ВЗЛЁТ
Образовательный центр
Гимназии им. Е. М. Примакова

Технологии искусственного интеллекта и политика постправды

Воронцов Константин Вячеславович

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

Технологии ИИ, которые меняют мир

Яндекс Найти

Google

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

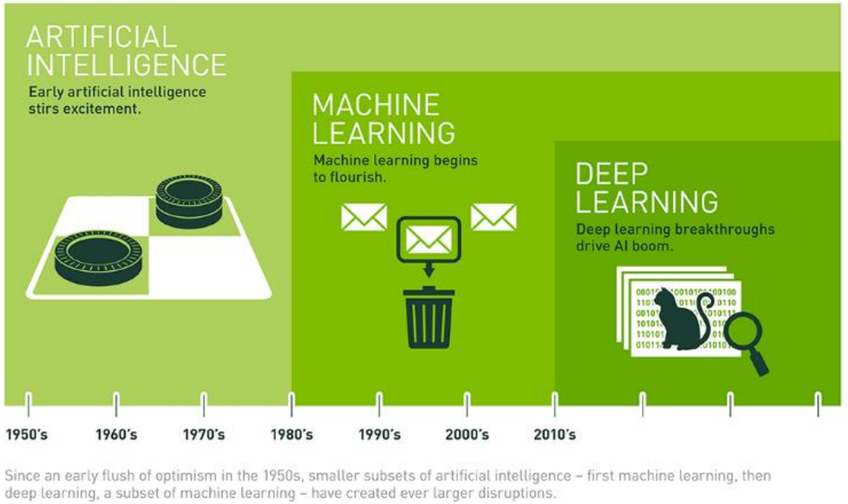
Клаус Мартин Шваб,
президент Всемирного
экономического форума

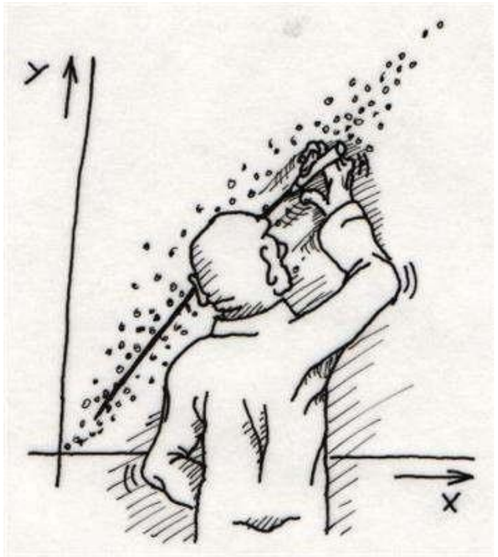
Машинное обучение, большие данные «и много других страшных слов»

- Статистический анализ данных (Statistical Data Analysis)
- Искусственный интеллект (Artificial Intelligence) 1955
- Распознавание образов (Pattern Recognition)
- Машинное обучение (Machine Learning) 1959
- Статистическое обучение (Statistical Learning)
- Интеллектуальный анализ данных (Data Mining) 1989
- Машинный интеллект (Machine Intelligence) 2000
- Бизнес-аналитика (Business Intelligence, Business Analytics)
- Предсказательная аналитика (Predictive Analytics) 2007
- Большие данные (Big Data) 2008
- Аналитика больших данных (Big Data Analytics)
- Наука о данных (Data Science) 2011

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
- около 100 000 научных публикаций в год

Задача машинного обучения с учителем

Этап №1 – обучение (train)

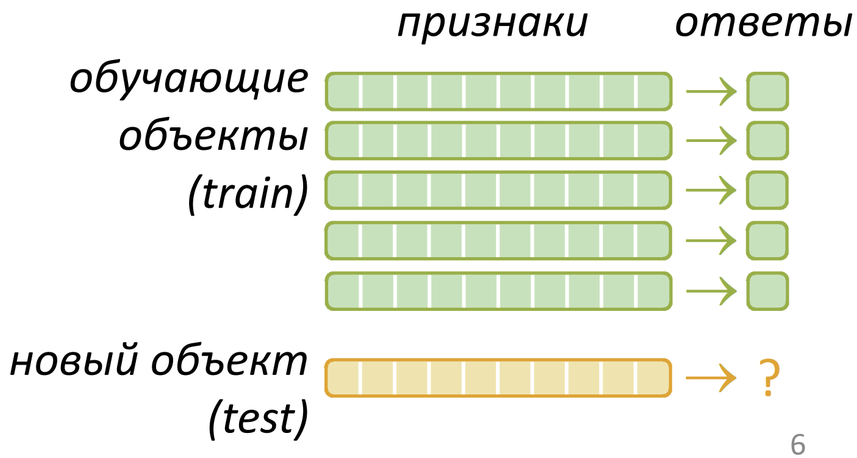
- **На входе:**
  данные – выборка пар «*объект* → *ответ*»,
  каждый объект описывается *вектором признаков*
- **На выходе:**
  модель, предсказывающая ответ по объекту

Задача поставлена,
если у неё есть «ДНК»:

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

Этап №2 – применение (test)

- **На входе:**
  данные – вектор признаков нового объекта
- **На выходе:**
  предсказание ответа на новом объекте



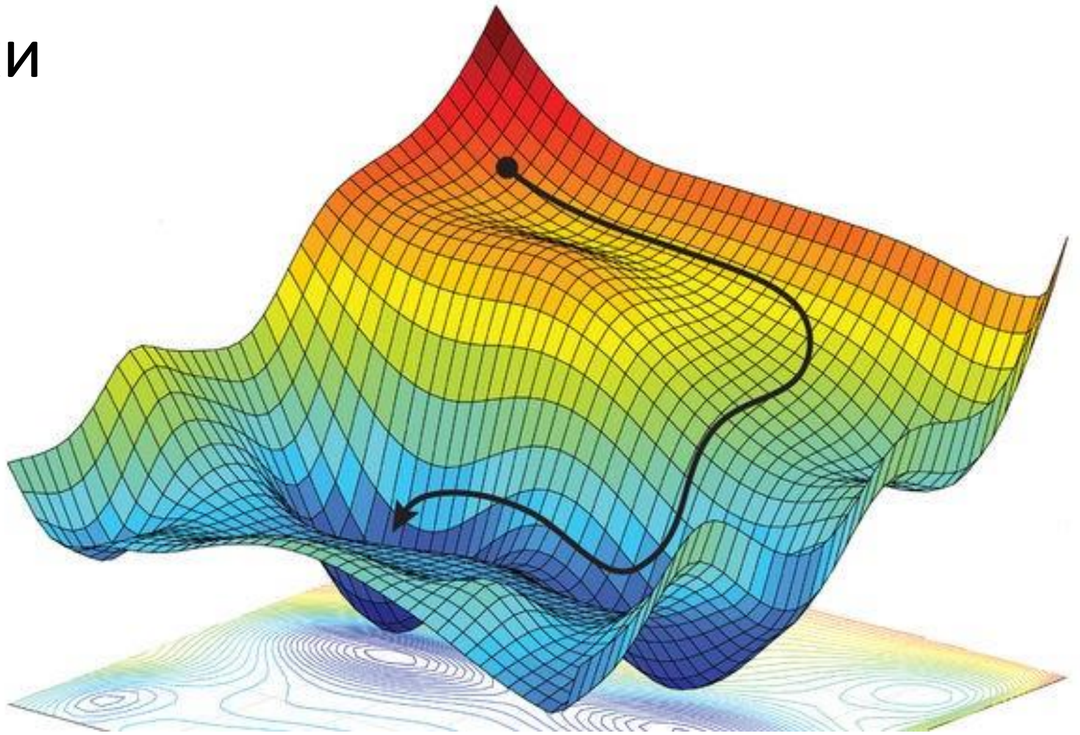
Машинное обучение – это оптимизация

x – вектор объекта обучающей выборки

w – параметры модели

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

$Q(w)$ – критерий качества модели



Задача обучения модели:

