Всероссийский Фестиваль «Наука 0+»

Искусственный интеллект: мифы, реальность, перспективы

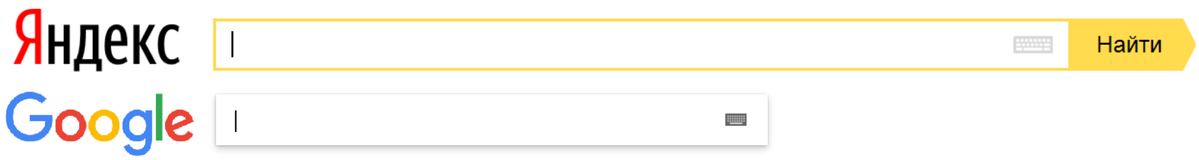
Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,
зав. лаб. Машинного обучения и семантического анализа
Института Искусственного Интеллекта МГУ,
профессор, и.о. зав. кафедрой Математических методов прогнозирования ВМК МГУ,
профессор, зав. кафедрой Машинного обучения и цифровой гуманитаристики МФТИ,
г.н.с. ФИЦ «Информатика и управление» РАН

Искусственный интеллект: мифы, реальность, перспективы

1. Задачи машинного обучения
   - Бум искусственного интеллекта и нейронных сетей
   - Классические задачи машинного обучения
   - Обучаемая векторизация сложно структурированных данных
2. Задачи обработки и понимания естественного языка
   - Задачи разметки текста
   - Задача конкурса ПРО//ЧТЕНИЕ
   - Модели внимания и трансформеры
3. Задачи понимания общественно-политического дискурса
   - Языковые явления эпохи постправды
   - Задачи разметки текста
   - Детектирование пропаганды, манипуляций, поляризации мнений

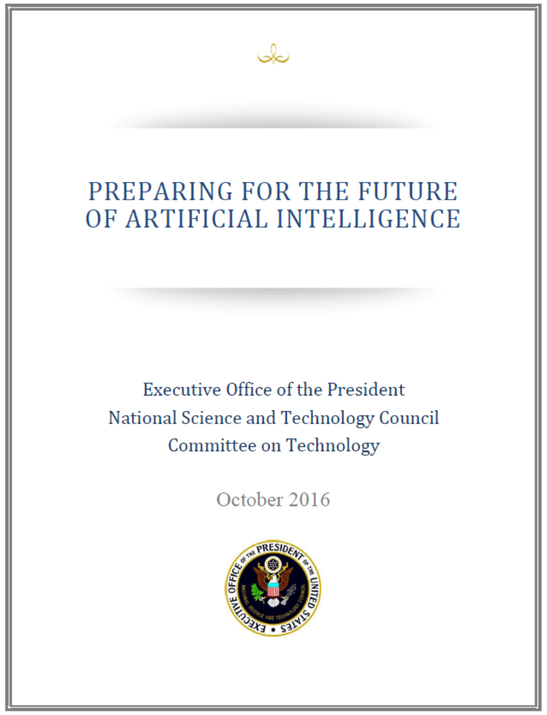
Технологии ИИ, которые меняют мир

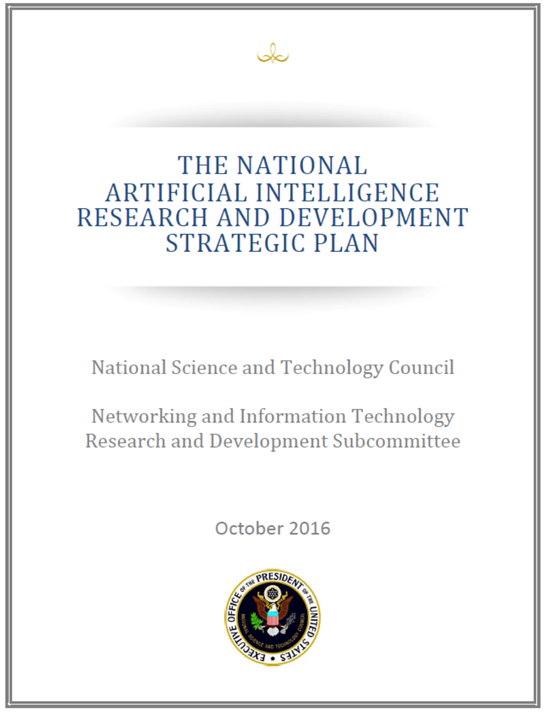


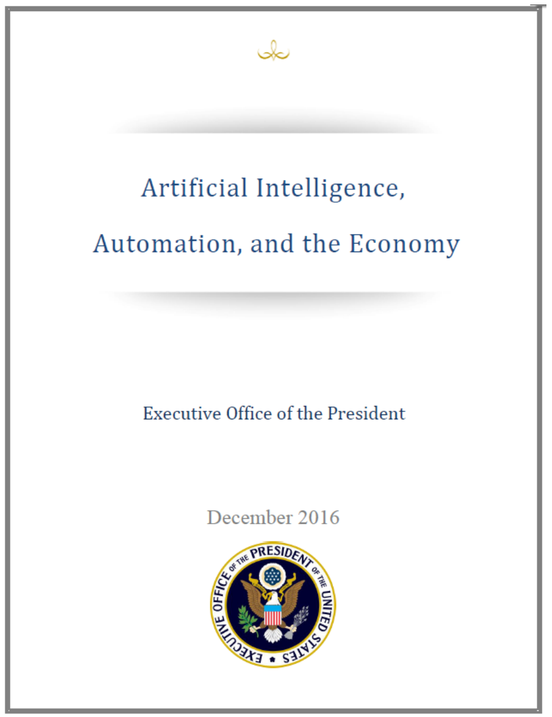
«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, *искусственном интеллекте* и *машинном обучении*» (2016)

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»

Национальная стратегия развития ИИ в РФ

10 окт 2019

УКАЗ

ПРЕЗИДЕНТА РОССИЙСКОЙ ФЕДЕРАЦИИ

О развитии искусственного интеллекта в Российской Федерации

В целях обеспечения ускоренного развития искусственного интеллекта в Российской Федерации, проведения научных исследований в области искусственного интеллекта, повышения доступности информации и вычислительных ресурсов для пользователей, совершенствования системы подготовки кадров в этой области п о с т а н о в л я ю:

1. Утвердить прилагаемую Национальную стратегию развития искусственного интеллекта на период до 2030 года.

2. Правительству Российской Федерации:

а) до 15 декабря 2019 г. обеспечить внесение изменений в национальную программу "Цифровая экономика Российской Федерации", в том числе разработать и утвердить федеральный проект "Искусственный интеллект";

б) представлять Президенту Российской Федерации ежегодно доклад о ходе реализации Национальной стратегии развития искусственного интеллекта на период до 2030 года;

в) предусматривать при формировании в 2020 - 2030 годах проектов федеральных бюджетов на очередной финансовый год и на плановый период бюджетные ассигнования на реализацию настоящего Указа.

УТВЕРЖДЕНА
Указом Президента
Российской Федерации
от 10 октября 2019 г. № 490

НАЦИОНАЛЬНАЯ СТРАТЕГИЯ

развития искусственного интеллекта на период до 2030 года

I. Общие положения

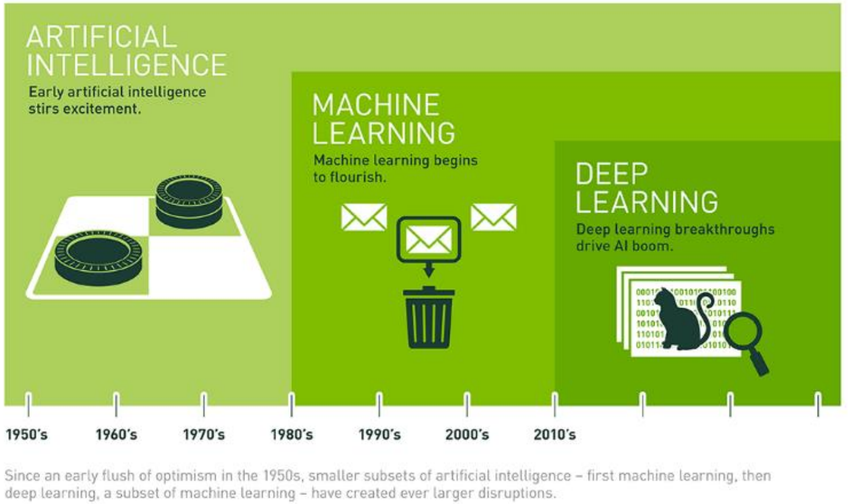
1. Настоящей Стратегией определяются цели и основные задачи развития искусственного интеллекта в Российской Федерации, а также меры, направленные на его использование в целях обеспечения национальных интересов и реализации стратегических национальных приоритетов, в том числе в области научно-технологического развития.

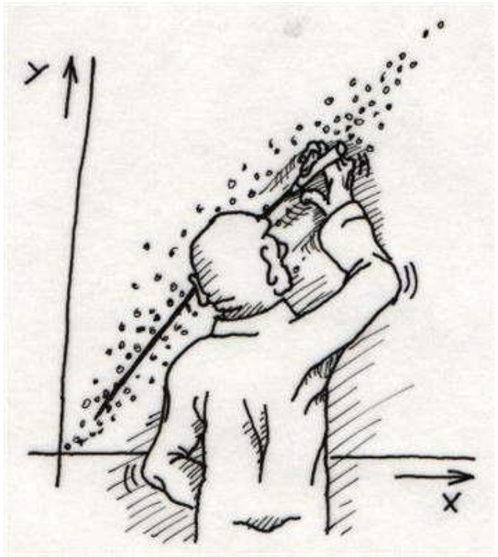
2. Правовую основу настоящей Стратегии составляют Конституция Российской Федерации, Федеральный закон от 28 июня 2014 г. № 172-ФЗ "О стратегическом планировании в Российской Федерации", указы Президента Российской Федерации от 7 мая 2018 г. № 204 "О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года", от 9 мая 2017 г. № 203 "О Стратегии развития информационного общества в Российской Федерации на 2017 - 2030 годы", от 1 декабря 2016 г. № 642 "О Стратегии научно-технологического развития Российской Федерации" и иные нормативные правовые акты Российской Федерации, определяющие направления применения информационных технологий в Российской Федерации.

3. Настоящая Стратегия является основой для разработки (корректировки) государственных программ Российской Федерации, государственных программ субъектов Российской Федерации, федеральных и региональных проектов, плановых и программно-целевых документов государственных корпораций, государственных компаний, акционерных обществ с государственным участием,

Машинное обучение (Machine Learning, ML)

- одна из ключевых информационных технологий будущего
- наиболее успешное направление ИИ, 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
- более 100 000 научных публикаций в год

Задачи машинного обучения с учителем

Этап №1 – обучение (train)

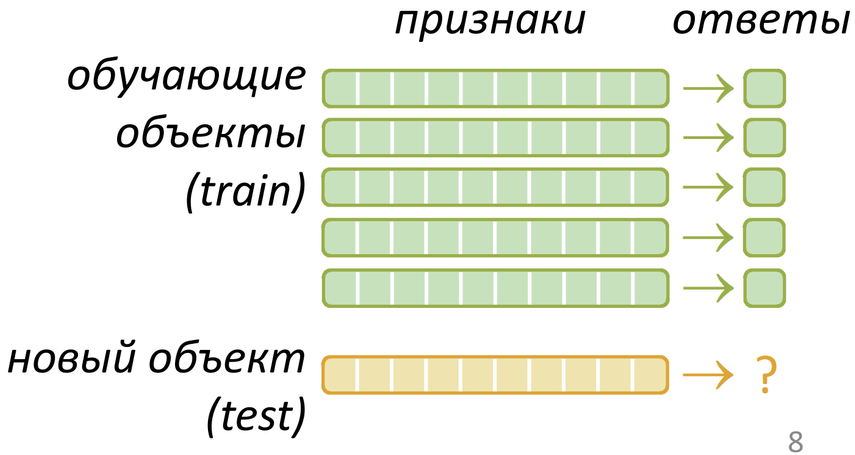
- **На входе:**
  данные – выборка пар «*объект* → *ответ*»,
  каждый объект описывается *вектором признаков*
- **На выходе:**
  модель, предсказывающая ответ по объекту

Задача поставлена,
если у неё есть «ДНК»:

- ***Дано***
- ***Найти***
- ***Критерий***

Этап №2 – применение (test)

