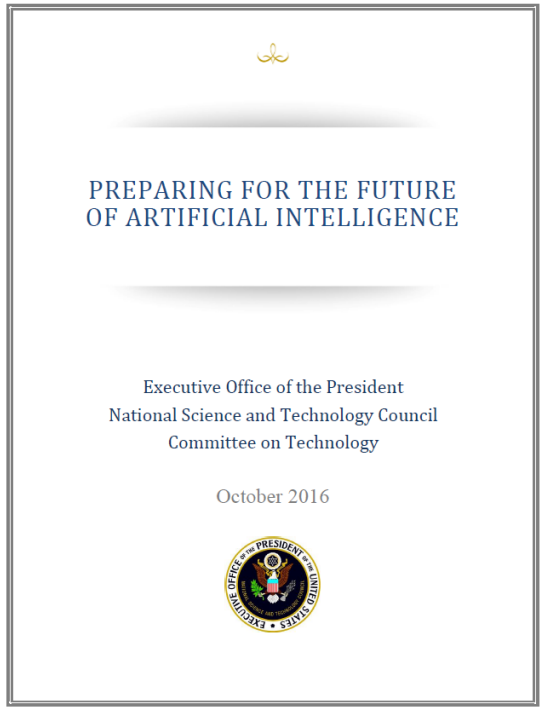
Машинное обучение в LegalTech

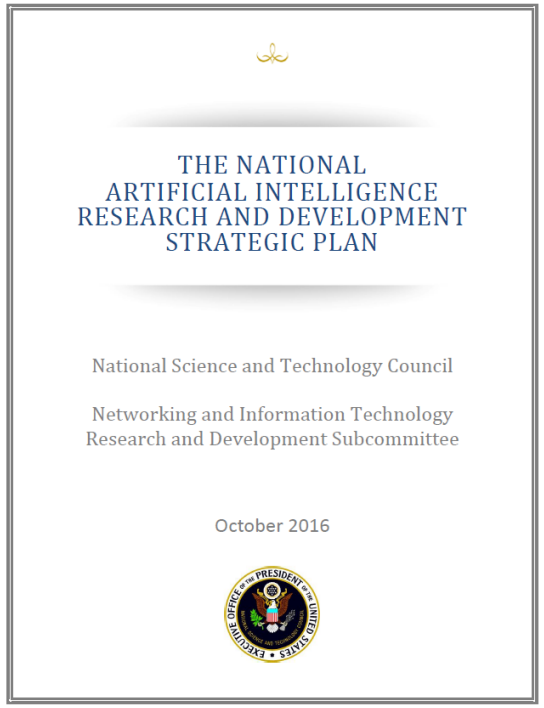
1. Задачи, возможности и ограничения машинного обучения
   - Постановки задач, терминология и методология машинного обучения
   - Бум искусственного интеллекта и нейронных сетей
   - Задачи и технологии анализа текстов естественного языка
   - Разведочный информационный поиск
2. LegalTech: задачи, технологии, сервисы
3. Уберизация юридического консультирования

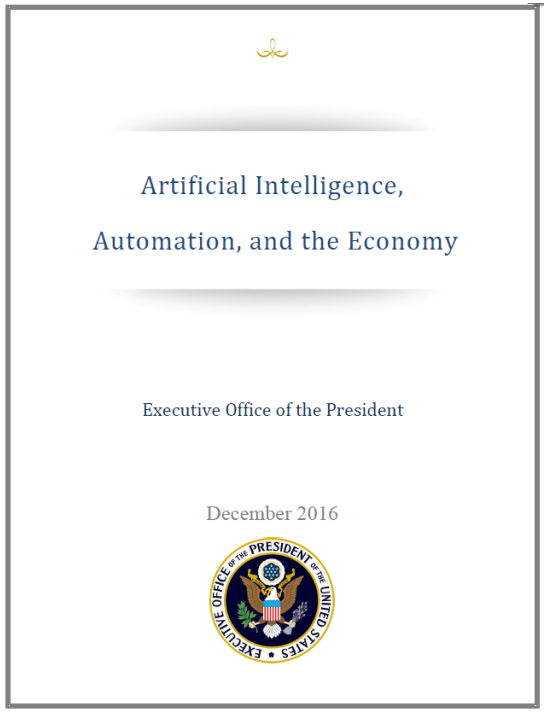
«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, *искусственном интеллекте* и *машинном обучении*» (2016)