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

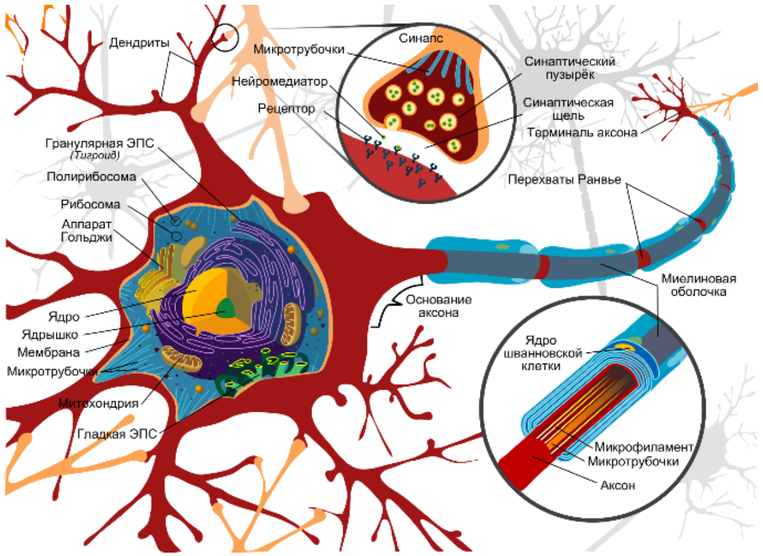
Способ решения – численные методы оптимизации

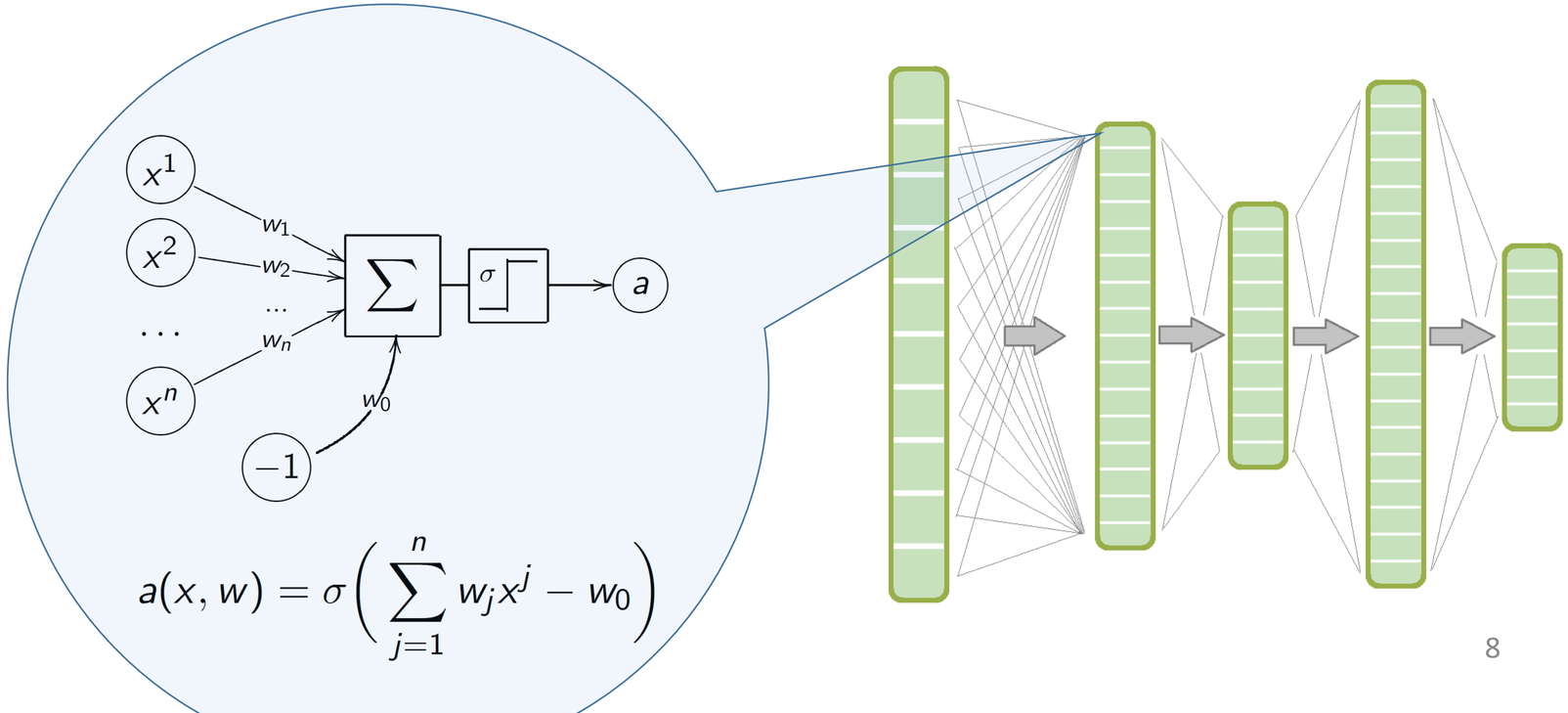
Искусственные нейронные сети

На каждом слое сети вектор объекта преобразуется в новый вектор

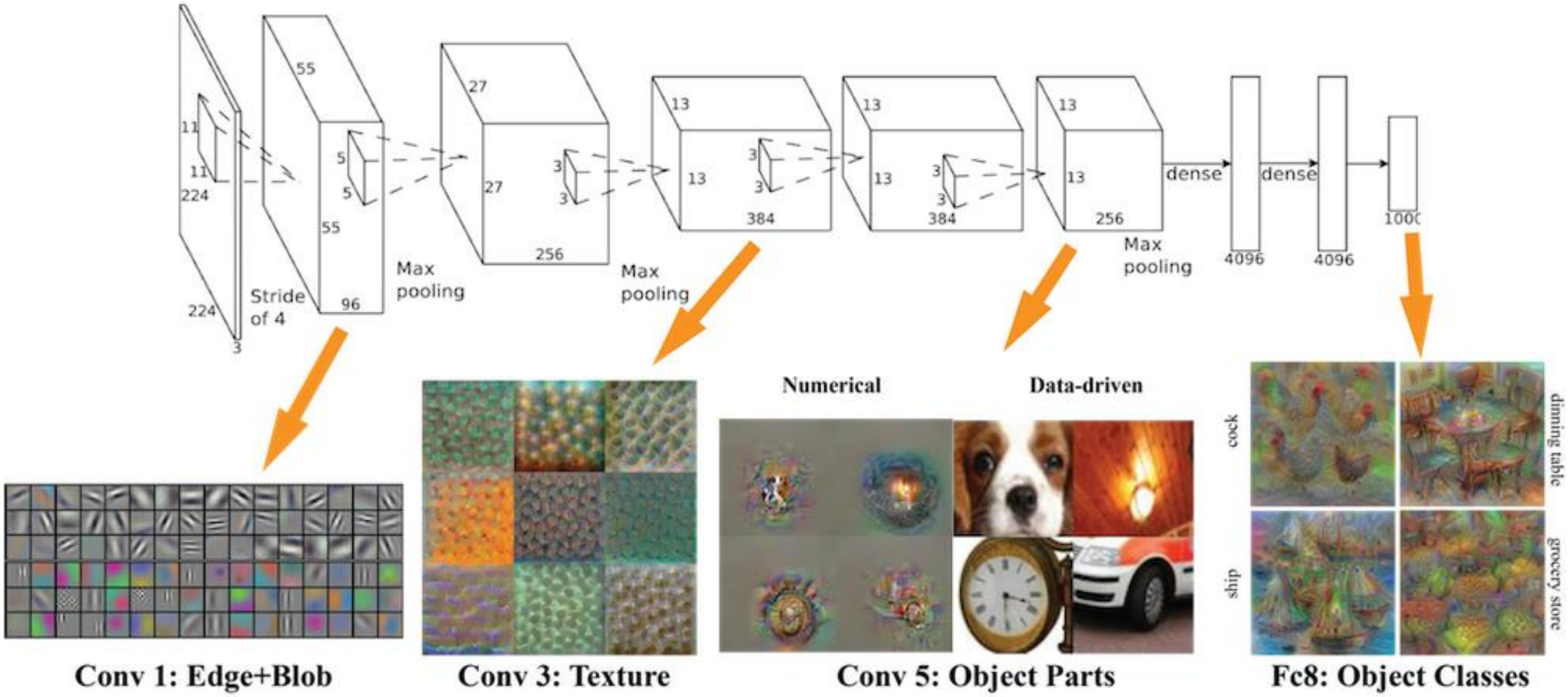
Эти преобразования обучаемые, их параметры входят в w