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



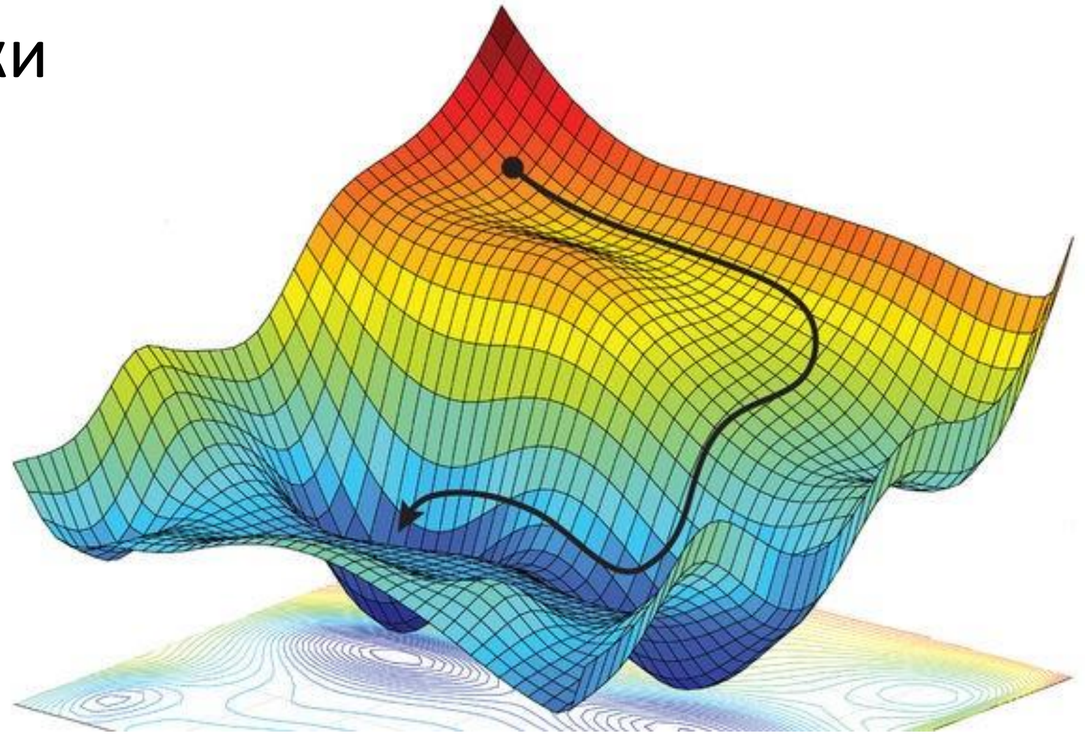
Машинное обучение – это оптимизация

x – вектор объекта обучающей выборки

w – параметры модели

$\mathrm{Loss}(x, w)$ – функция потерь

$Q(w)$ – критерий качества модели



Задача на этапе обучения модели:

$$Q(w) = \sum_{x} \mathrm{Loss}(x, w) \to \min$$

Способ решения – численные методы оптимизации

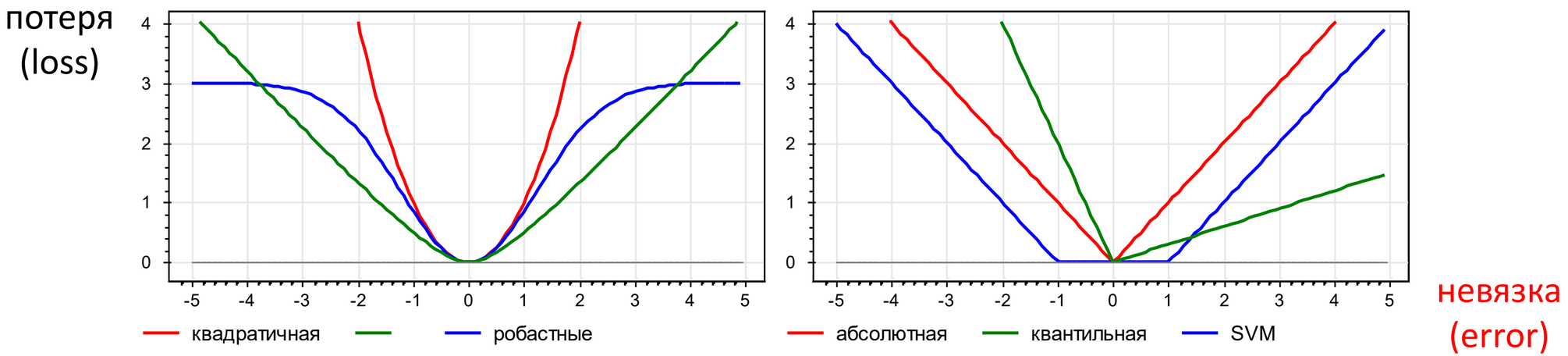
Задачи восстановления регрессии (regression)

x – вектор объекта обучающей выборки, y – числовой ответ

$a(x, w)$ – модель регрессии с параметрами w

Например, $a(x, w) = \sum_j w_j x_j$ — линейная модель регрессии

$\text{Loss}(x, w) = (a(x, w) - y)^2$ – квадратичная функция потерь



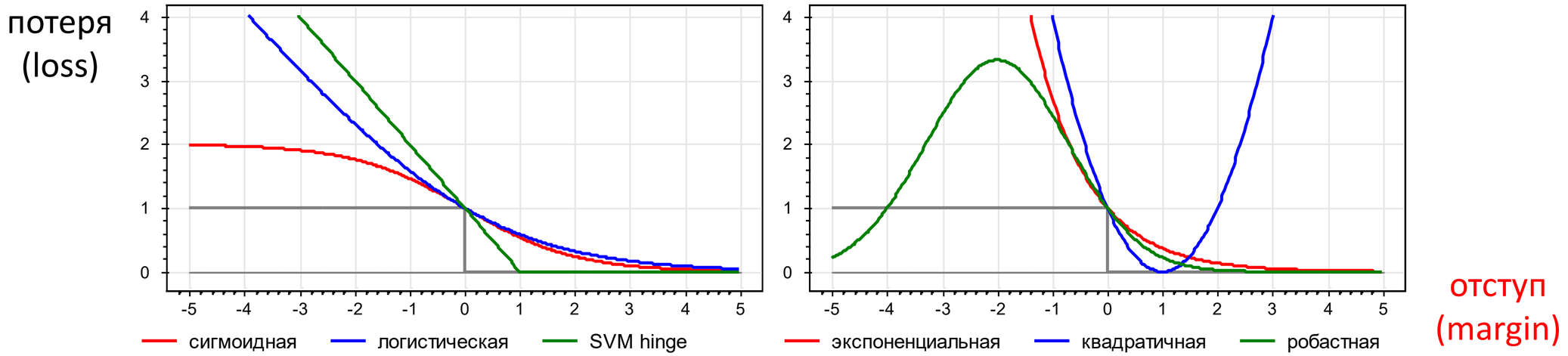
Задачи классификации (classification)

x – вектор объекта обучающей выборки, y – ответ ($+1$ или -1)

$a(x, w)$ – модель классификации с параметрами w

Например, $a(x, w) = \text{sign}\left(\sum_j w_j x_j\right)$ — линейная модель

$\text{Loss}(x, w) = \max\left(0, 1 - y\sum_j w_j x_j\right)$ — функция потерь «hinge»