Клаус Мартин Шваб,
президент Всемирного
экономического форума

Отчёты Белого дома США, май-октябрь 2016



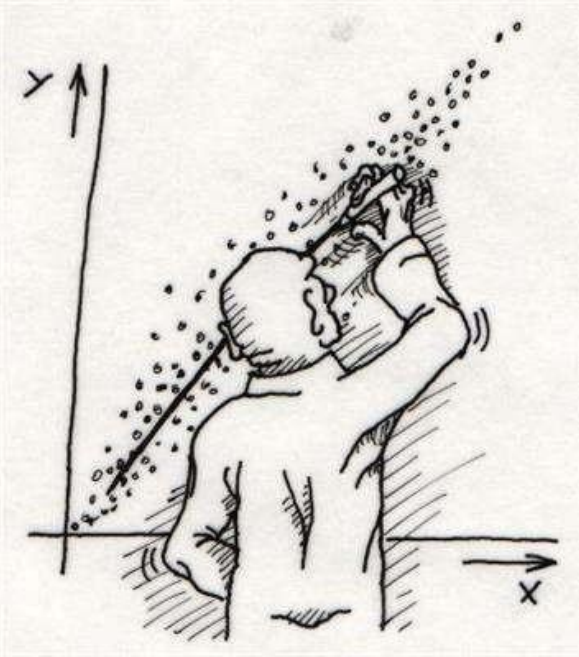




«Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future»

Машинное обучение (Machine Learning, ML)

- одна из ключевых информационных технологий будущего
- наиболее успешное направление *искусственного интеллекта*, вытеснившее экспертные системы и инженерию знаний



- проведение функции через заданные точки в сложно устроенных пространствах
- математическое моделирование в условиях, когда знаний мало, данных много
- тысячи различных методов и алгоритмов
- около 100 000 научных публикаций в год

Основная задача машинного обучения

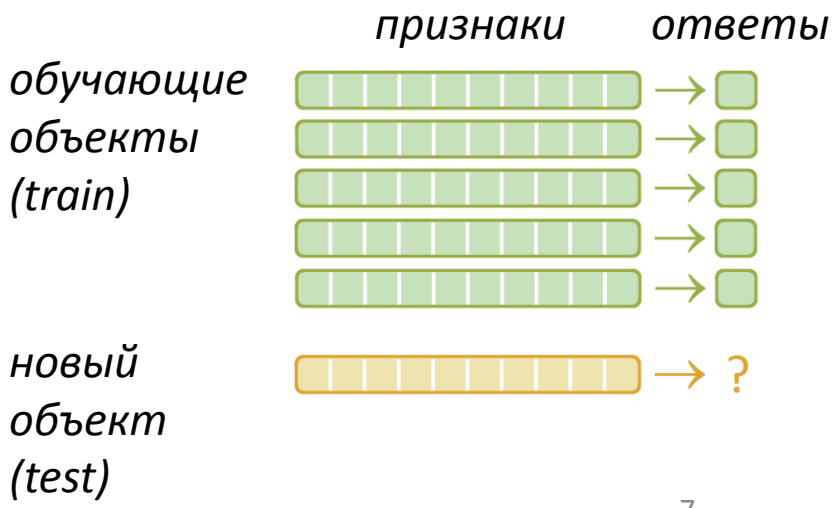
Этап №1 – обучение с учителем

- **На входе:**
  данные – выборка прецедентов «*объект* → *ответ*»,
  каждый объект описывается набором *признаков*
- **На выходе:**
  модель, предсказывающая ответ по объекту

Если нет данных,
то нет
и машинного
обучения

Этап №2 – применение

- **На входе:**
  данные – новый объект
- **На выходе:**
  предсказание ответа на новом объекте