Каждое преобразование (нейрон) – взвешенная сумма признаков



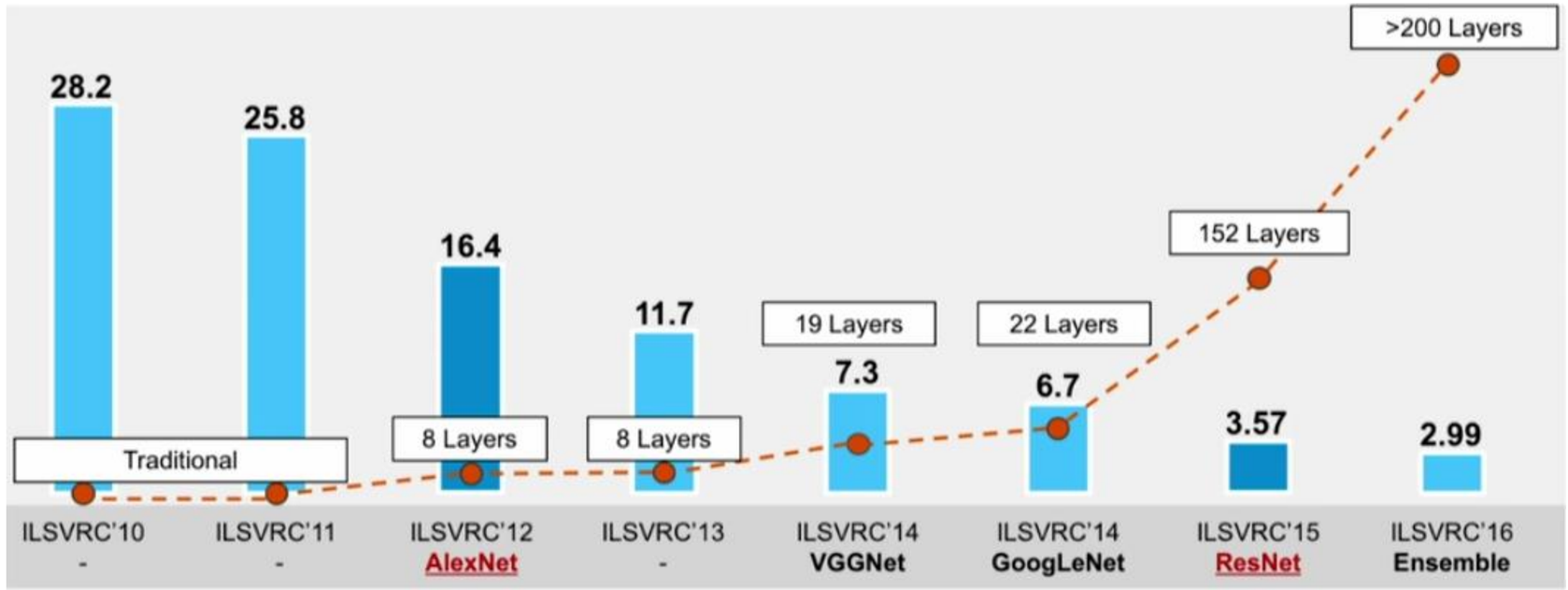


Глубокие свёрточные нейронные сети
для классификации изображений (2012 г.)



Что такое «большие данные». Пример

ImageNet: открытая выборка 14M изображений, 20K категорий

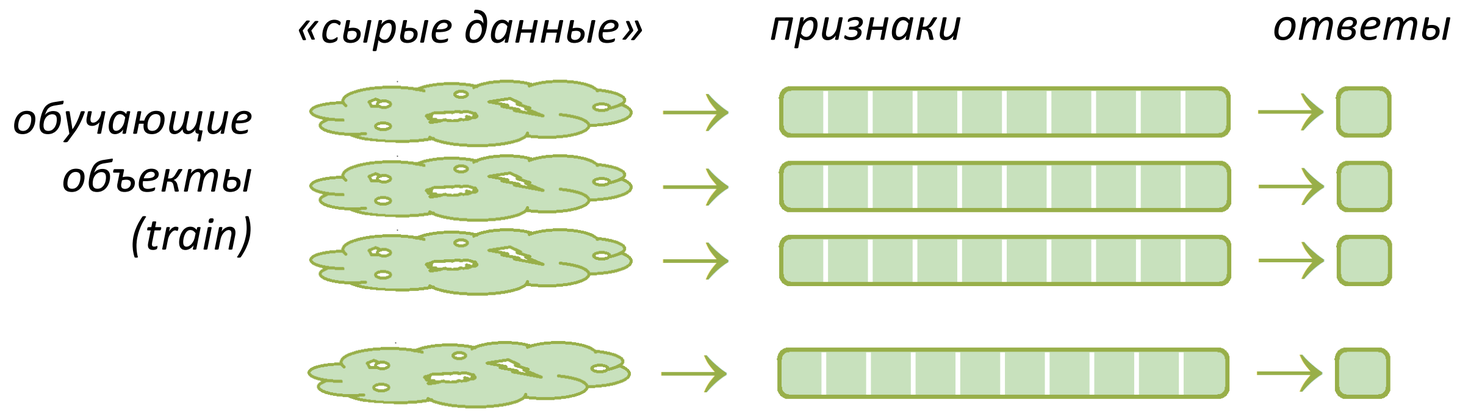


Старт в 2009 г. Человеческий уровень ошибок 5% пройден в 2015 г.

Глубокие нейронные сети

Вход: сложно структурированные «сырые» данные об объектах

Выход: векторные представления объектов и ответы



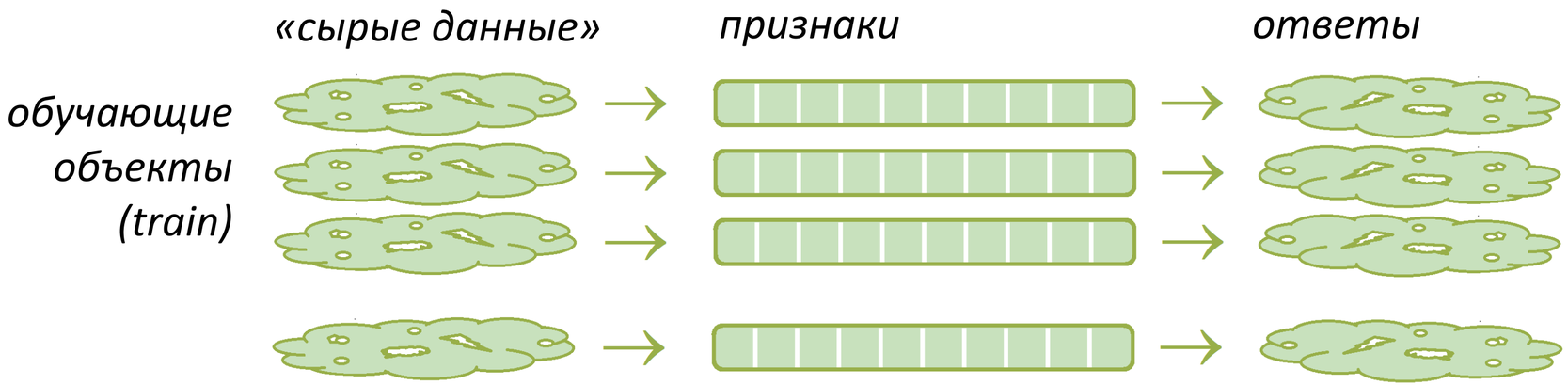
Deep Learning – это всего лишь обучаемая векторизация сложных объектов