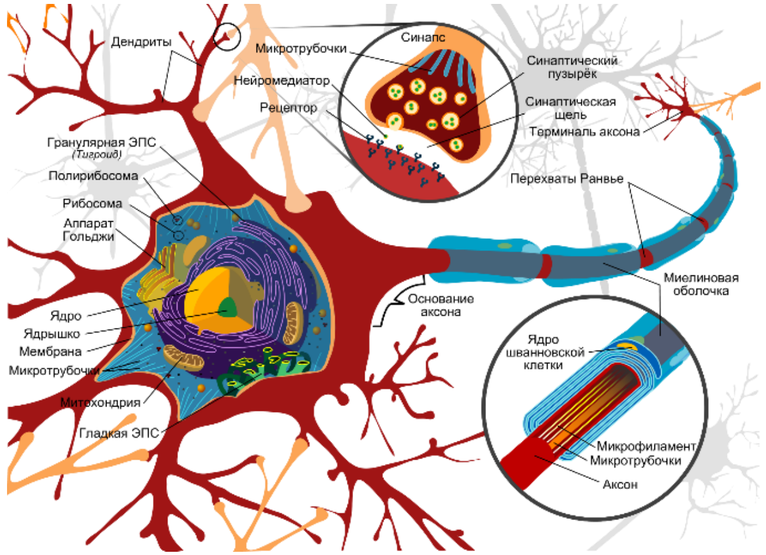


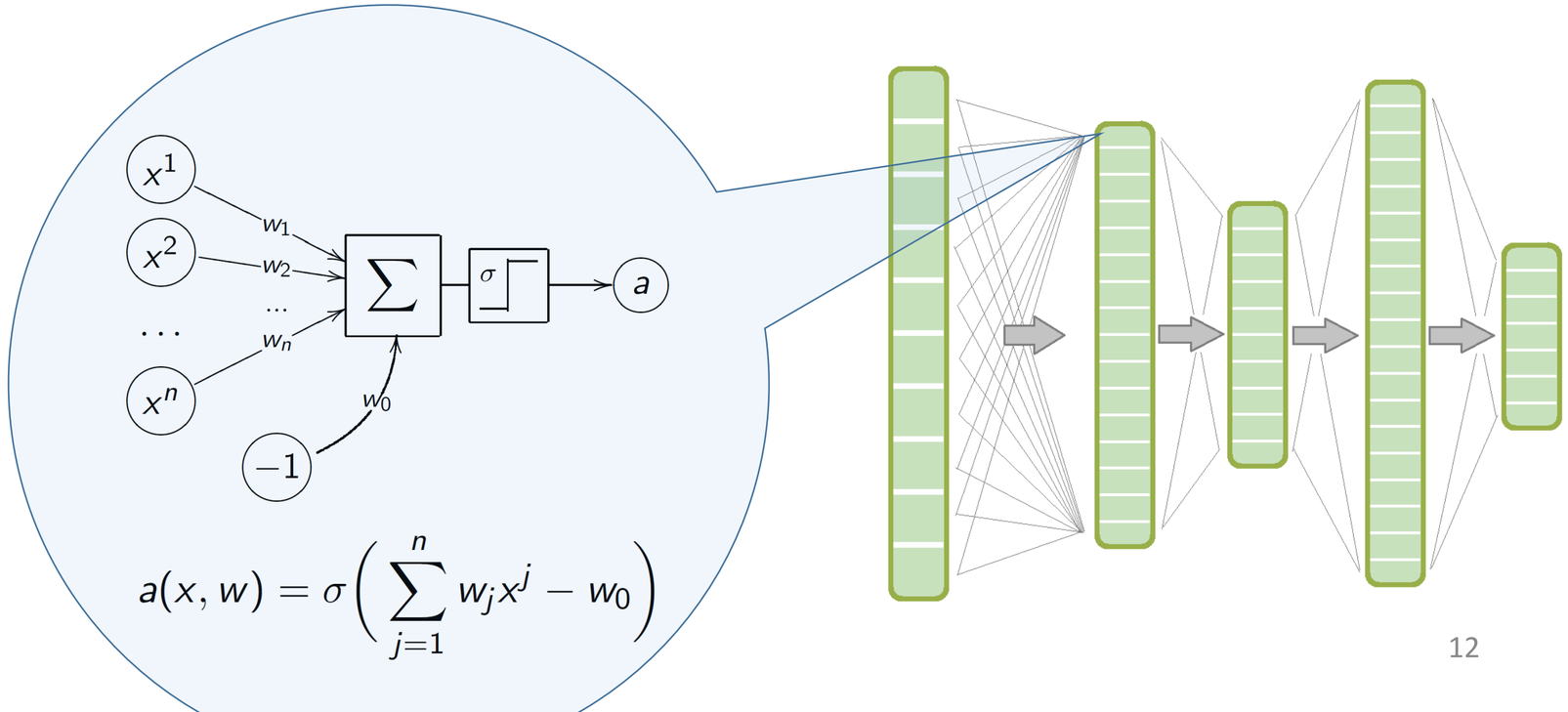
Искусственные нейронные сети

На каждом слое сети вектор объекта преобразуется в новый вектор

Эти преобразования обучаемые, их параметры входят в w

Каждое преобразование (нейрон) – взвешенная сумма признаков

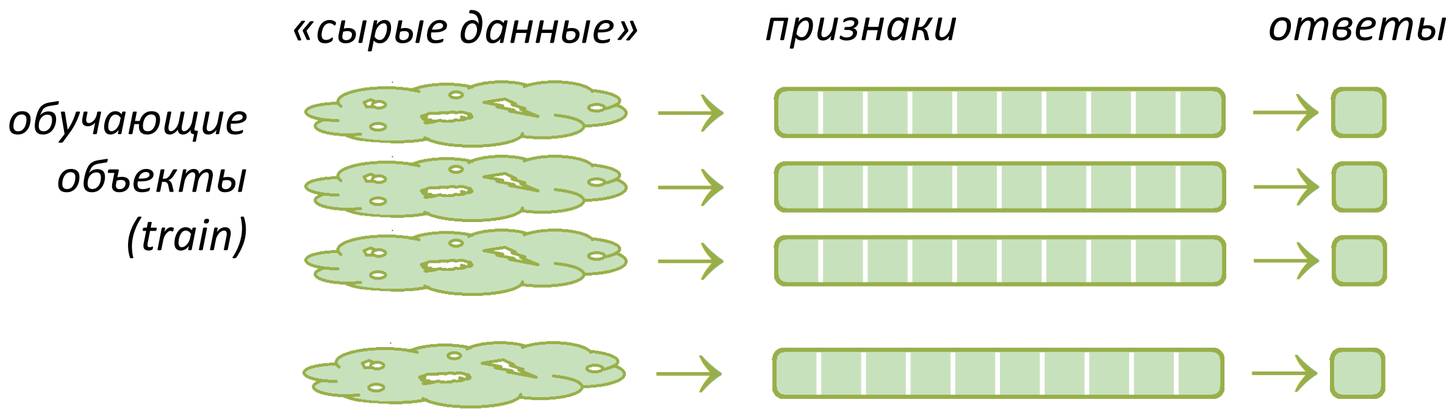




Глубокие нейронные сети

Вход: сложно структурированные «сырые» данные объектов

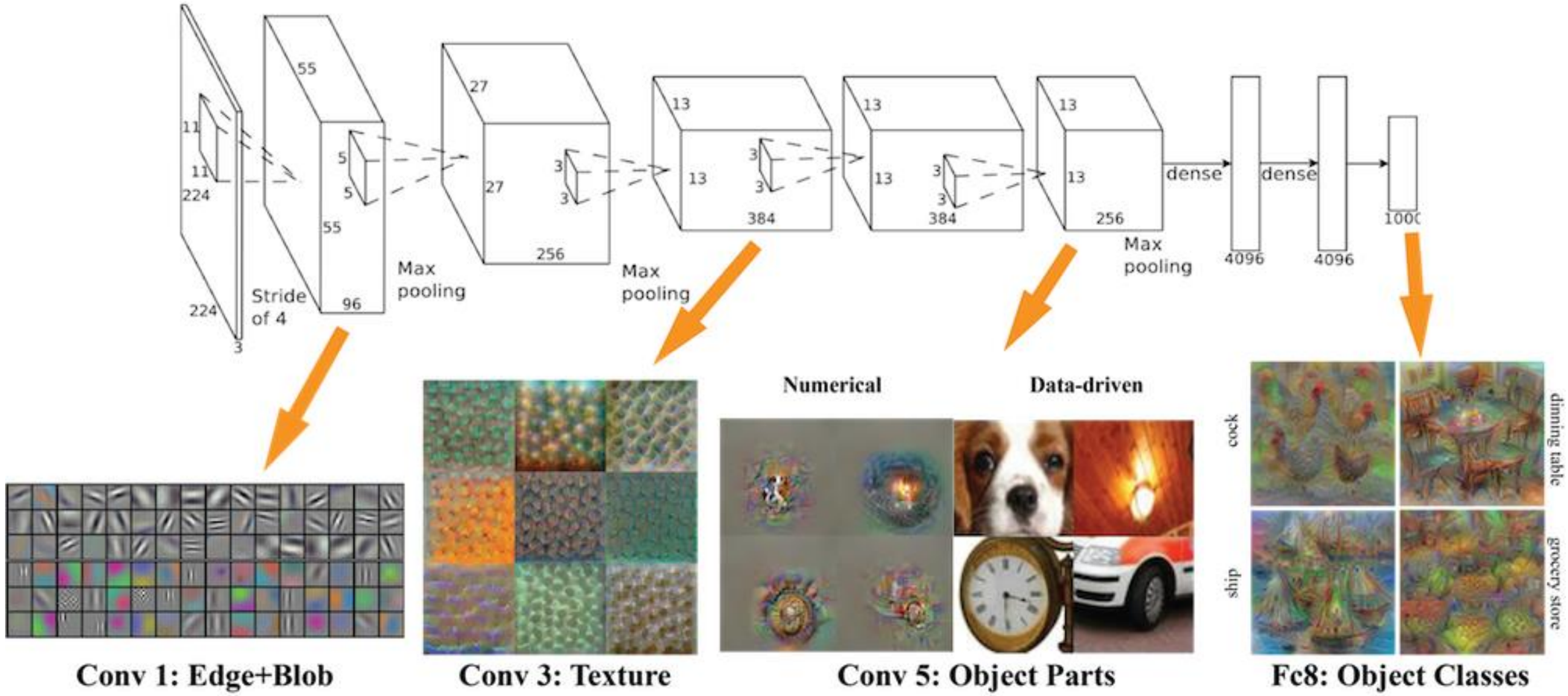
Выход: векторные представления объектов, затем ответы



Deep Learning – это обучаемая векторизация объектов со сложной внутренней структурой

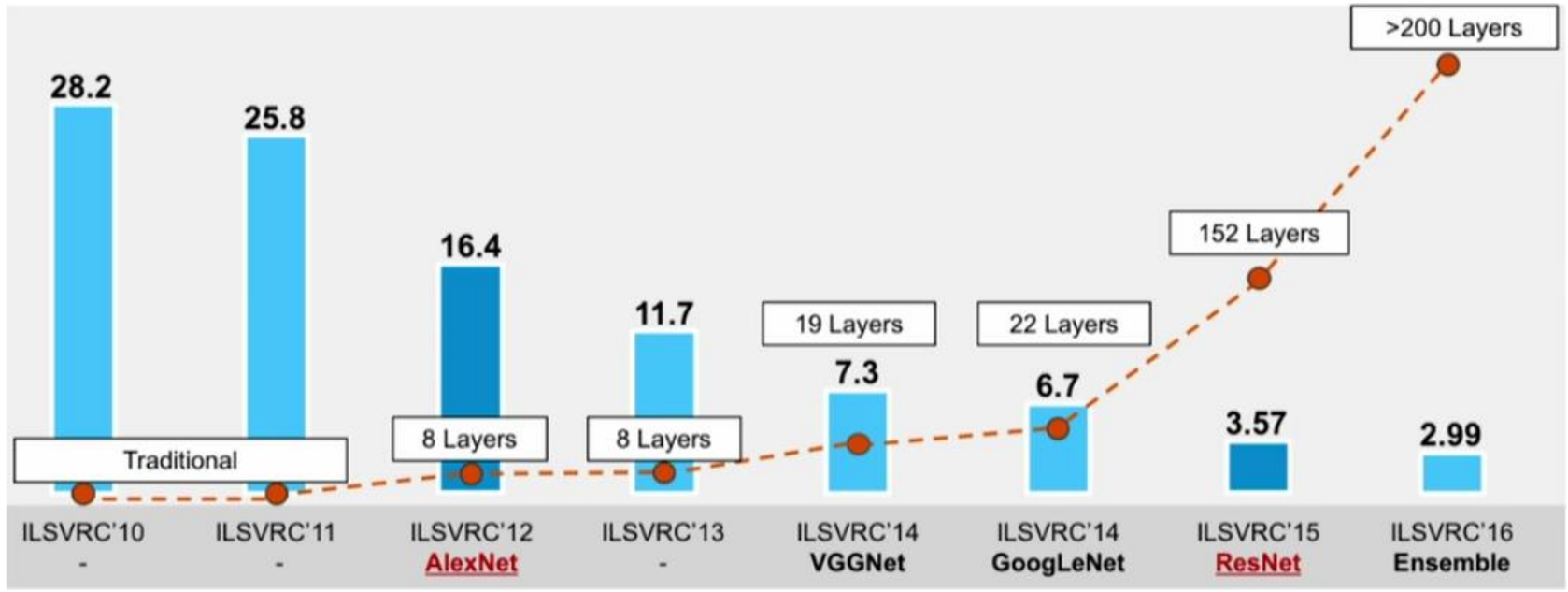
Примеры сложно структурированных объектов:
тексты, изображения, видео, временные ряды, транзакции, графы, ...

Глубокие свёрточные нейронные сети
для классификации изображений



Что такое «большие данные». Пример

ImageNet: открытая выборка 14M изображений, 20K категорий



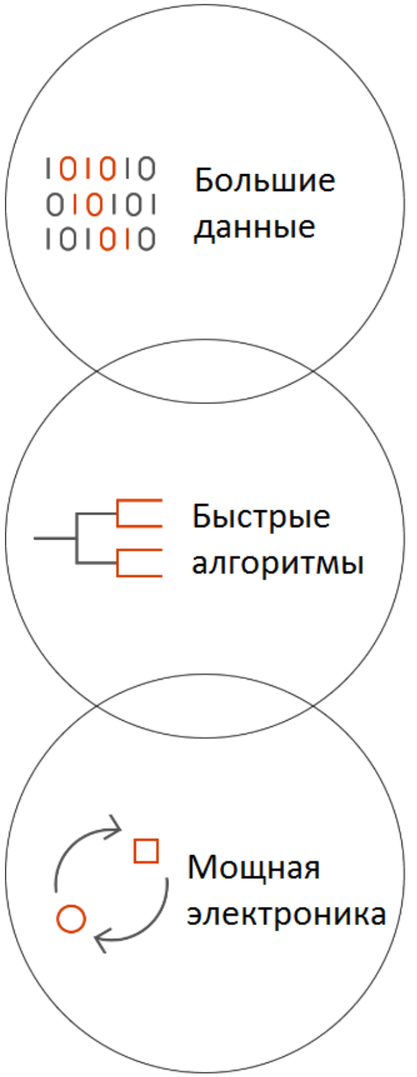
Старт в 2009 г. Человеческий уровень ошибок 5% пройден в 2015 г.

Три составляющих успеха Deep Learning

- Повсеместное применение компьютерных технологий
  → *накопление больших выборок данных*
  в частности, ImageNet

- Развитие математических методов и алгоритмов
  → *накопление критической массы опыта*
  методы оптимизации, контроль переобучения

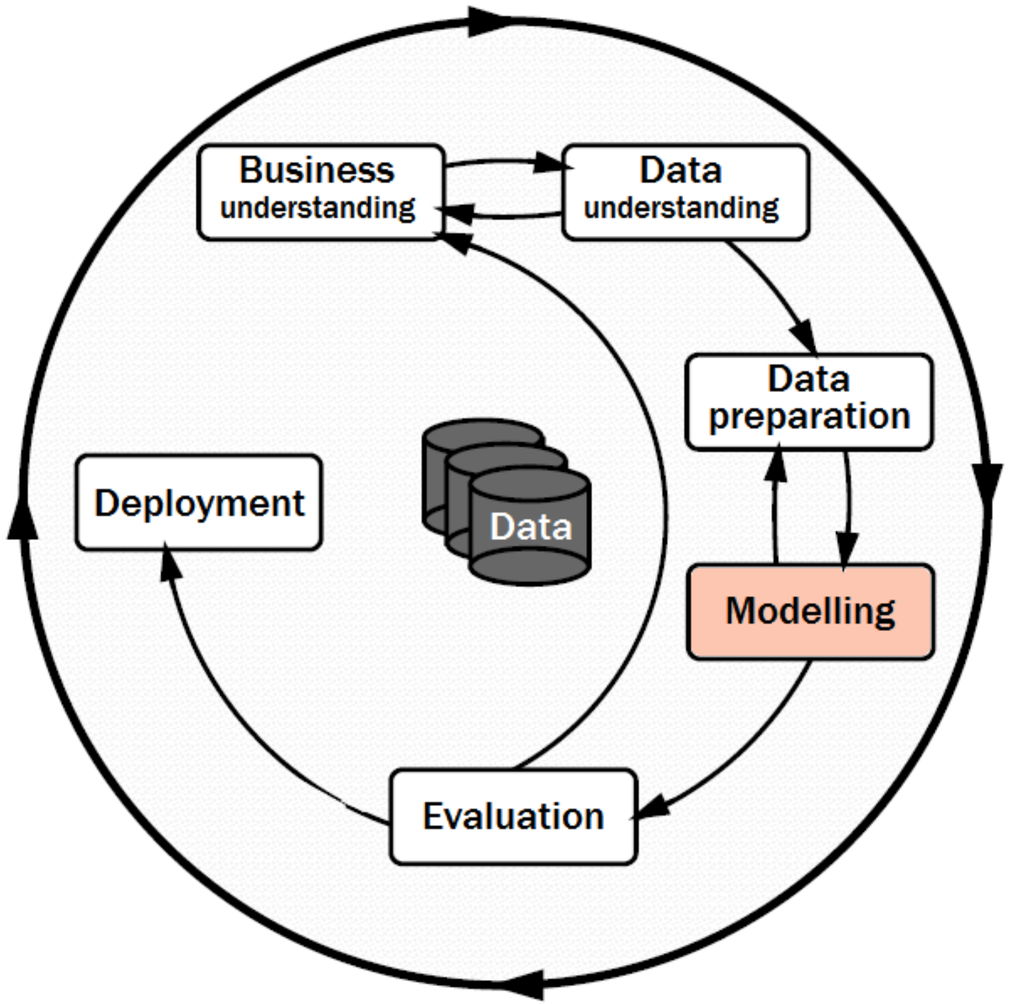
- Достижения микроэлектроники
  → *рост вычислительных мощностей по закону Мура*
  в частности, GPU



Этапы решения задач ML и их автоматизация

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

(SPSS, Teradata, Daimler AG, NCR Corp., OHRA)



- понимание прикладной задачи
- понимание данных
- конструирование признаков
  → обучаемая векторизация (DL)
- обучаемые модели (ES → ML)
- оценивание решения
  → AutoML
- внедрение и эксплуатация
  → бесшовная эволюция моделей (RL)

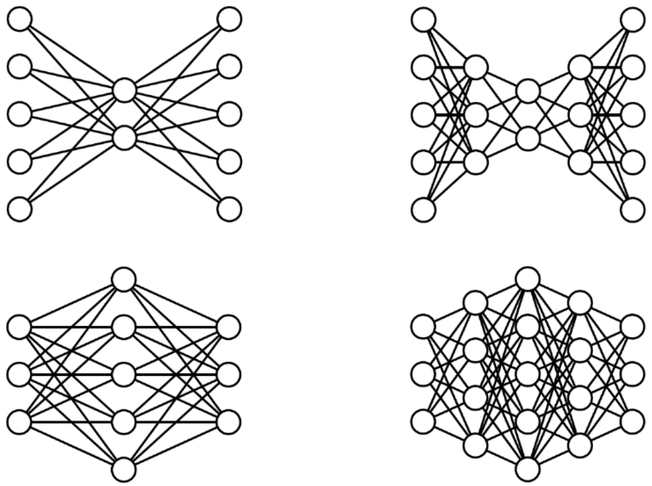
Обучаемая векторизация (autoencoders)