Примеры задач машинного обучения

- **Кредитный скоринг:**
  объект – данные о заёмщике
  ответ – вероятность дефолта, решение по кредиту

- **Информационный поиск в Интернете:**
  объект – данные о паре «запрос и документ»
  ответ – оценка релевантности документа запросу

Google bing Yandex Baidu 百度

- **Рекомендательные системы в Интернете / TV:**
  объект – данные о паре «пользователь, товар / фильм»
  ответ – оценка вероятности покупки / просмотра

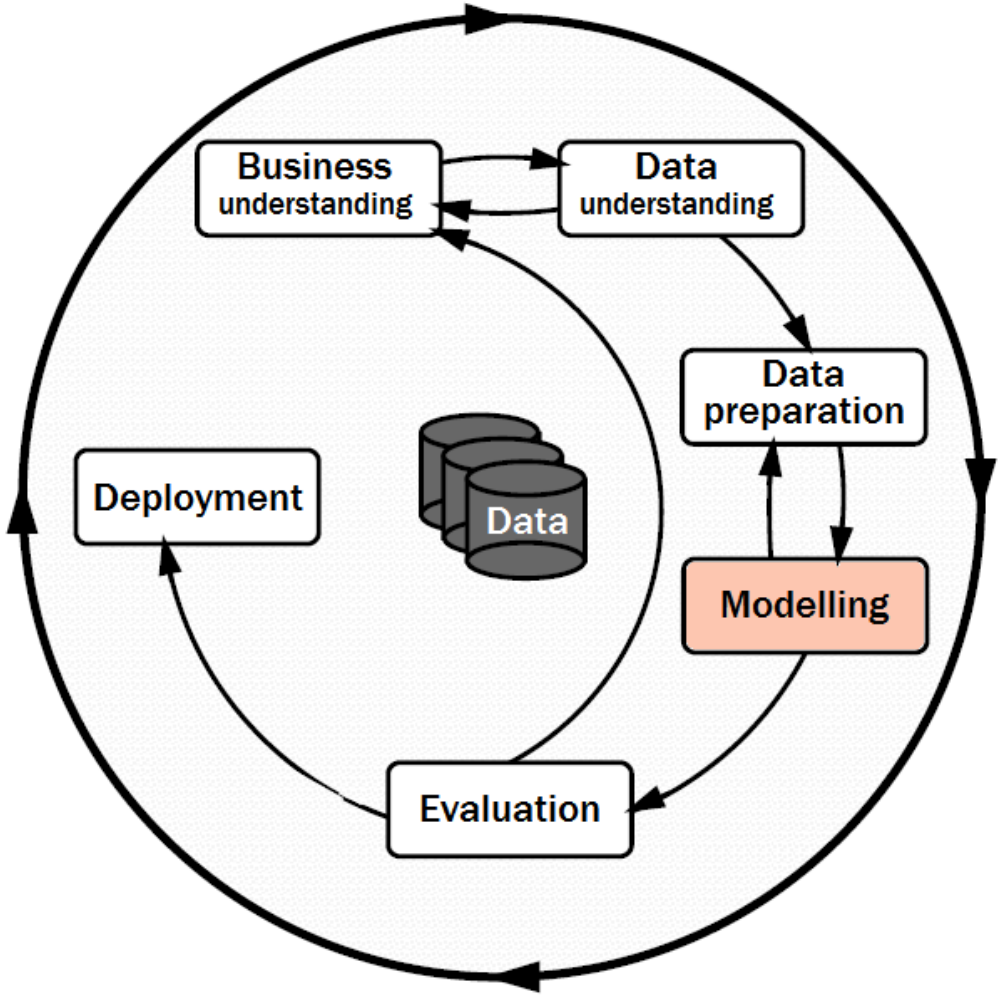
amazon.com Prime
NETFLIX
hulu

Примеры задач ML в LegalTech

- **Предсказание результатов судебных процессов:**
  объект – описание дела: документы, записи о событиях
  ответ – вероятность выиграть дело
- **Информационный поиск:**
  объект – описание дела или вопрос на естественном языке
  ответ – ранжированный список релевантных НПА и/или схожих дел
- **Рекомендательный сервис:**
  объект – пара «описание дела, профиль юриста / юрфирмы»
  ответ – оценка релевантности
- **Автоматическая генерация ответов на вопросы:**
  объект – текст вопроса на естественном языке
  ответ – текст ответа, включая фрагменты НПА