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

Генеративные глубокие нейронные сети

Вход: сложно структурированные объекты

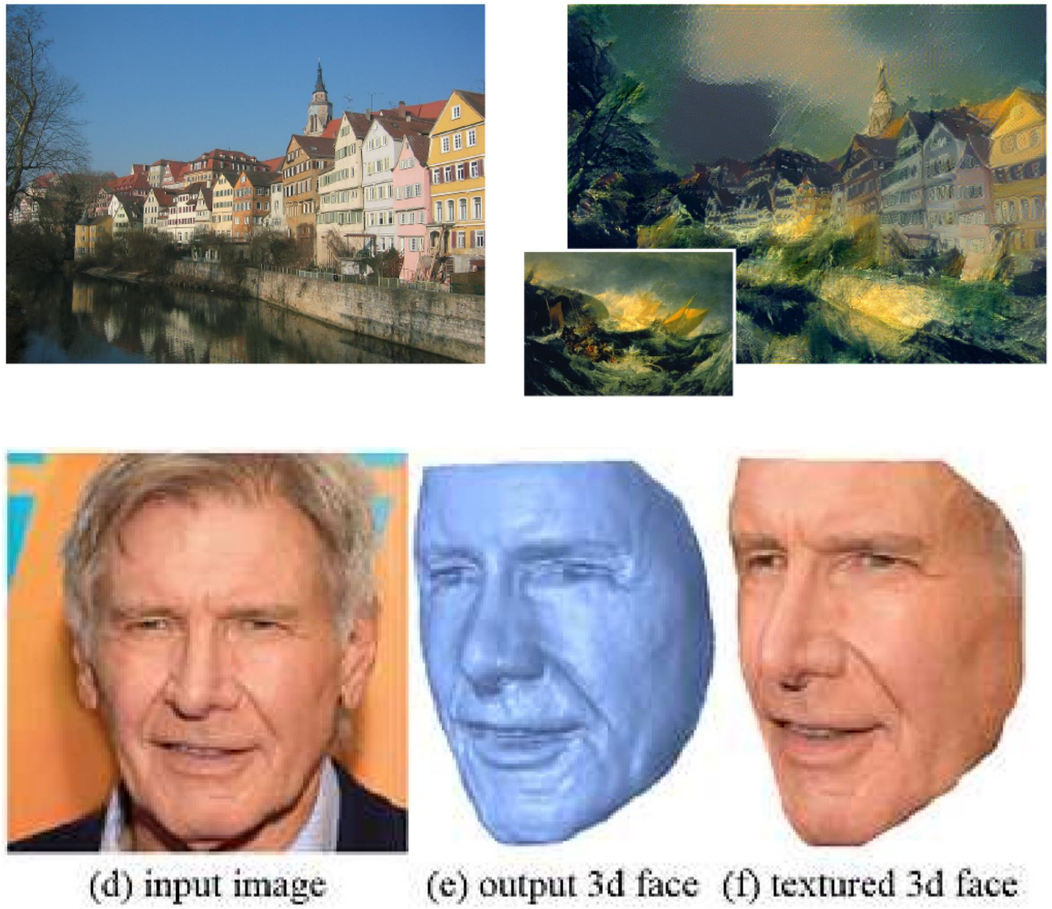
Выход: сложно структурированные ответы

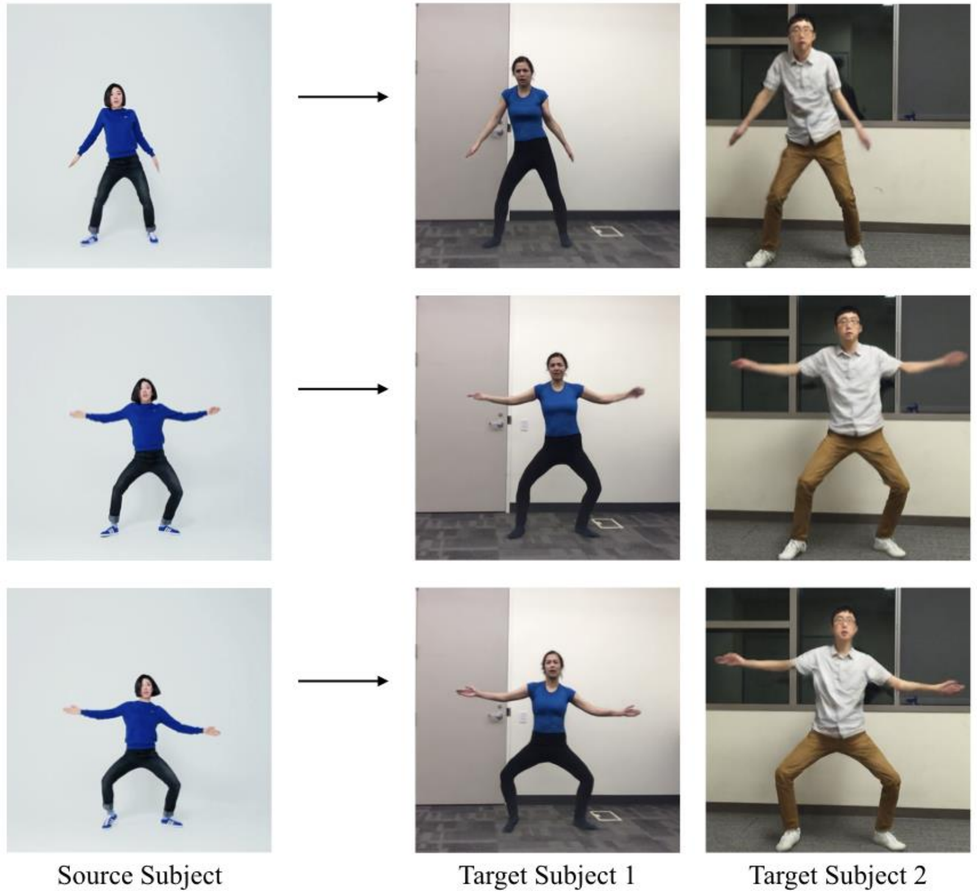


Примеры задач: синтез изображений, перенос стиля, машинный перевод, диалоговый интеллект, реферирование текстов

Модели: seq2seq, CNN, RNN, LSTM, GAN, BERT, GPT-3 , GPT-4 и др.

Синтез изображений и видео





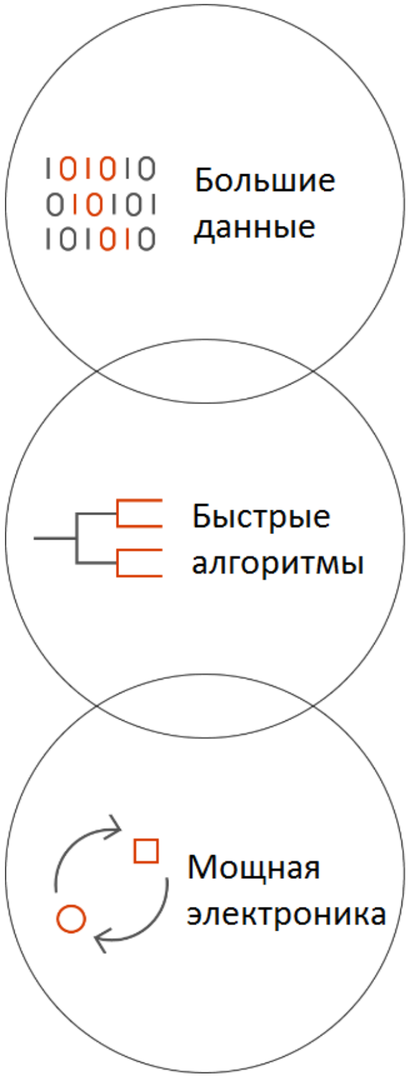
Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros. Everybody Dance Now. ICCV-2019.