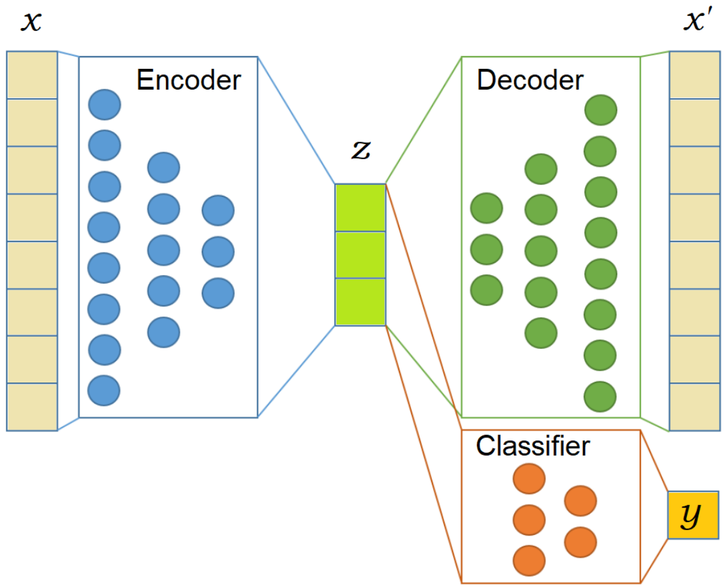
x – вектор объекта обучающей выборки, ответов не дано

$z = f(x, w)$ – модель кодирования x в векторное представление z

$x' = g(z, w')$ – модель декодирования z в реконструкцию x'

$\text{Loss}(x, w) = \|g(f(x, w), w') - x\|$ – точность реконструкции объекта





Предобучение (transfer learning)

$z = f(x, w)$ – часть модели, универсальная для широкого класса задач

$y = g(z, w')$ – часть модели, специфичная для своей задачи

$\min\limits_{w,\, w'}: \sum_x \text{Loss}_1(g_1(f(x, w), w'))$ – обучение по большим данным

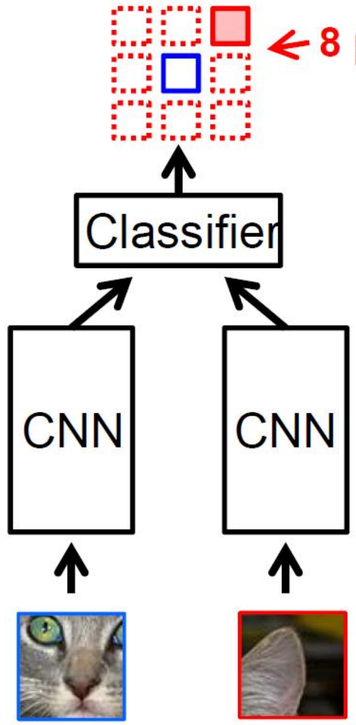
$\min\limits_{w'} \sum_{x'} \text{Loss}_2(g_2(f(x', w), w'))$ – обучение по своим данным

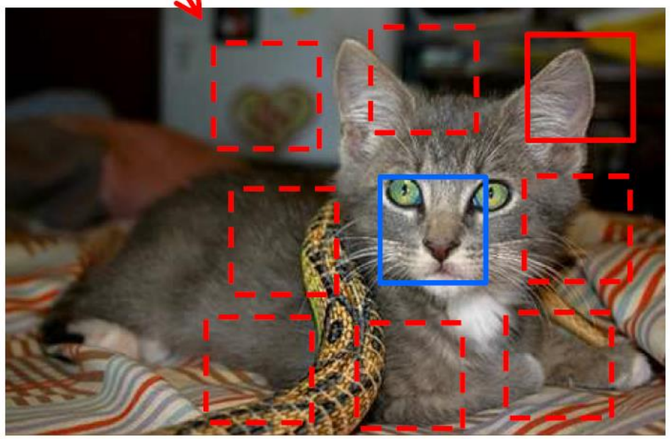
Sinno Jialin Pan, Qiang Yang.
A Survey on Transfer Learning. 2009

Самостоятельное обучение (self-supervised)

Модель векторизации $z = f(x, w)$ обучается предсказывать взаимное расположение пар фрагментов одного изображения

Преимущество:
сеть выучивает векторные представления объектов без размеченной обучающей выборки





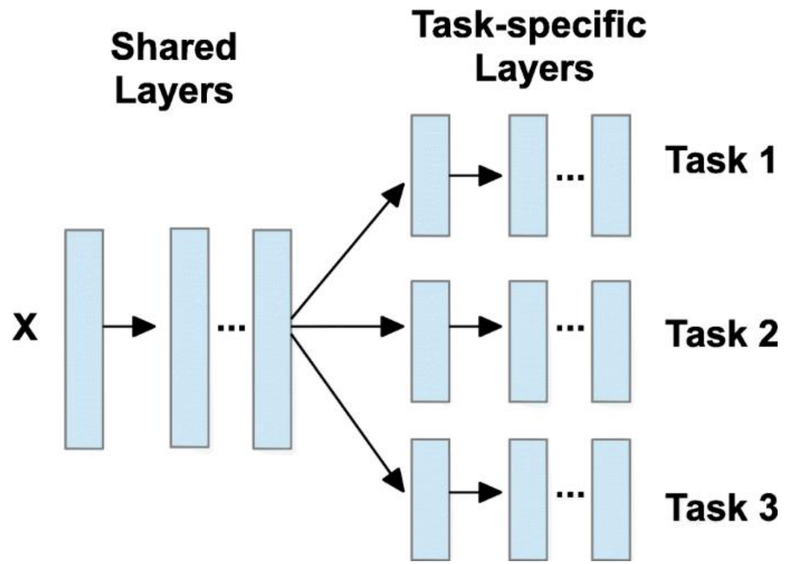
Unsupervised visual representation learning by context prediction, Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Многозадачное обучение (multi-task learning)

$z = f(x, w)$ – модель векторизации, универсальная для всех задач

$y = g_t(z, w'_t)$ – часть модели, специфичная для t-й задачи

$\min\limits_{w, w'_t}: \sum_t \sum_x \text{Loss}_t(g_t(f(x, w), w'_t))$ – обучение по всем задачам



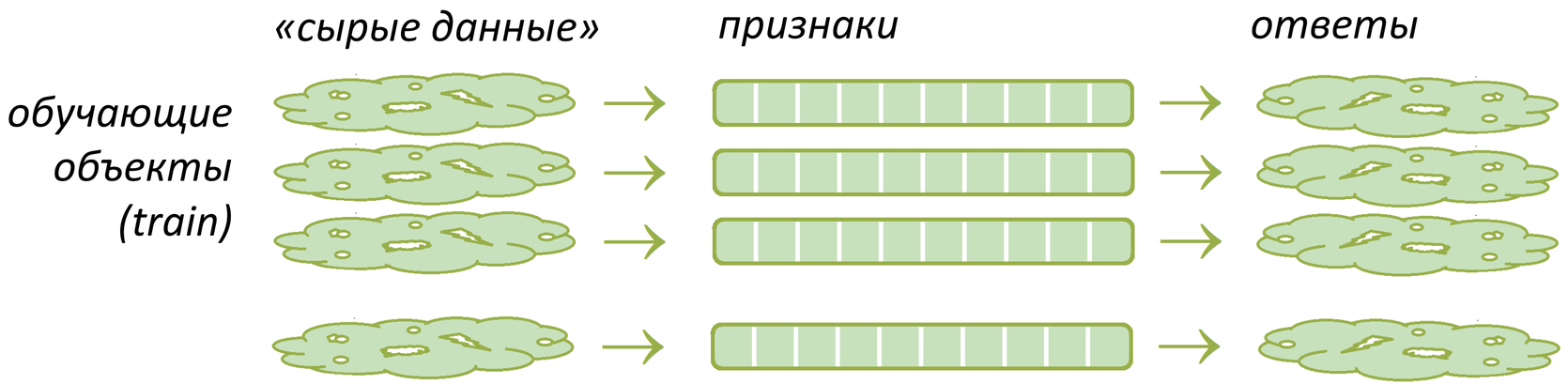
M.Crawshaw. Multi-task learning with deep neural networks: a survey. 2020

Y.Wang et al. Generalizing from a few examples: a survey on few-shot learning. 2020

Нейронные сети для синтеза объектов

Вход: сложно структурированные объекты

Выход: сложно структурированные ответы



Примеры: синтез изображений и видео, перенос стиля, чат-боты, машинный перевод, ответы на вопросы, суммаризация текстов,…

Модели: seq2seq, CNN, RNN, LSTM, GAN, BERT, GPT-3 и др.

Генеративная состязательная сеть (GAN)

$x = g(z, w)$ – модель генерации реалистичного объекты x из шума z

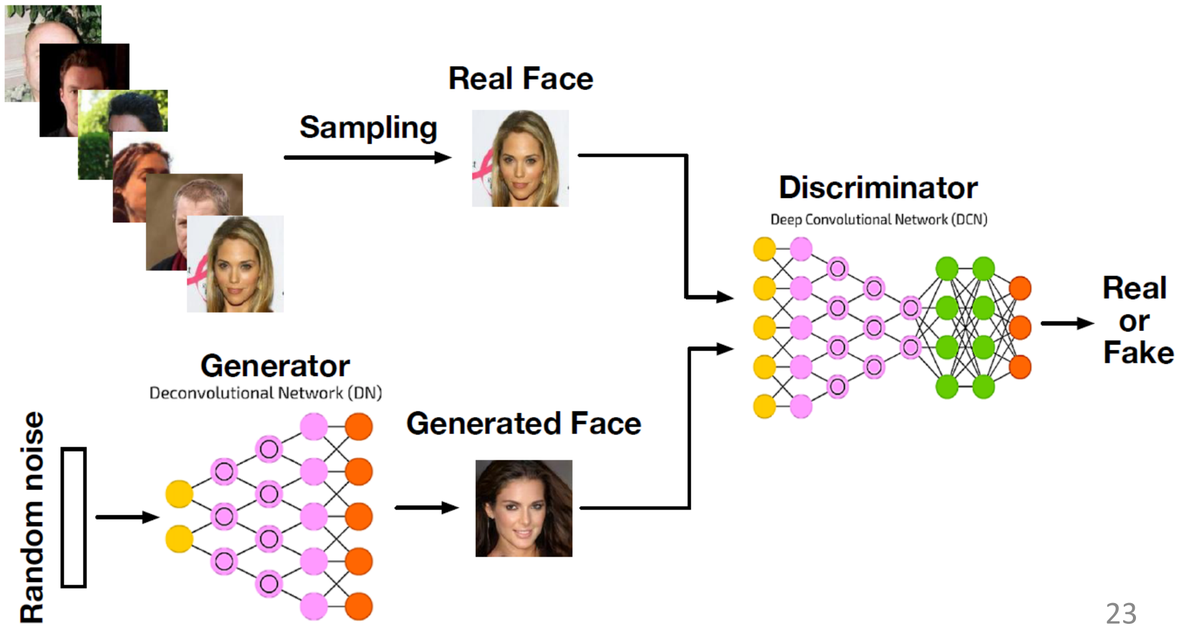
$f(x, w')$ – модель классификации x «реальный/сгенерированный»

$\min_{w} \max_{w'} \sum_x \ln f(x, w') + \ln\left(1 - f(g(z, w), w')\right)$ – совместное обучение

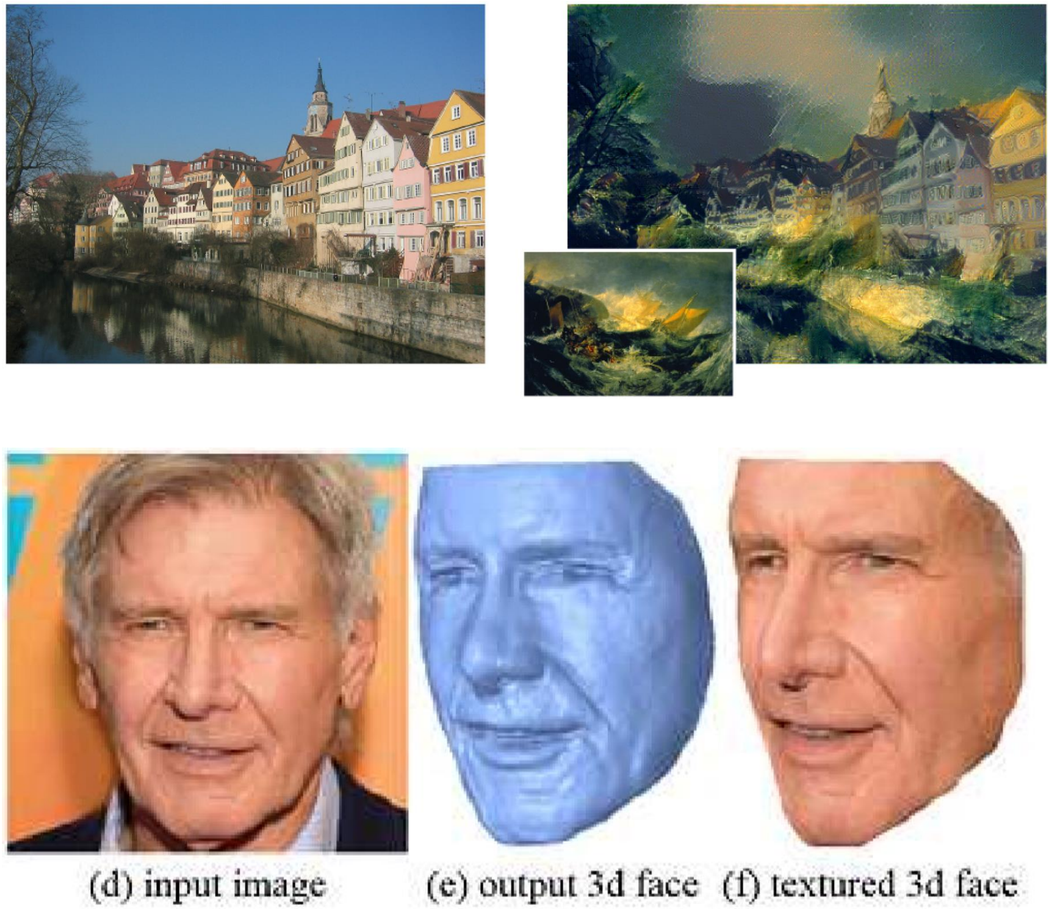
Antonia Creswell et al. Generative Adversarial Networks: an overview. 2017.