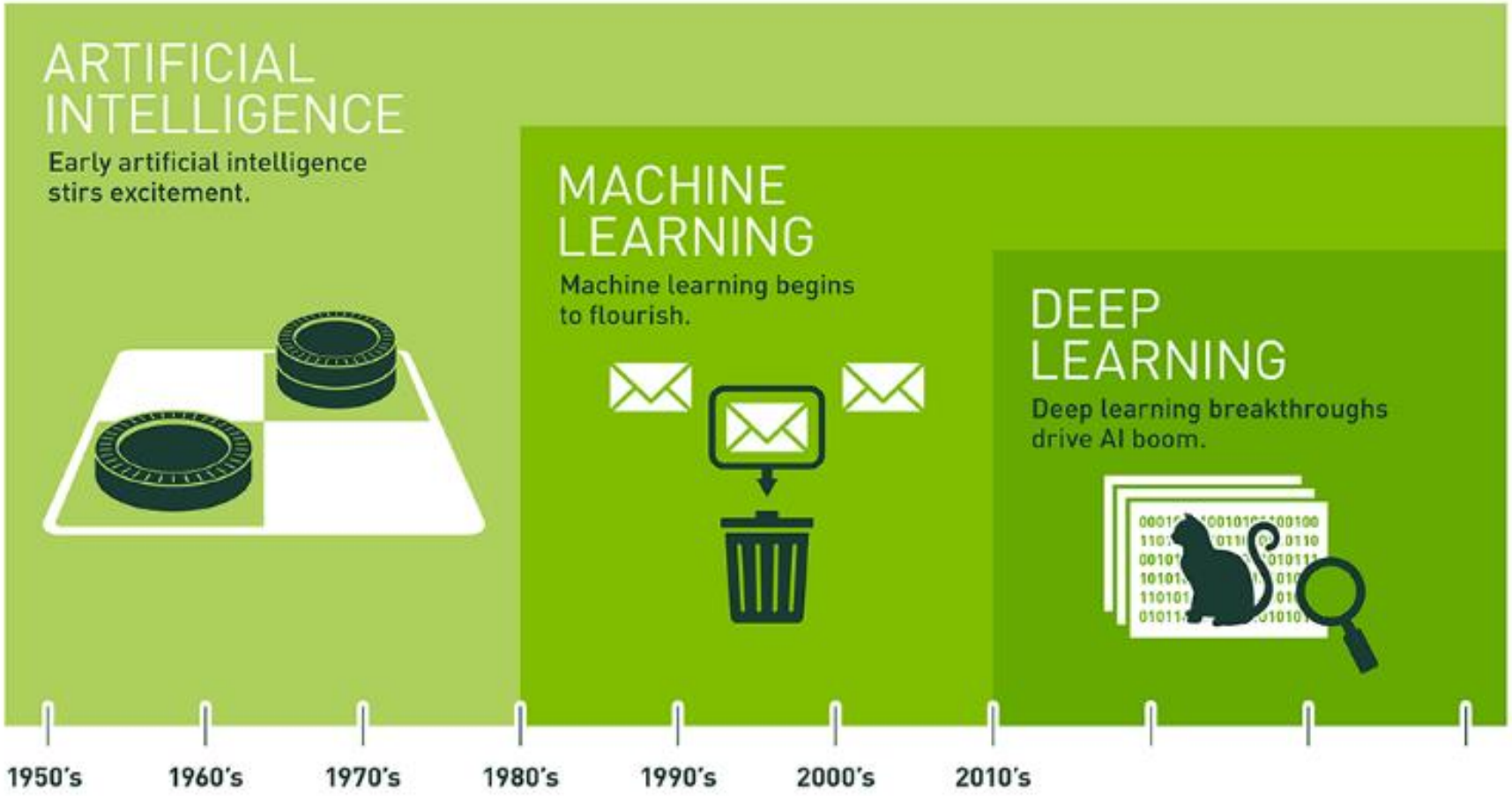
Этапы решения задач анализа данных

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



- понимание бизнес-задач
- понимание данных
- предобработка данных
- инженерия признаков
- построение моделей
- оптимизация параметров
- контроль переобучения
- (кросс-)валидация решения
- внедрение и эксплуатация

Эволюция искусственного интеллекта



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Глубокое обучение – одна из новейших технологий машинного обучения

Глубокие нейронные сети обеспечили прорыв в компьютерном зрении

ImageNet: открытая выборка
15M размеченных изображений

Google: Распознавание кадров с котами на видео из Youtube

Бум искусственного интеллекта

1997: IBM Deep Blue обыграл чемпиона мира по шахматам

2005: Беспилотный автомобиль: DARPA Grand Challenge

2006: Google Translate – статистический машинный перевод

2011: 40 лет DARPA CALO привели к созданию Apple Siri

2011: IBM Watson победил в ТВ-игре «Jeopardy!»

2011–2015: ImageNet: 25% → 3,5% ошибок против 5% у людей