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 высокой размерности

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

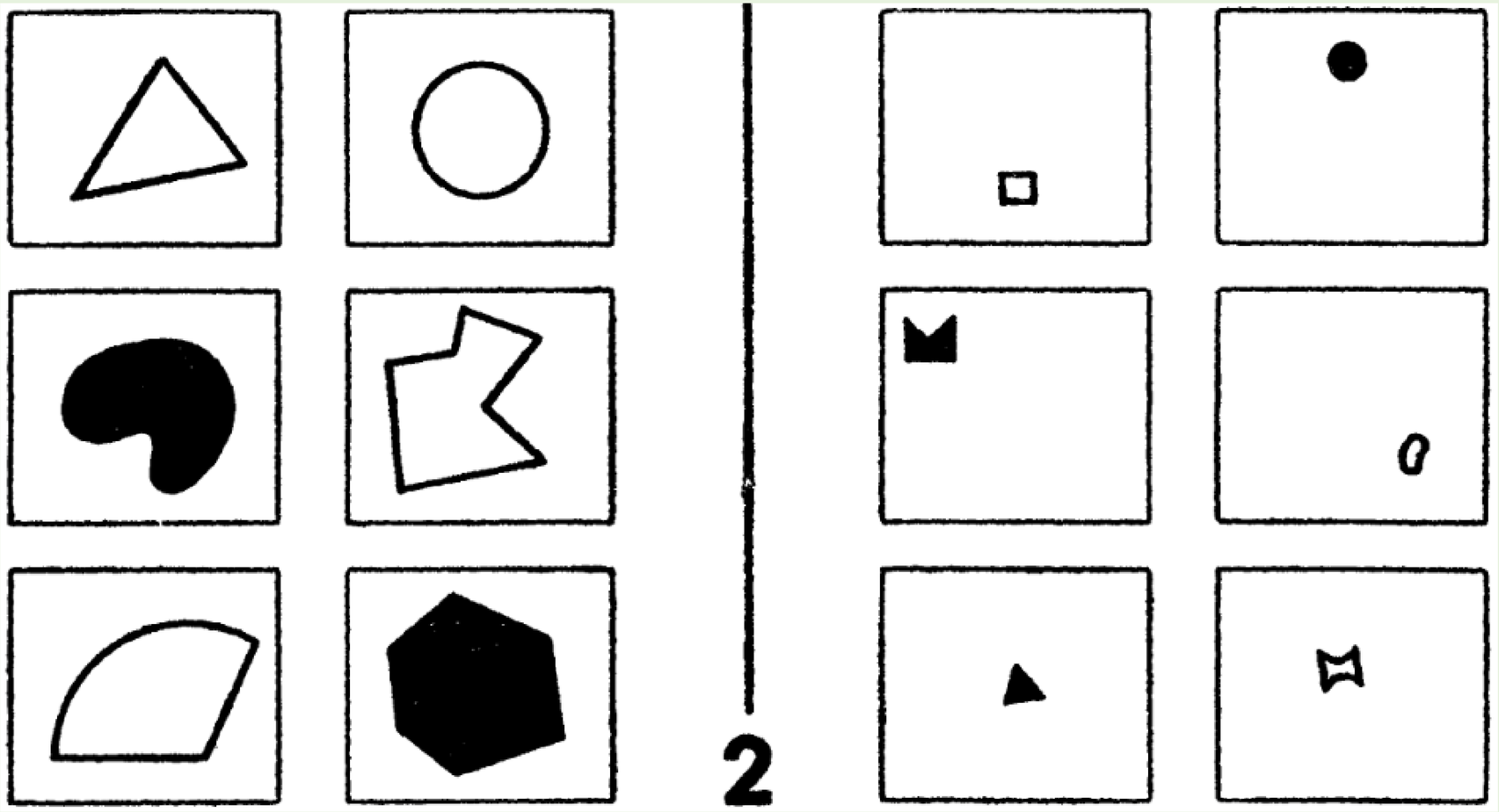


Развлекательная игра! =)

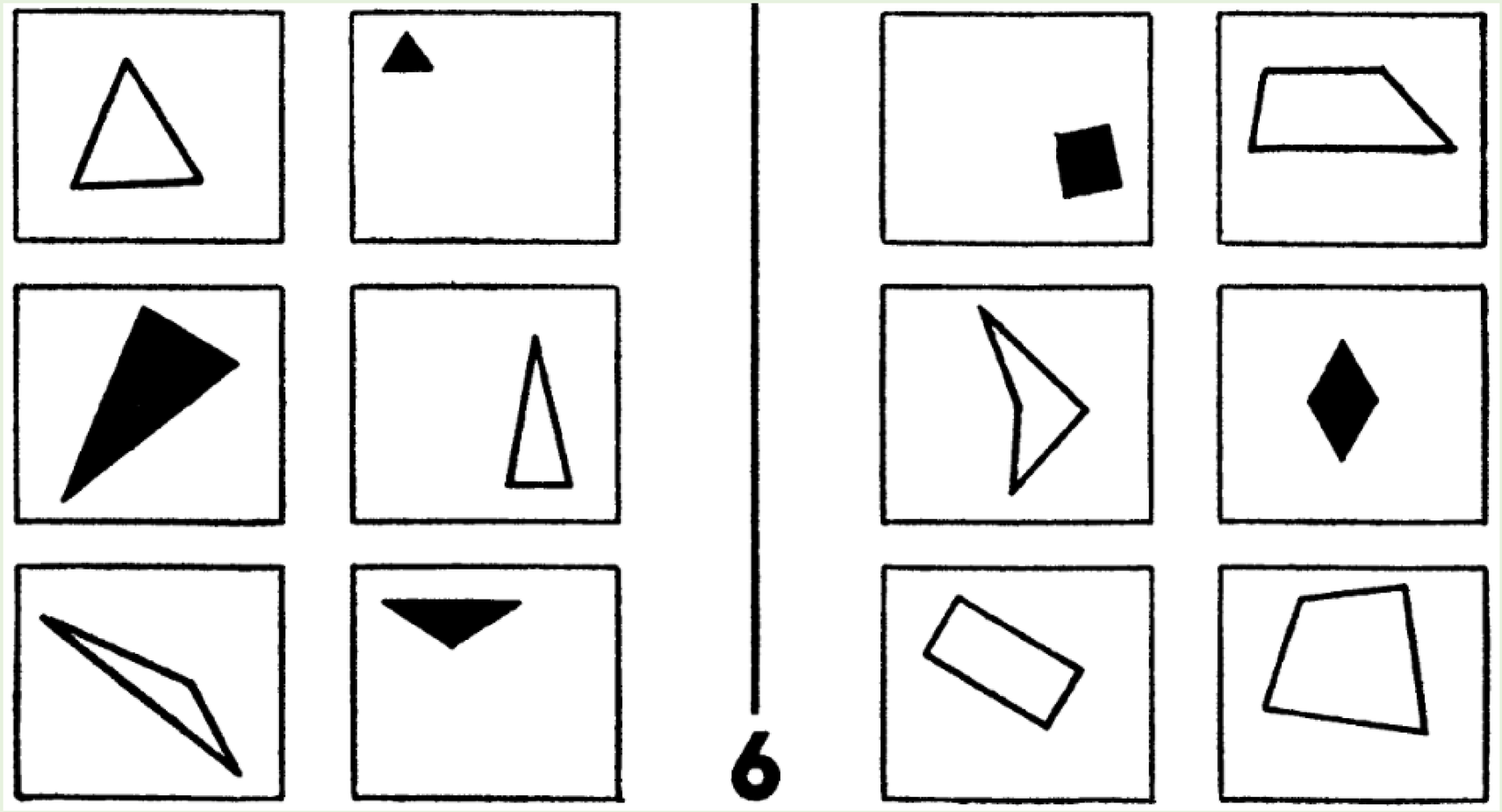
Представим себя на месте искусственного интеллекта...

Михаил Моисеевич Бонгард. Проблема узнавания. М.: Наука, 1967.