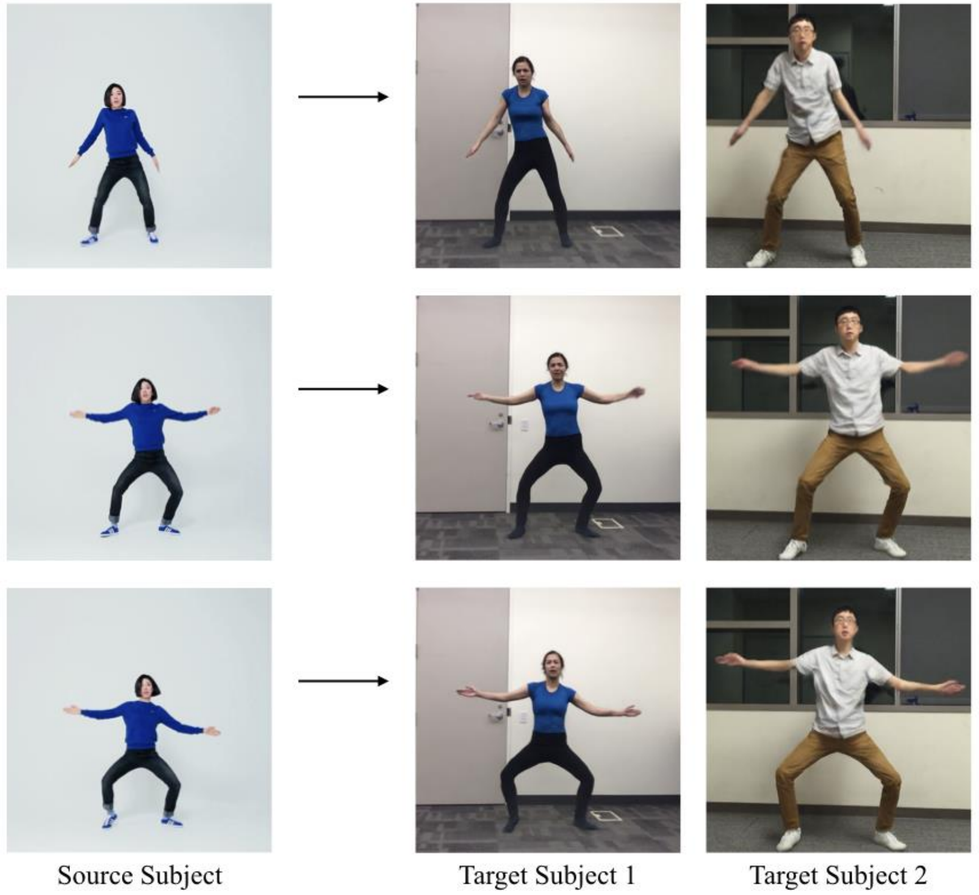
Zhengwei Wang et al. Generative Adversarial Networks: a survey and taxonomy. 2019.

Chris Nicholson. A Beginner's Guide to Generative Adversarial Networks. 2019.



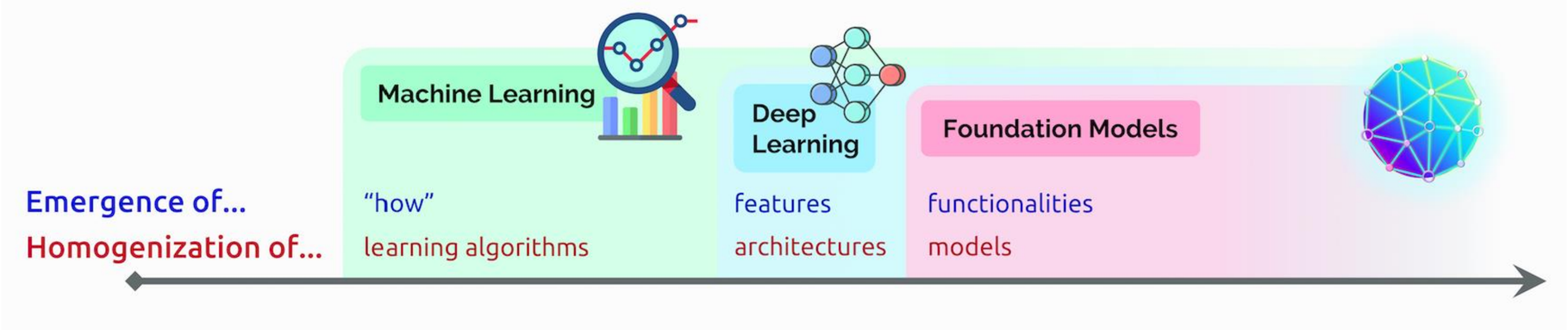
Синтез изображений и видео

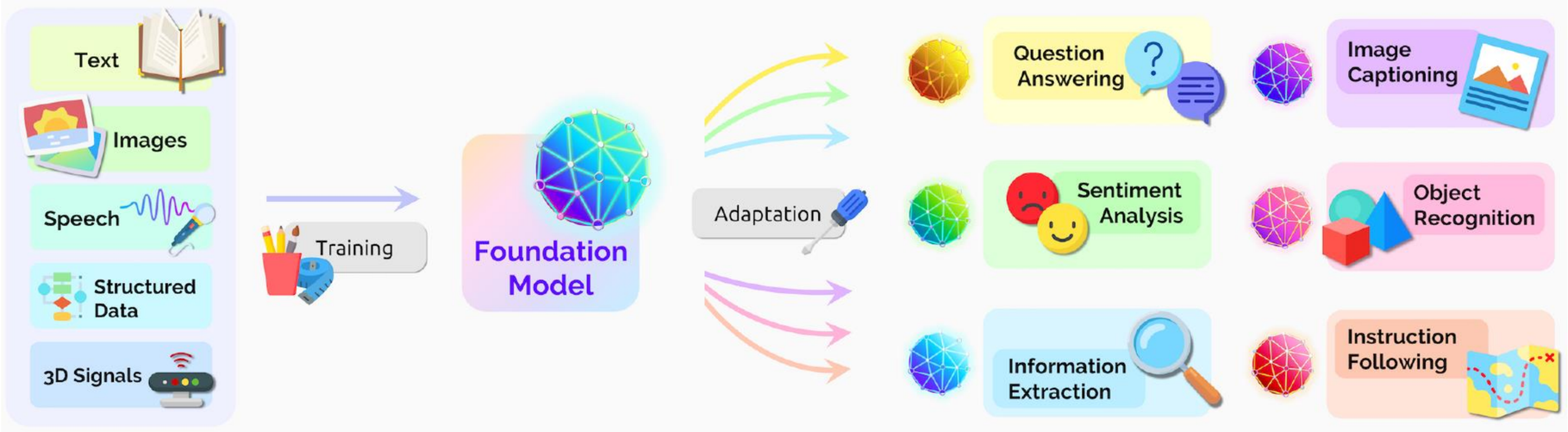




Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros. Everybody Dance Now. ICCV-2019.

Фундаментальные модели (Foundation Models)





R.Bommasani et al. (CRFM, Stanford University). On the opportunities and risks of foundation models. 2021

Искусственный интеллект: мифы, реальность, перспективы

1. Задачи машинного обучения
   - Бум искусственного интеллекта и нейронных сетей
   - Классические задачи машинного обучения
   - Обучаемая векторизация сложно структурированных данных
2. Задачи обработки и понимания естественного языка
   - Задачи разметки текста
   - Задача конкурса ПРО//ЧТЕНИЕ
   - Модели внимания и трансформеры
3. Задачи понимания общественно-политического дискурса
   - Языковые явления эпохи постправды
   - Задачи разметки текста
   - Детектирование пропаганды, манипуляций, поляризации мнений

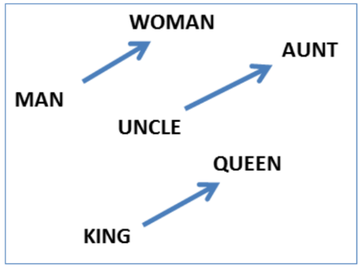
Эволюция подходов в обработке текстов

Декомпозиция задач по уровням «пирамиды NLP»

- морфологический анализ, лемматизация, опечатки, …
- синтаксический анализ, выделение терминов, NER, …
- семантический анализ, выделение фактов, тем, …

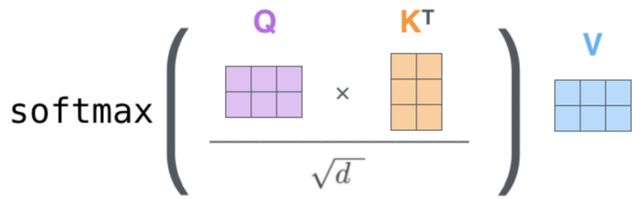
Модели векторизации слов (эмбедингов)

- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016], …
- тематические модели LDA [Blei, 2003], ARTM [2014], …



Нейросетевые модели контекстной векторизации

- рекуррентные нейронные сети: LSTM, GRU, …
- «end-to-end» модели внимания и трансформеры:
  машинный перевод [2017], BERT [2018], GPT-3 [2020], …



Примеры задач разметки текстов в NLP/ML

- распознавание номинативов (named entity recognition, NER)
- распознавание частей речи (part of speech tagging, POS)
- выделение тональности номинатива (sentiment analysis, SA)
- выделение синтаксических связей (syntax parsing)
- выделение семантических ролей (semantic role labeling, SRL)
- выделение текстовых полей данных (slot filling)
- выделение полей в библиографических записях
- сегментация научных или юридических текстов
- разрешение анафоры, кореферентности, эллипсиса

Выделение и тегирование фрагментов текста

Нотация BIOES (begin-inside-outside-end-single) для выделения начала и конца фрагмента

Для задачи распознавания именованных сущностей:

B-PER	I-PER	I-PER	I-PER	E-PER	OUT	OUT	S-LOC
Карл	Фридрих	Иероним	фон	Мюнхгаузен	родился	в	Боденвердере

Для задачи определения семантических ролей:

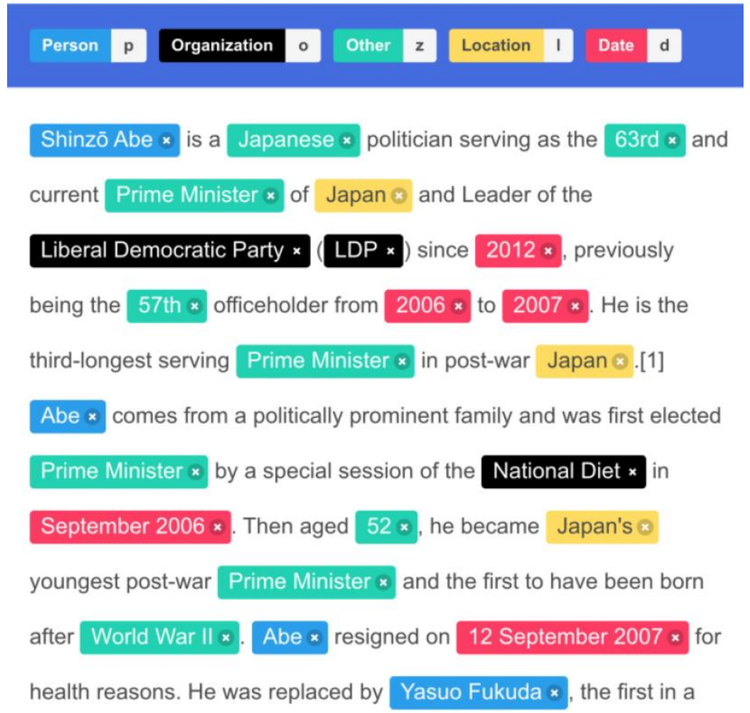
B_ACT	I_ACT	I_ACT	O	B_NUM_PER	O	B_LOC	I_LOC
Book	a	table	for	3	in	Domino's	pizza

Пример: разметка номинативов (NER)

Named entity — объект (сущность) реального мира, имеющий наименование и относящийся к определённой категории.