2015: Фонд OpenAI в $1 млрд. Илона Маска и Сэма Альтмана

2016: DeepMind, OpenAI: динамическое обучение играм Atari

2016: Google DeepMind обыграл чемпиона мира по игре го

2017: OpenAI обыграл чемпиона мира по компьютерной игре Dota 2

Открытые данные для ИИ

- Выгоды открытых данных
  - быстрое выявление центров компетенций, подбор кадров
  - быстрое выявление наилучших решений и «потолка качества»
  - формирование проф.сообществ, обучение на реальных кейсах
  - популяризация научных знаний в любых областях, где есть данные
- Конкурсы анализа данных
  - www.NetflixPrize.com (2006-2009) – первый крупный конкурс, $1 млн.
  - www.kaggle.com – самая известная платформа
  - www.FakeNewsChallenge.org – один из последних конкурсов
- DataRing.ru – отечественная конкурсная платформа
  - консалтинг по подготовке данных и условий конкурса
  - очистка, отбор, агрегирование, деперсонификация данных

NLP: обработка естественного языка

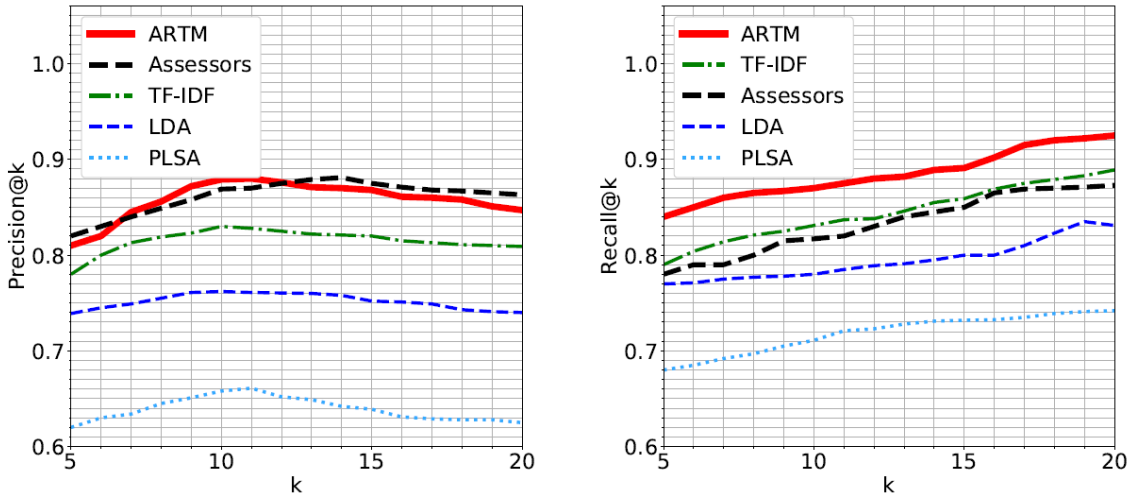
Стек технологий NLP (см. также nlpub.ru)