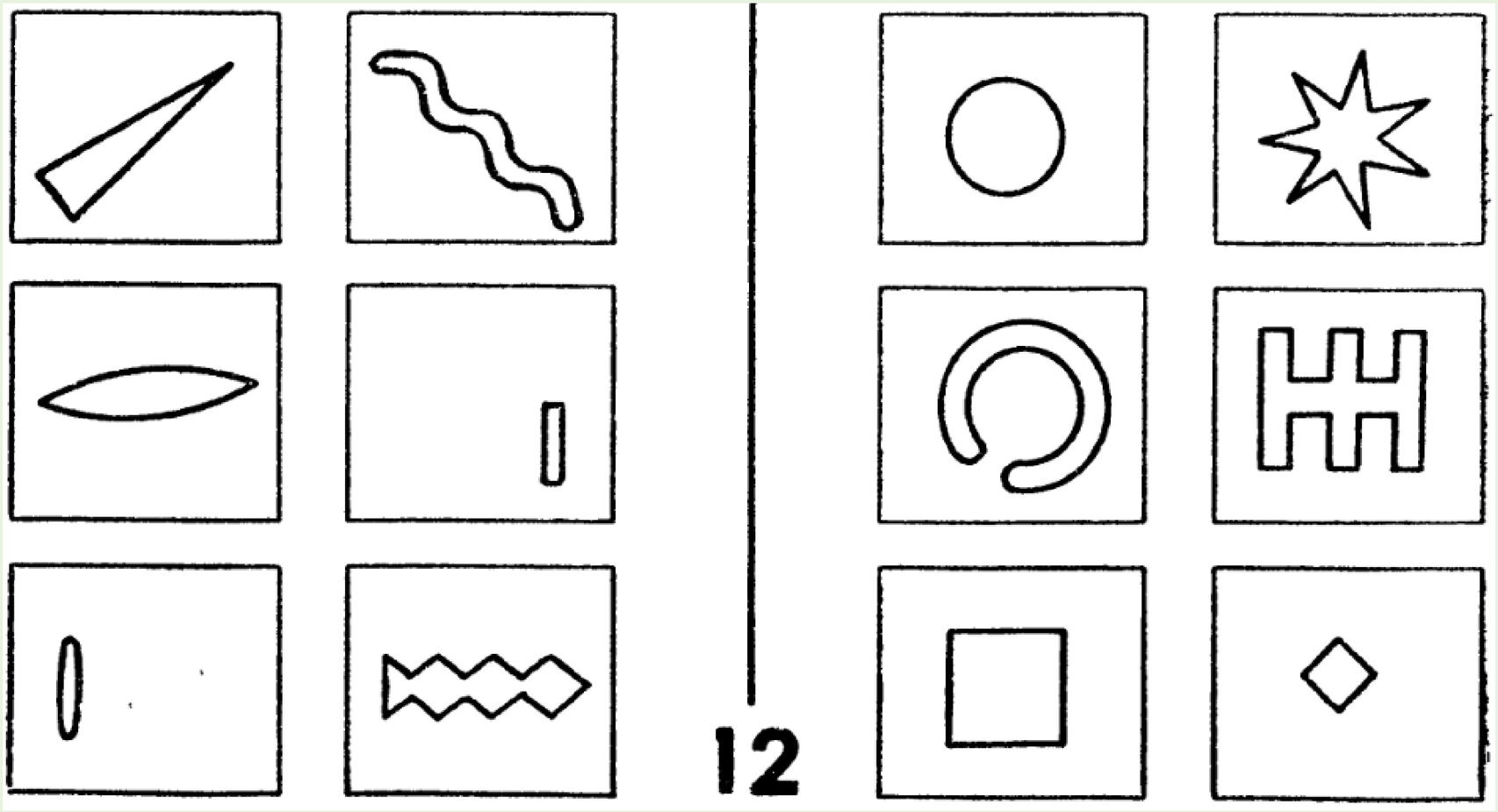
Тесты Бонгарда (1967). Требуется найти правило классификации. Дана обучающая выборка: два класса, по 6 объектов в каждом.



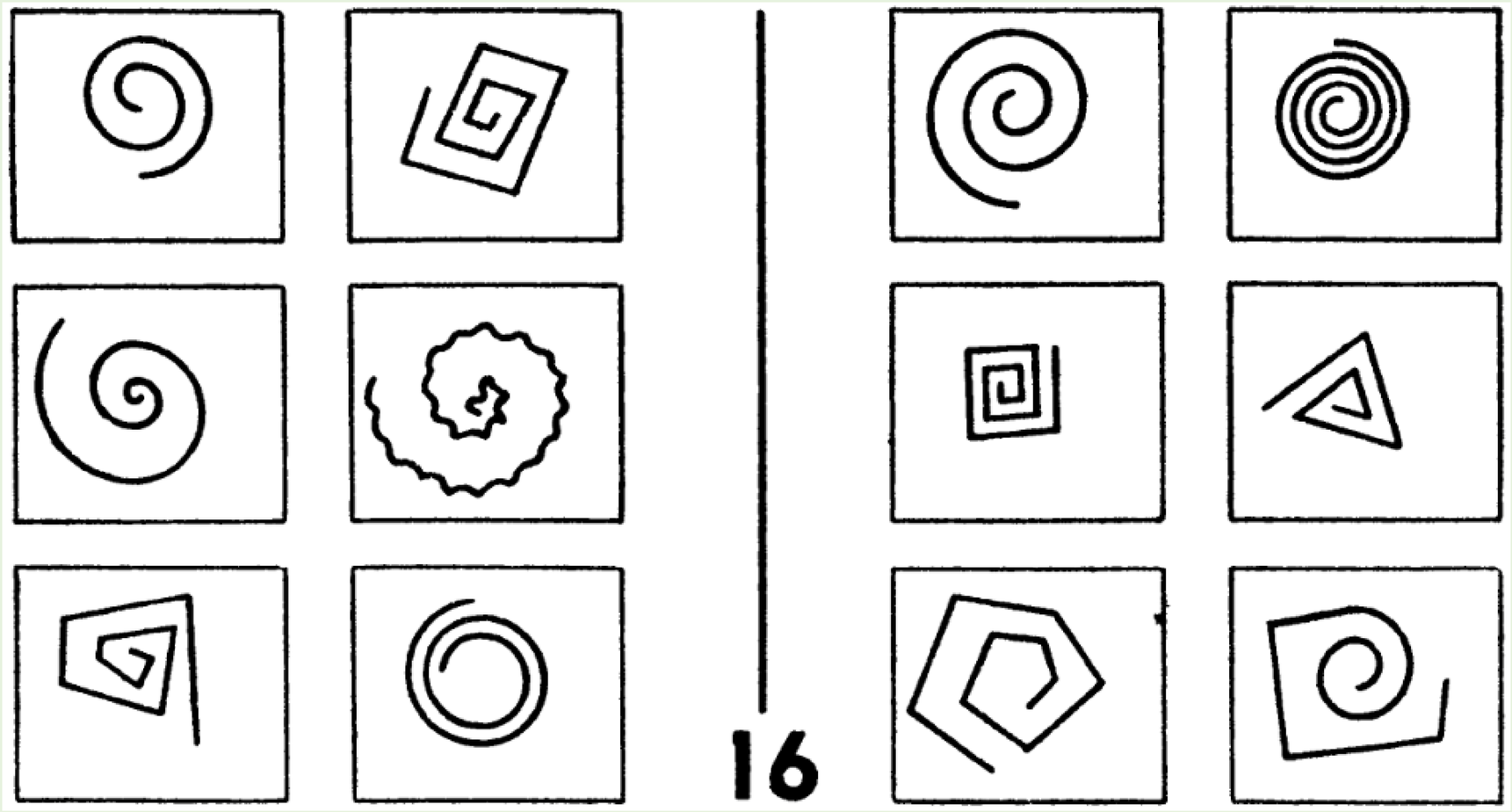
Что даёт нам уверенность, что мы нашли верное правило?
1. Точность классификации известных примеров