Примеры категорий
(зависят от предметной области):

- персона, организация, локация, время
- ссылка на нормативно-правовой акт
- заболевание, симптом, препарат
- биологический вид
- астрономический объект



Пример: разметка семантических ролей (SRL)

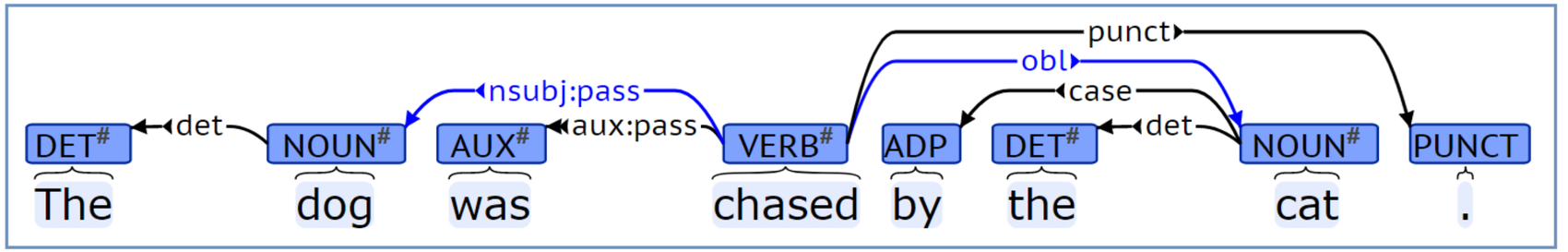
Задача: найти в предложении *актанты* — именные группы, обозначающие участников ситуации и их *семантические роли.*

Примеры семантических ролей:

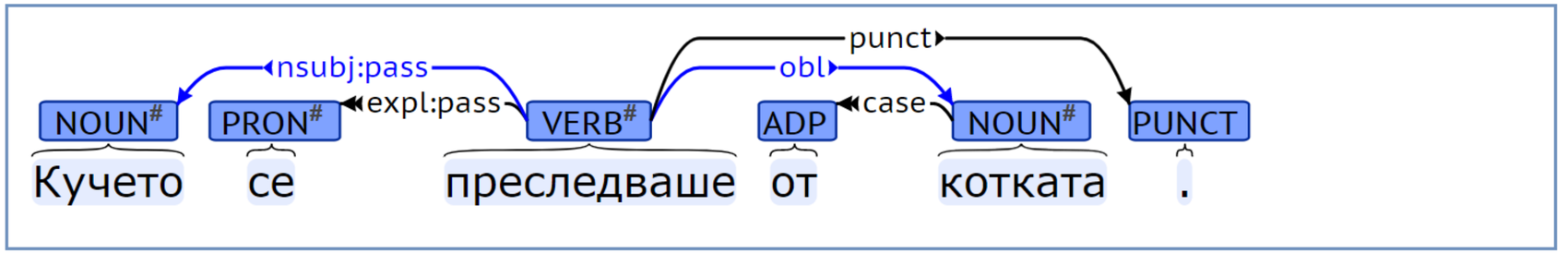
- **агенс**: одушевлённый инициатор и контролёр действия
- **пациенс**: участник, на которого направлено действие
- **бенефактив**: участник, получающий пользу или вред
- **адресат**: получатель сообщения (может быть бенефактивом)
- **инструмент**: посредством чего осуществляется действие
- **экспериенцер**: носитель чувств и восприятий
- **стимул**: источник восприятий
- **источник**: исходный пункт движения
- **цель**: конечный пункт движения

Пример: частеречная и синтаксическая разметка

английский:



болгарский:



теги частей речи

NOUN	noun	существительное	INTJ	interjection	междометие
PROPN	proper noun	имя собственное	ADP	adposition	предлог
ADJ	adjective	прилагательное	CONJ	conjunction	союз
VERB	verb	глагол	PART	particle	частица
ADV	adverb	наречие	PUNCT	punctuation	знак пунктуации
PRON	pronoun	местоимение	SYM	symbol	символ
NUM	numeral	числительное	X	other	иное

Конкурс ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Задача: разметка смысловых ошибок в сочинениях ЕГЭ по русскому языку, литературе, истории, обществознанию и английскому языку.

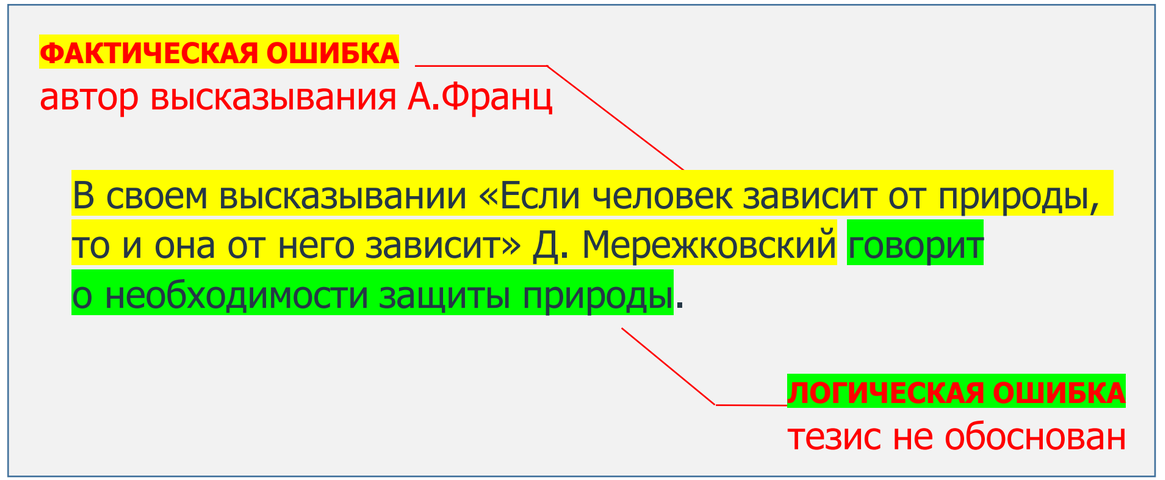
Период: декабрь 2019 — июнь 2022, три цикла испытаний.

Призовой фонд: 100М русский язык + 100М английский язык

Типов ошибок: 152
(р:70 л:16 о:23 и:20 а:23)

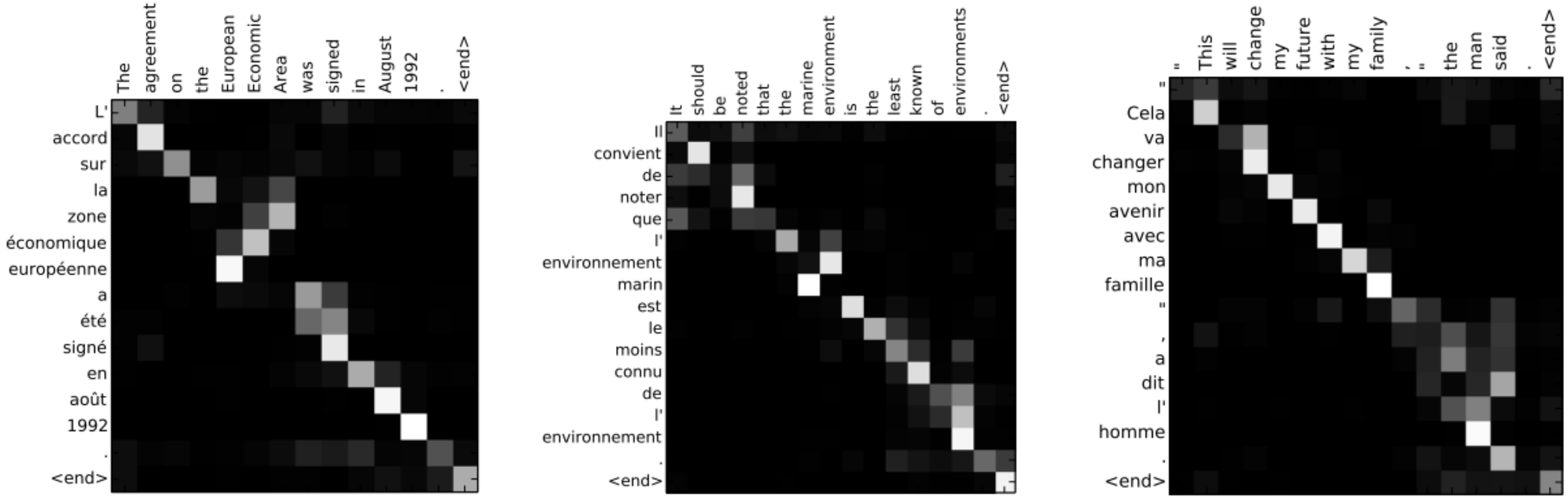
Подтипов ошибок: 236
(р:112 л:19 о:29 и:26 а:50)

Помимо выделения ошибок, надо давать их объяснения.



Конкурс ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Модели внимания: машинный перевод



Интерпретация моделей внимания: *матрица семантического сходства* A[t,i] показывает, на какие слова x[i] входного текста модель обращает внимание, когда генерирует слово перевода y[t]

Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015

Модели внимания: аннотирование изображений

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

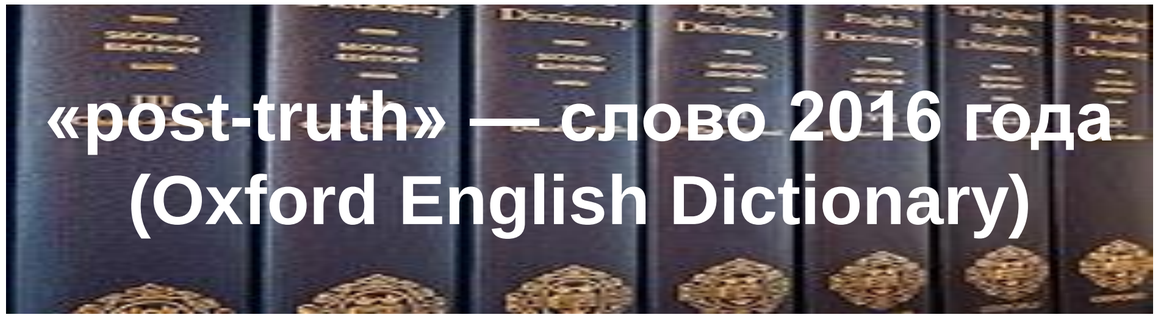
Интерпретация: на какие области модель обращает внимание, генерируя подчёркнутое слово в описании изображения

Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

Искусственный интеллект: мифы, реальность, перспективы

1. Задачи машинного обучения
   - Бум искусственного интеллекта и нейронных сетей
   - Классические задачи машинного обучения
   - Обучаемая векторизация сложно структурированных данных
2. Задачи обработки и понимания естественного языка
   - Задачи разметки текста
   - Задача конкурса ПРО//ЧТЕНИЕ
   - Модели внимания и трансформеры
3. Задачи понимания общественно-политического дискурса
   - Языковые явления эпохи постправды
   - Задачи разметки текста
   - Детектирование пропаганды, манипуляций, поляризации мнений

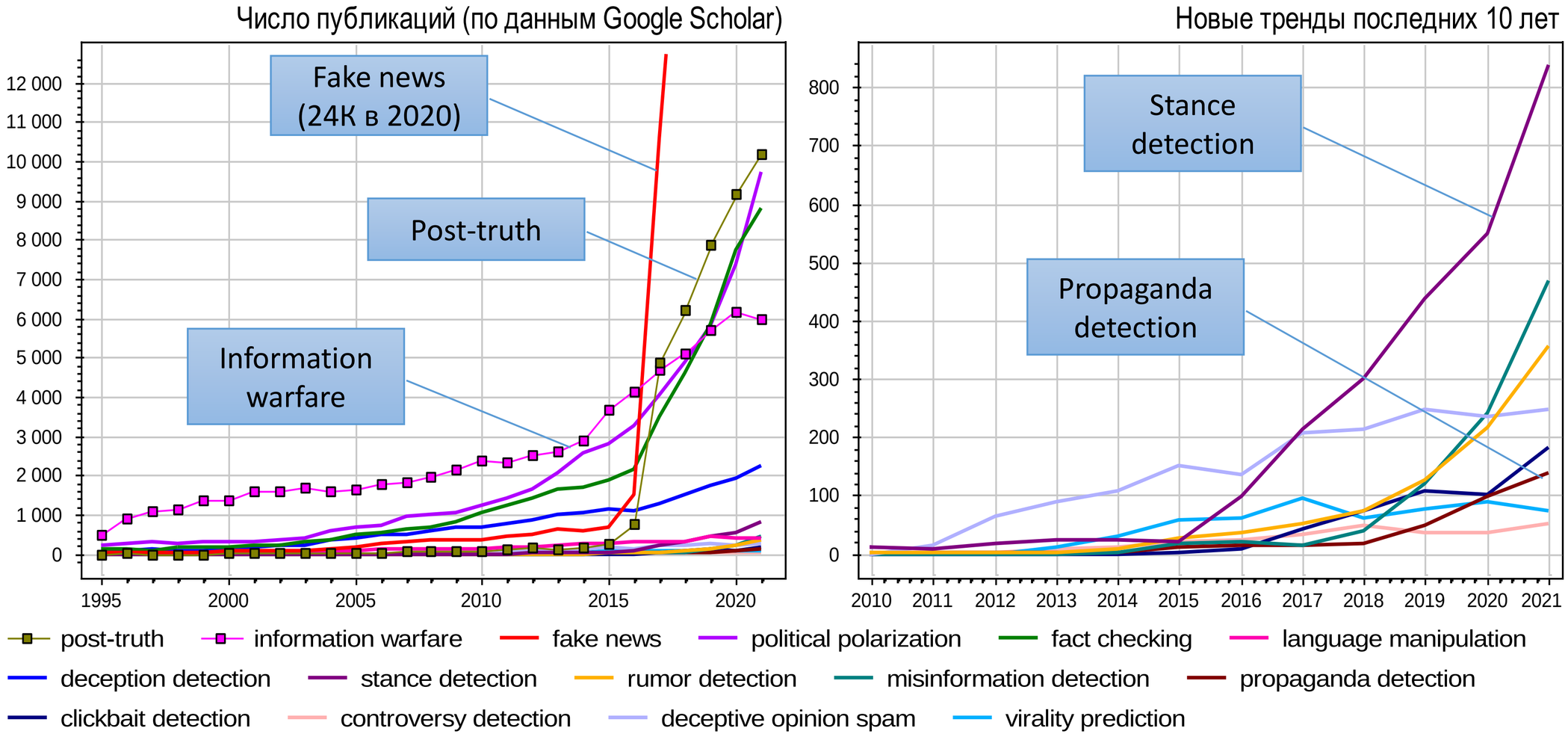
Постправда и информационные войны



- Факты становятся менее значимы, чем эмоции и личные убеждения
- Явление «информационных пузырей»
- Явление «неопровержимой лжи»
- Постправда маскируется под «другие грани истины»
- Постправда — новая форма пропаганды и инструмент «мягкой силы»