- Морфологический анализ и лемматизация (lemmatization)
- Синтаксический анализ (syntax analysis)
- Автоматическое выделение терминов (automatic term extraction)
- Распознавание именованных сущностей (named entity recognition)
- Сегментация текста (text segmentation)
- Классификация текстов (text classification)
- Кластеризация текстов (text clustering)
- Тематическое моделирование (topic modeling)
- Семантические векторные представления слов (word embedding)
- Семантический анализ и построение онтологий (ontology learning)
- Аннотирование и суммаризация (text summarization)
- Обучаемое ранжирование (learning to rank)
- Информационный поиск (information retrieval)
- Ответы на вопросы, машинный перевод, чат-боты (sequence-to-sequence)

Пример. Разведочный информационный поиск

- Длинные запросы (1 стр. А4)
- 100 запросов
- 3 асессора на каждый запрос
- 30 минут в среднем на запрос
- Разметка на Яндекс.Толока
- Коллекции техно-новостей

Результат: *точность* (precision) и *полнота* (recall) поиска



A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Машинное обучение в LegalTech

1. Задачи, возможности и ограничения машинного обучения
2. LegalTech: задачи, технологии, сервисы
   - Обзор по материалам Skolkovo LegalTech 2017
   - Информационно-поисковые сервисы
   - Сервисы на основе искусственного интеллекта
3. Уберизация юридического консультирования

Что в LegalTech не является AI/ML

FreshDoc.ru конструктор документов	Конструктор документов
Робот-юрист	Конструктор документов
Симплоер *Право следующего поколения*	Смарт-контракты
JEFFIT	Автоматизация рутины в юридических фирмах
PatentBot	Быстрая регистрация торговых марок
XSUD	Учёт судебных дел, доступ к данным госорганов
Preferentum	Проверка документов на соответствие закону

* По материалам конференции Skolkovo LegalTech, 1 декабря 2017

Что в LegalTech является или становится AI/ML

Европейская Юридическая Служба

Автоматические ответы на вопросы

Правовед.RU

Автоматические ответы на вопросы (нейросеть)

PLATFORMA
Финансирование судебных процессов

Финансирование судебных процессов, возможен рекомендательный поиск инвесторов и адвокатов

FLEXBBY

Документооборот на основе модели событий

КонсультантПлюс
надежная правовая поддержка

Гарант
Информационно-правовое обеспечение

Preferentum

ПРАВО RU

Специализированные модели информационного поиска

* По материалам конференции Skolkovo LegalTech, 1 декабря 2017

Машинное обучение в LegalTech

1. LegalTech: задачи, технологии, сервисы
2. Задачи, возможности и ограничения машинного обучения
3. Уберизация юридического консультирования
   - По пути ImageNet: как собрать открытый датасет вопросов и ответов
   - Гибридный интеллект: как заинтересовать всех участников процесса

Уберизация юридического консультирования

Веб-сервис вопросов-ответов + подбор вопросов под юриста

Вопросы и ответы – тексты на естественном языке

Централизованное накопление выборки «вопрос - ответ»

Если клиент не удовлетворён автоматическим ответом,
то вопрос переадресуется юристам

- По пути ImageNet:
  как собрать открытый датасет вопросов и ответов
- Гибридный интеллект:
  кесарю кесарево, а боту ботово
- Как заинтересовать всех участников в повышении качества данных

Слагаемые успеха в AI-трансформации

- Большие открытые данные
- Открытый обмен данными и знаниями в проф.сообществе
- Очистка, грамотная подготовка, анонимизация данных
- Организация конкурсов анализа данных или хакатонов
- Привлечение экспертизы в области AI-ML-NLP
- Переход от статичного обучения к активному

Педро Домингос. «Верховный алгоритм». 2016.