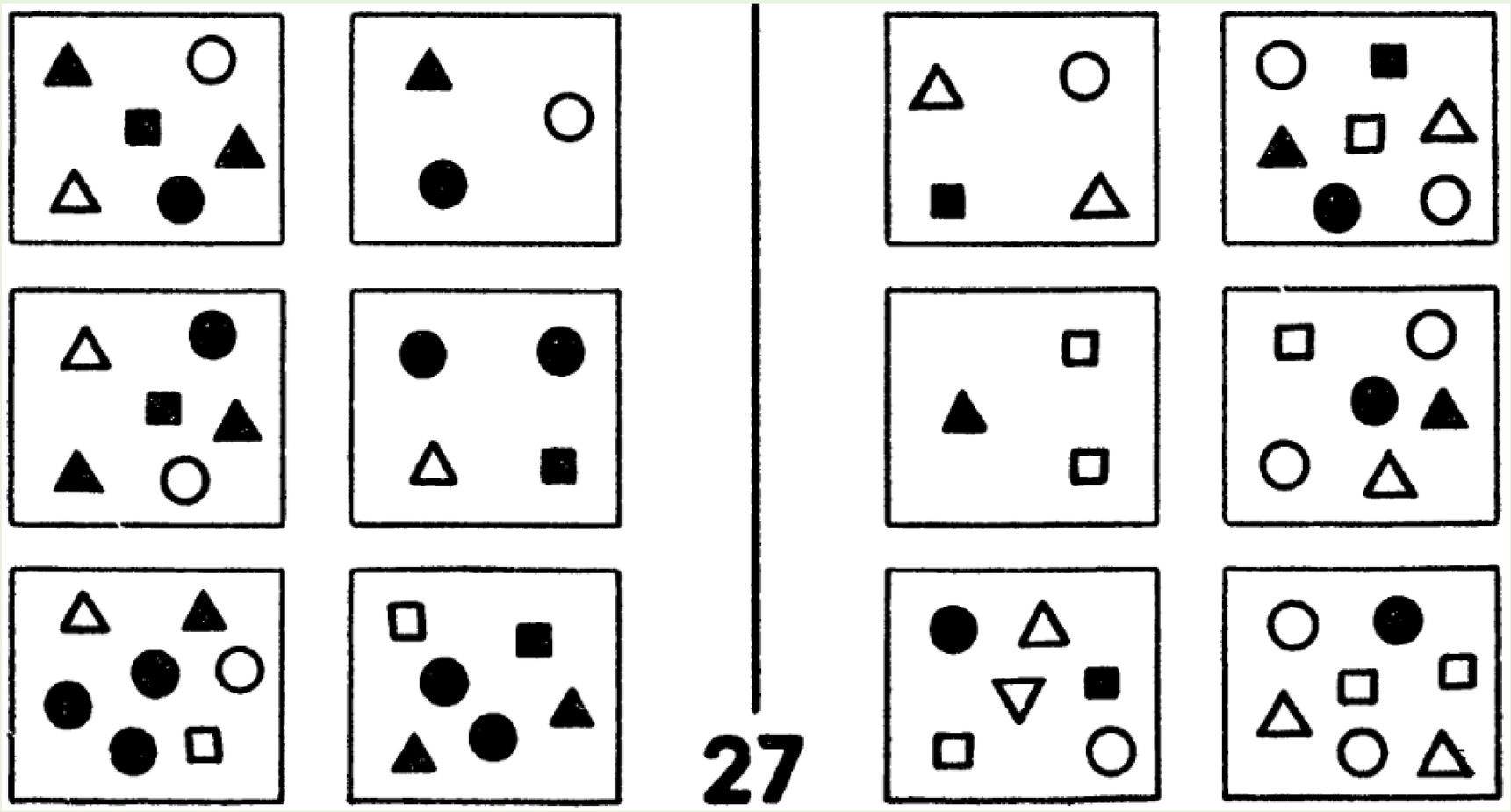
Что ещё даёт нам уверенность, что мы нашли верное правило?
2. Простота и общность правила (неизбыточность модели)



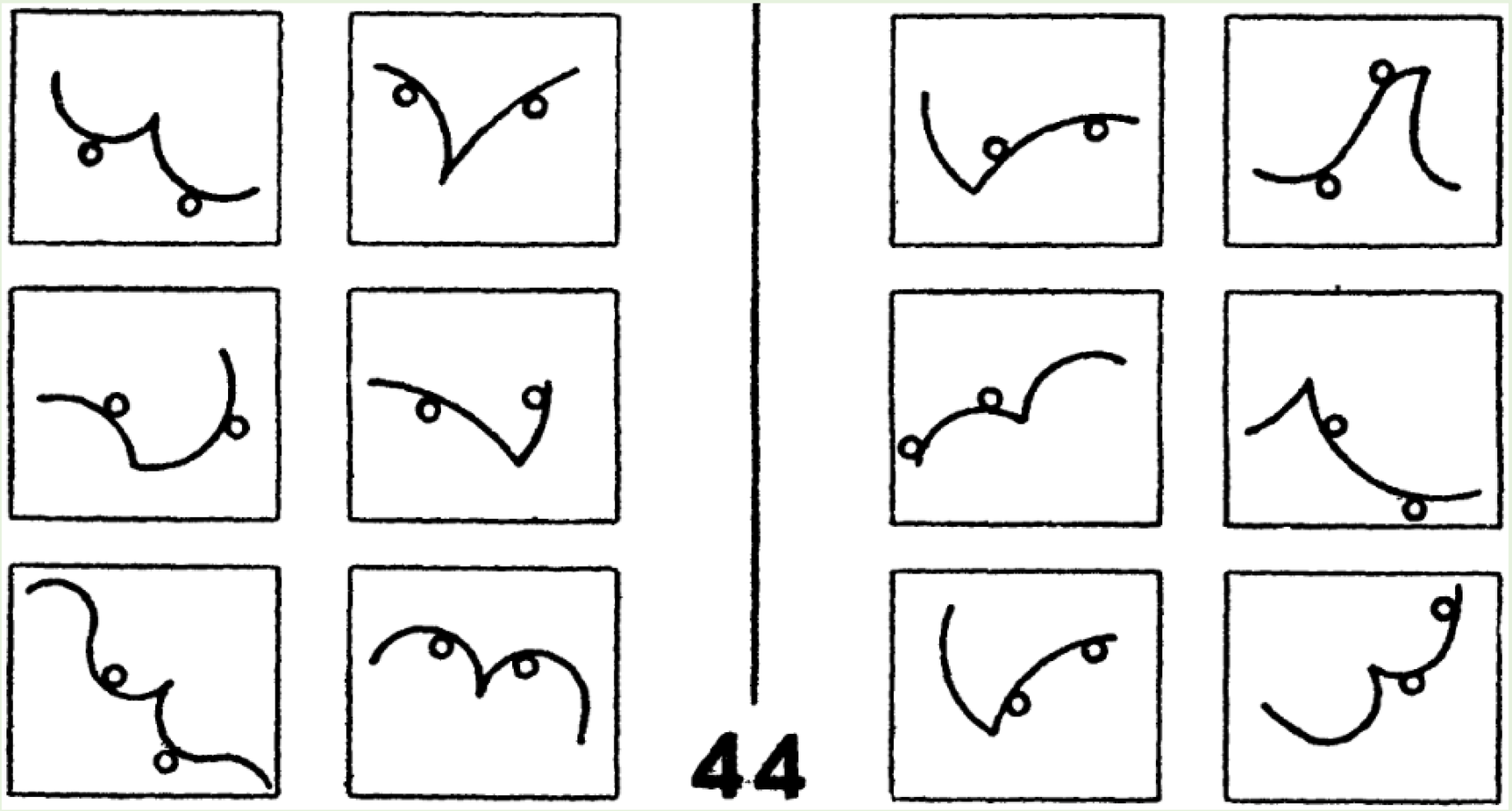
Мы решаем эти задачи почти мгновенно, но они всё ещё сложны для компьютера. Чем мы пользуемся?