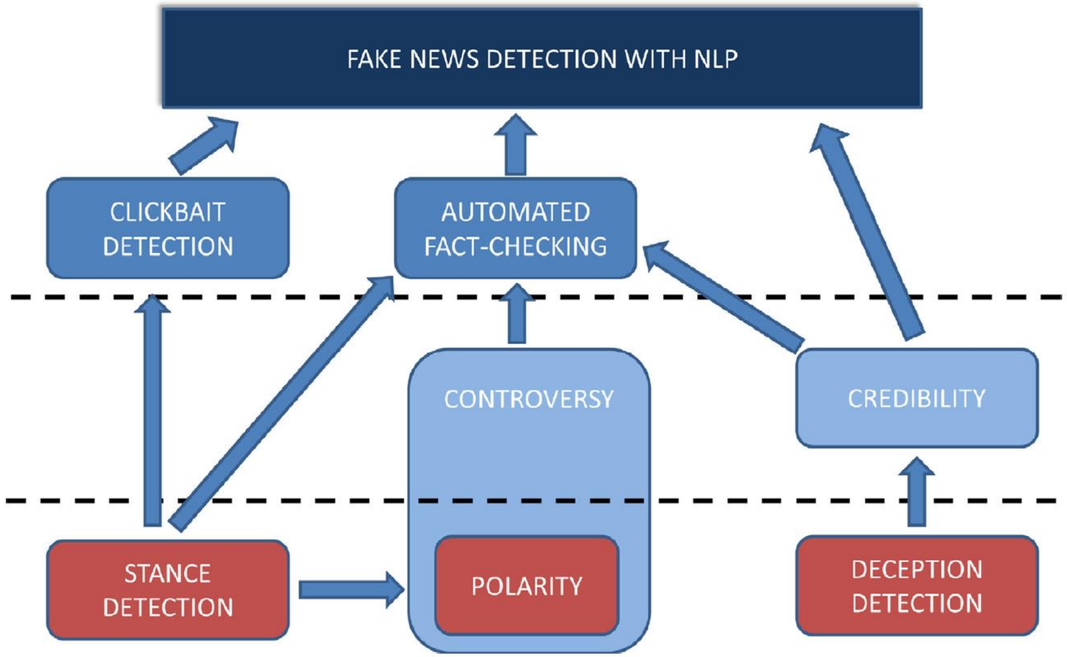
Fake News и смежные области исследований

(библиометрический анализ по данным Google Scholar)



Область исследований «Fake News Detection»

1. Deception Detection
   выявление обмана в тексте новости
2. Automated Fact-Checking
   автоматическая проверка фактов
3. Stance Detection
   выявление позиции за/против запроса (claim)
4. Controversy Detection
   выявление и кластеризация разногласий
5. Polarization Detection
   классификация позиций по многим темам
6. Clickbait Detection
   выявление противоречий заголовка и текста
7. Credibility Scores
   оценка достоверности источника или новости



E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems With Applications, Elsevier, 2020.

1. Deception Detection (выявление обмана)

- **История:** более 50 лет исследований в психологии и криминологии
- **Задача** классификации текста на два класса: *обман / не обман*
- **Обучающие выборки:**
  - Контролируемый эксперимент: люди *врут / не врут* на заданную тему
  - Материалы судебных заседаний (датасет DECOUR)
  - Отзывы на товары/услуги, проверяемые с помощью краудсорсинга
- **Признаки** – лингвистические маркеры (Linguistic-Based Cues, LBC)
- **Критерии:** Accuracy или F-мера 70–92% в зависимости от задачи
- На небольших датасетах классический ML лучше и проще DL
- Проблема переноса моделей на другие датасеты

Типы лингвистических маркеров

Манипулятивные и суггестивные приёмы

- многословие: плеоназмы, лишние слова, тавтологии, расщепления сказуемого
- избыточные повторы слов и фраз
- повышенная когнитивная сложность текста, перегруженные синтаксические конструкции
- повышенная экспрессивность, преобладание негативной тональности
- категоричность, психологическое давление

Уход от личной ответственности

- безличные глаголы, глаголы абстрактной семантики, модальные глаголы, объективация
- неконкретность, уклончивость, безличность, неопределённость высказываний

Подача информации

- оторванность от контекста: пониженная детализация места, времени, событий
- упрощение, пониженное лексическое разнообразие, лексическая недостаточность
- замалчивание фактов, сообщение ложных сведений (fact-checking, см. далее)

2. Automated Fact-Checking (проверка фактов)

- **История:** ручной fact-checking давно используется в журналистике
- **Задача** классификации текста целиком, по порядковой шкале:
  True, Mostly True, Half True, Mostly False, False
- **Обучающие выборки:**
  - Платформы для проверки фактов: Politifact, FullFact, FactCheck и др.
  - Соревнования: CLEF-2018,19,20,21, FEVER, SemEval (Rumour-Eval)
  - Датасеты: NELA-GT-2018,19, FakeNewsNet, Snopes и др.
- **Вспомогательная задача:** стоит ли отправлять текст на проверку?
  Три класса: *Non-Factual Sentence, Unimportant, Check-Worthy*
  (пример: ClaimBuster, https://idir.uta.edu/claimbuster, 2015)

3. Stance Detection (выявление позиции)

- **История:** задача textual entailment (текстового следования) – классификация пар текстов «текст *t* ⇒ гипотеза *h*» на три класса: «*h* следует из *t*», «*h* противоречит *t*», «*h* не относится к *t*»
- **Задача:** классификация текста *h* относительно запроса (claim) *t*: *agree, disagree, discusses (позиция не высказана), unrelated*
- **Обучающие выборки:**
  - SNLI: 570K пар предложений: entail, contradict, independent
  - Датасеты: Emergent, SemEval-2016 6A(stance), FakeNewsChellenge FNC-1
- **Критерии:** F1-мера до 97% на новостях; Accuracy до 68% на Twitter

4. Controversy / 5. Polarization Detection

Две специальные разновидности задачи Stance Detection

- **Controversy Detection** (выявление полемики, разногласий):
  - кластеризация мнений без учителя
  - выделение сообществ сторонников каждого мнения в социальной сети
  - количественное оценивание объёма и динамики сообществ
- **Polarization Detection** (выявление поляризованности общества):
  - выявление разногласий по совокупности запросов или тем
- **Обучающие выборки:**
  - Датасеты социальных сетей, обычно Twitter
  - Википедия
- **Критерии:** Accuracy 73–83% (на Википедии, методом kNN)

6. Clickbait Detection (обнаружение кликбейта)

- **История:** задача появилась в 2016 году. Обнаружение заголовков или ссылок-приманок, не соответствующих сути контента
- **Задача:** классификация пары «заголовок, текст» на два класса Задача аналогична Textual Entailment и Stance Detection
- **Признаки:** гиперболизация, противоречия, web-трафик
- **Обучающие выборки:**
  - Датасеты: Webis-Clickbait 2017 (32К заголовков) и др.
  - Соревнование: Clickbait challenge 2017
- **Критерии:** F1-мера до 68%; Accuracy до 86%

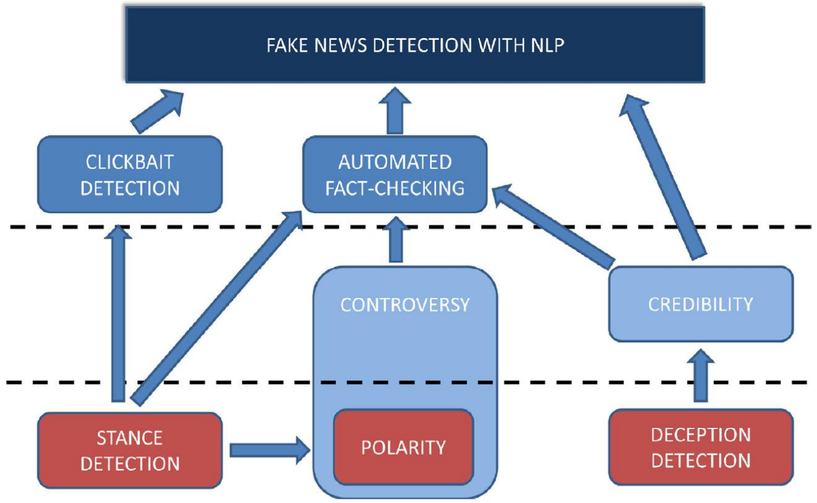
7. Credibility Scores (Оценивание надёжности)

- **История:** старая задача в социологии, психологии, маркетинге
- **Задача:** оценить уровень доверия (credibility, trustworthiness) для источника (СМИ, блогера, пользователя) или отдельной новости
- **Признаки:**
  - распространение ненадёжного контента (spam, deception, fake и др.)
  - вероятность быть ботом (по диспропорции рассылок и качеству контента)
  - стиль контента, геолокация и образовательный уровень читателей
- **Обучающие выборки:**
  - много несопоставимых датасетов, отсутствует «золотой стандарт»
- **Критерии:** AUC до 89%; accuracy до 81%; MSE до 0.33
  - много критериев, не хватает методологического единства

Чего-то не хватает…

1. **Fake News** – не единственный и не самый сильный инструмент политики постправды.
2. **Пропаганда** использует не только фейки, но и полуправду, замалчивание, манипулятивные воздействия и т.д.
3. **Информационные войны** нацелены на разрушение социокультурного кода и сложившейся общественной идеологии.

- Как распознавать манипулятивные воздействия и идеологические атаки?
- Как находить разногласия и замалчивание?
- Насколько расширится типология задач?



E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems With Applications, Elsevier, 2020.

Типология деструктивного дискурса и система подзадач ML/NLP для его детекции

воздействия → фейки → пропаганда → инф.война

1.  детекция приёмов манипулирования
2.  детекция замалчивания
3.  **детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)**
4.  **детекция кликбэйта (clickbait detection)**
5.  **автоматическая проверка фактов (auto fact-checking)**
6.  **детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)**
7.  выявление конструктов картины мира: ценностей, идеологем, мифологем
8.  оценивание возможных психо-эмоциональных реакций
9.  выявление целевых аудиторий воздействия
10.  **оценивание и предсказание скорости распространения (virality prediction)**
11.  **оценивание достоверности источников (credibility scores)**
12.  детекция деструктивных воздействий (угрозы, призывы, провокации, вербовка, экстремизм)

Четыре основных типа подзадач ML/NLP

1. **Классификация текста (сообщения/предложения) целиком**
   - deception detection, fact-checking, text credibility
2. **Классификация пары текстов**
   - stance, controversy, polarization, clickbait detection
   - выявление противоречий, разногласий, замалчивания
3. **Разметка текста (выделение и классификация фрагментов)**
   - поиск лингвистических маркеров (linguistic-based cues) в тексте
   - детекция приёмов манипулирования
   - выявление идеологем, ценностей, элементов социокультурного кода
   - выявление психо-эмоциональных реакций и целевых аудиторий
   - выявление мнений, тональных оценочных суждений
4. **Кластеризация или тематическое моделирование**
   - кластеризация мнений по заданной теме (controversy detection)
   - выявление поляризации общественного мнения (polarization detection)

Выявление пропаганды (propaganda detection)

Чтобы выявлять пропаганду, нужно иметь модель пропаганды:

1. *Подмена и/или дополнение* фактов мнениями
2. *Фрагментирование*: часть фактов замалчивается
3. *Деконтекстуализация*: изымается контекст, без которого корректное понимание смысла фактов невозможно
4. *Реконтекстуализация*: конструируется новый контекст, выгодный манипулятору

Подзадачи ML/NLP:

- Выделение и различение фактов и мнений
- Выявление замалчиваний путём сравнения с другими источниками
- Выявление идеологем, используемых для реконтекстуализации

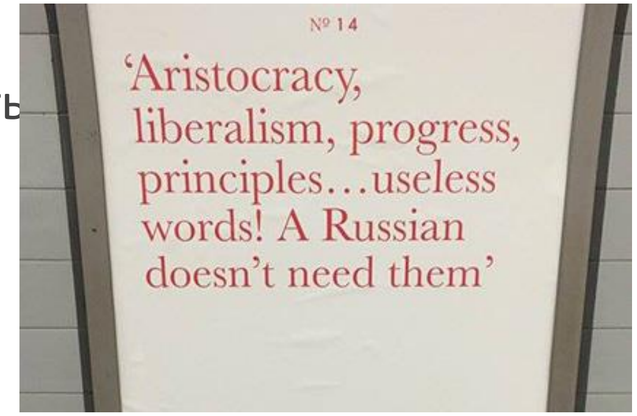
Обучающие выборки:

- Тексты новостей с размеченными фрагментами (факты, мнения, идеологемы)

Пример деконтекстуализации:
Цитаты классиков в лондонском метро

«Аристократизм, либерализм, прогресс, принципы, — ~~говорил между тем Базаров, — подумаешь, сколько иностранных...~~ и бесполезных слов! Русскому человеку они даром не нужны» [И.С.Тургенев, «Отцы и дети»]

«С нашей точки зрения, эти книги должны быть прочитаны во всем мире. Тот факт, что на долю русских писателей приходится такой большой процент наших изданий, свидетельствует о качестве русской литературы. Мы хотели побудить людей самостоятельно искать романы. Замысел кампании в том, чтобы прославить эти чудесные вещи»
— объяснение представителя книжного издательства Penguin



При этом на билборде не указано ни что это Тургенев, ни что эти слова принадлежат литературному герою.

Задача выявления приёмов манипулирования

Структура манипуляции:

- фрагмент-мишень
- фрагмент-воздействие
- тип манипуляции

Пример из СМИ:
«**Зеленский** просто **играет роль президента, а не является президентом**[обесценивание], – считает экс-депутат Верховной рады Борислав Береза»

Типы манипуляций (всего 18 типов):

- негативизация (обесценивание, дисфемизмы, ярлыки, депрессивы и т.п.)
- позитивизация (героизация, эвфемизация, лозунги и т.п.)
- деавторизация (замалчивание источника, маскировка под ссылку и т.п.)
- паралогизация (алогизм, ложное следование, подмена тезиса и т.п.)

Классификация приёмов манипулирования

1. Негативизация

1.1 Навешивания ярлыков
1.2 Дисфемизмы
1.3 Аналогия с негативным объектом
1.4 Антифразис
1.5 Прием обесценивания
1.6 Негативирующая гиперболизация
1.7 Моделирование негативного сценария
1.8 Вкрапление депрессивов

2. Позитивизация

2.1 Эвфемизация
2.2 Лозунговые слова и словосочетания
2.3 Позитивирующая гиперболизация

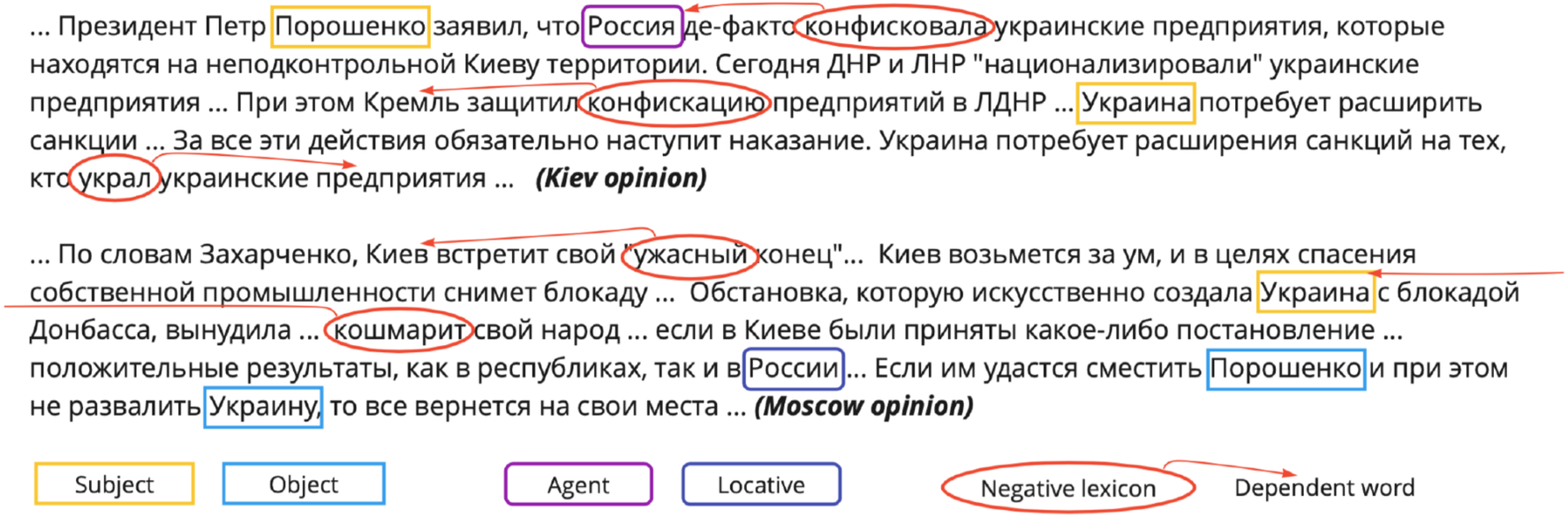
3. Деавторизация

3.1 Маскировка под ссылку на авторитет
3.2 Ссылки на неопределенный источник
3.3 Ссылки на неназванных свидетелей

4. Паралогизация

4.1 Ложная причинно-следственная связь
4.2 Прием «после этого не значит поэтому»
4.3 Подмена тезиса
4.4 Высказывание о состоянии другого

Задача выделения мнений в теме или событии



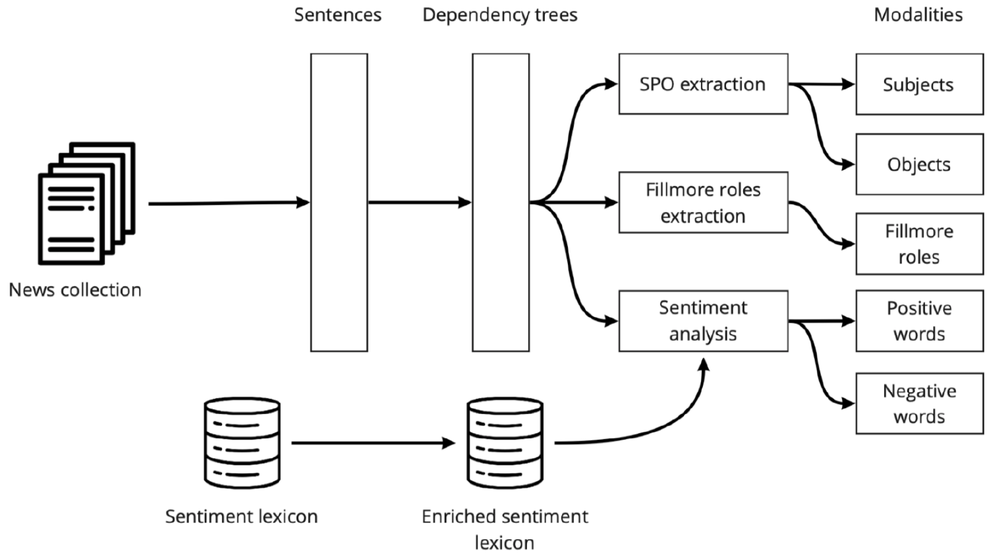
Слова «Порошенко», «Россия», «Украина» встречаются одинаково часто

«Порошенко» — субъект в первом тексте и объект во втором

«Россия» — агенс в первом тексте и локация во втором

Негативная тональность: «Россия», «Кремль» в 1-ом, «Киев», «Украина» во 2-ом

Задача выделения мнений в теме или событии



Modalities	*Pr*	*Rec*	*F*1
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

LPR Business

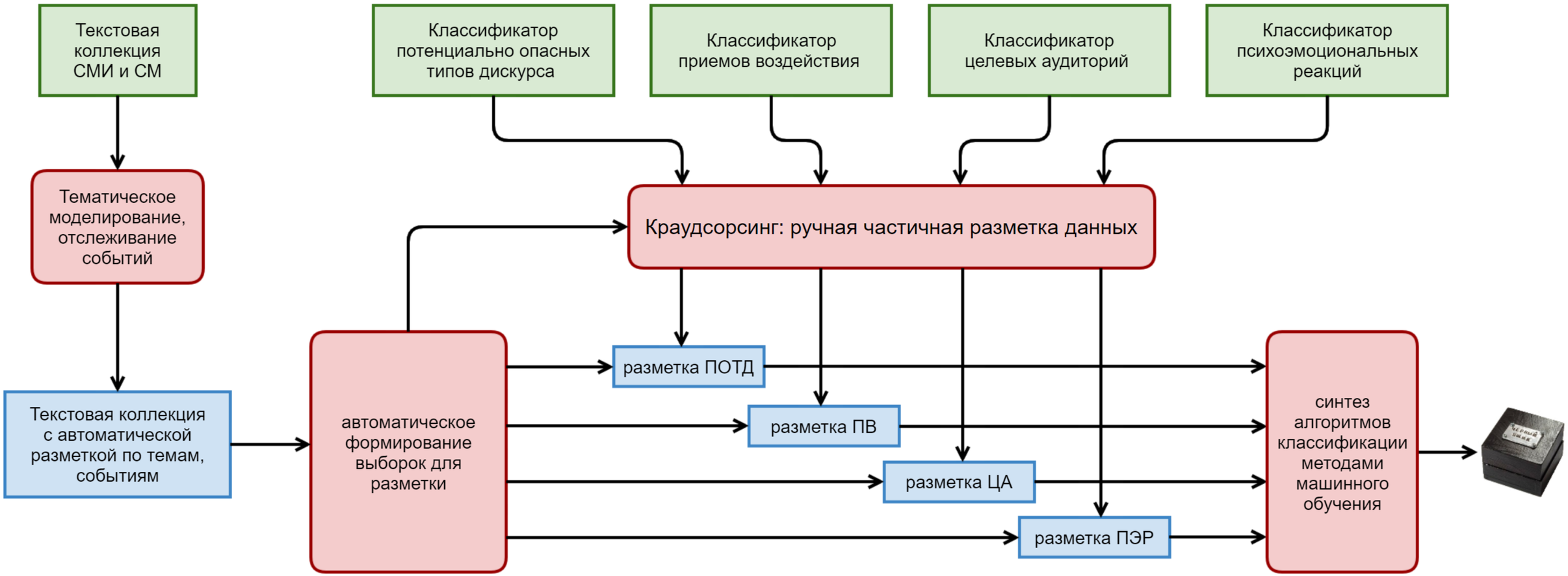
Modalities	*Pr*	*Rec*	*F*1
TF-IDF	0.57	0.97	0.72
SPO	0.56	0.99	0.72
FR	0.67	0.97	0.79
Sent	0.56	0.55	0.55
SPO+FR	0.72	0.99	0.83
SPO+Sent	0.57	0.99	0.72
FR+Sent	0.73	0.97	0.83
All	**0.77**	**0.94**	**0.85**

Paris Trump

Мнение формализуется как устойчивое сочетание слов, терминов, именованных сущностей, их семантических ролей по Филлмору и их тональных окрасок.
Все они используются в модели тематической векторизации как модальности.

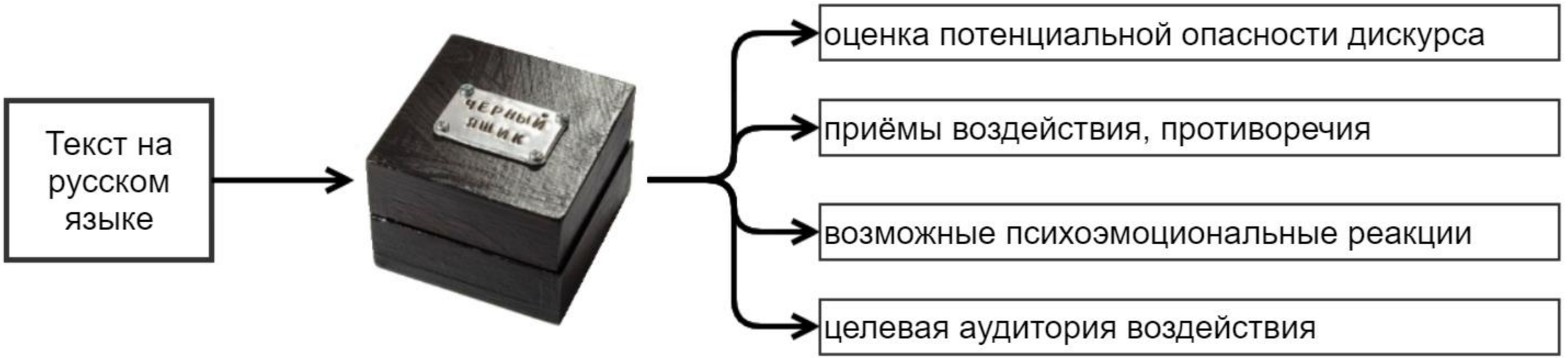
Feldman D. G., Sadekova T. R., Vorontsov K. V. Combining Facts, Semantic Roles and Sentiment Lexicon in A Generative Model for Opinion Mining. Dialogue 2020.

Разметка текстовых данных — магистральный путь формализации гуманитарных знаний



На выходе — модель классификации угроз в медийном информационном пространстве

Модель, обученная по размеченным обучающим выборкам,
может быть использована в автоматическом режиме для мониторинга и фильтрации деструктивного дискурса в информационном пространстве



Выводы

1. ML — это оптимизация параметров предсказательных моделей
2. ИИ — не «интеллект», а обучаемая векторизация данных
3. Перспективы развития ИИ — автоматизация процесса CRISP-DM
4. Предобученные модели внимания / трансформеры позволяют теперь решать сложные задачи понимания естественного языка
5. В том числе стоить модели для мониторинга и детекции угроз в медийном информационном пространстве
6. Разметка текстовых данных — магистральный путь формализации гуманитарных знаний в таких задачах
7. Методология разметки и оценивания идёт к стандартизации

Спасибо за внимание!

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН,
руководитель лаборатории МОСА
(Машинного Обучения и Семантического Анализа)
Института Искусственного Интеллекта МГУ
voron@mlsa-iai.ru

http://www.MachineLearning.ru/wiki?title=User:Vokov