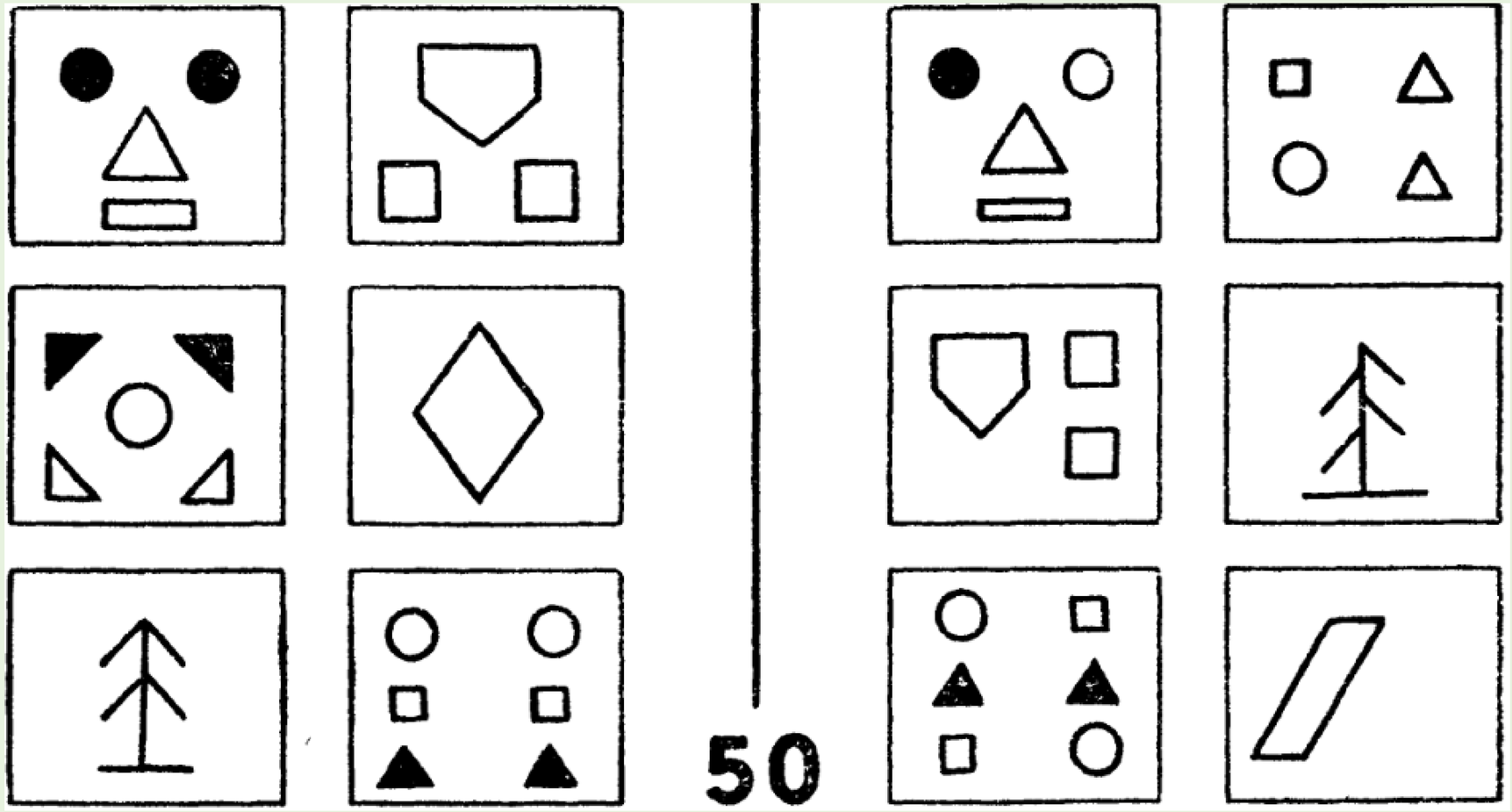
Нужно ли закладывать знания предметной области в явном виде?
Или возможно выработать все необходимые понятия на примерах?



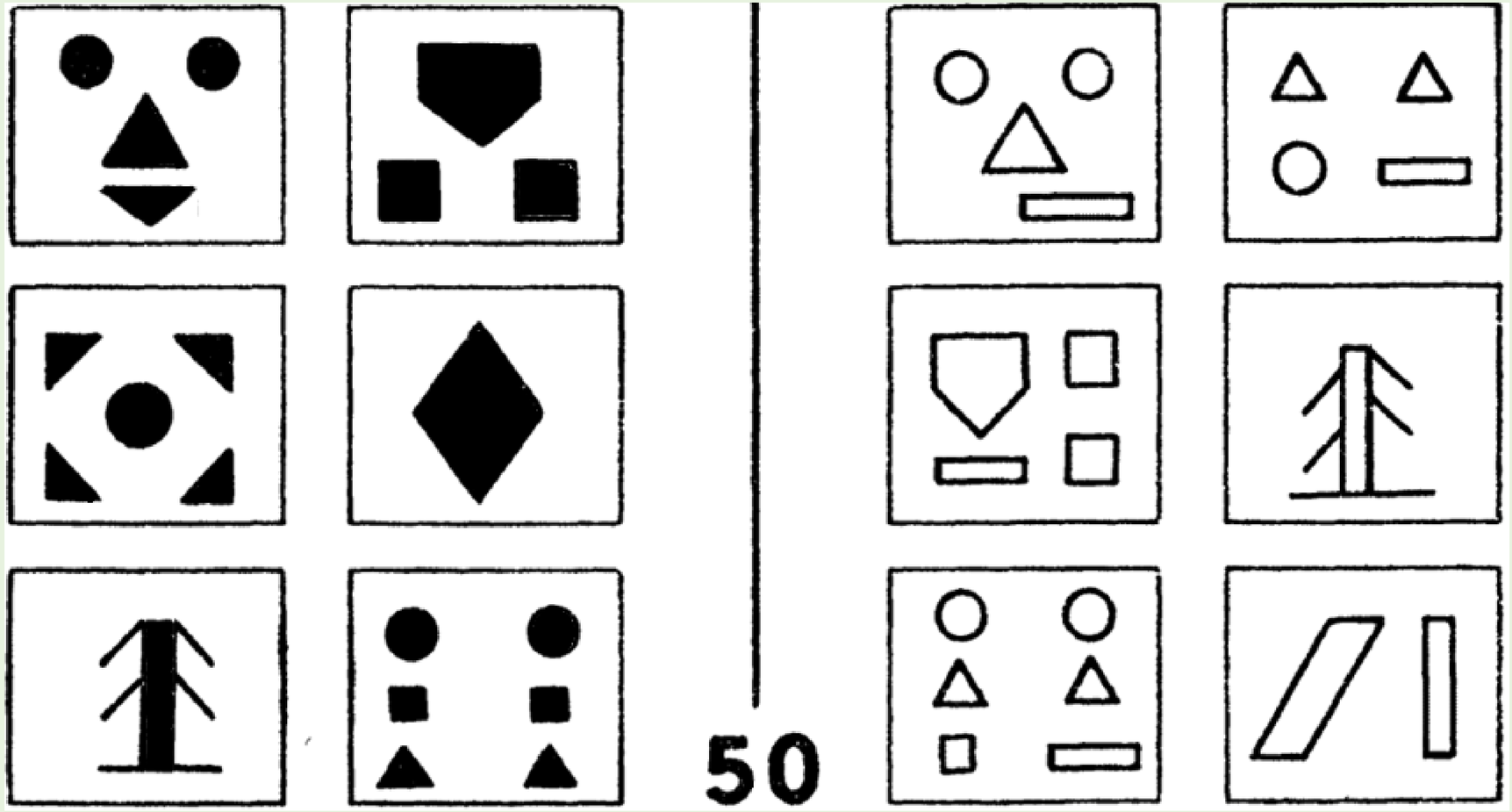
Как вычислять полезные признаки по сложным «сырым» данным?
Возможно ли поручить перебор признаков и моделей машине?



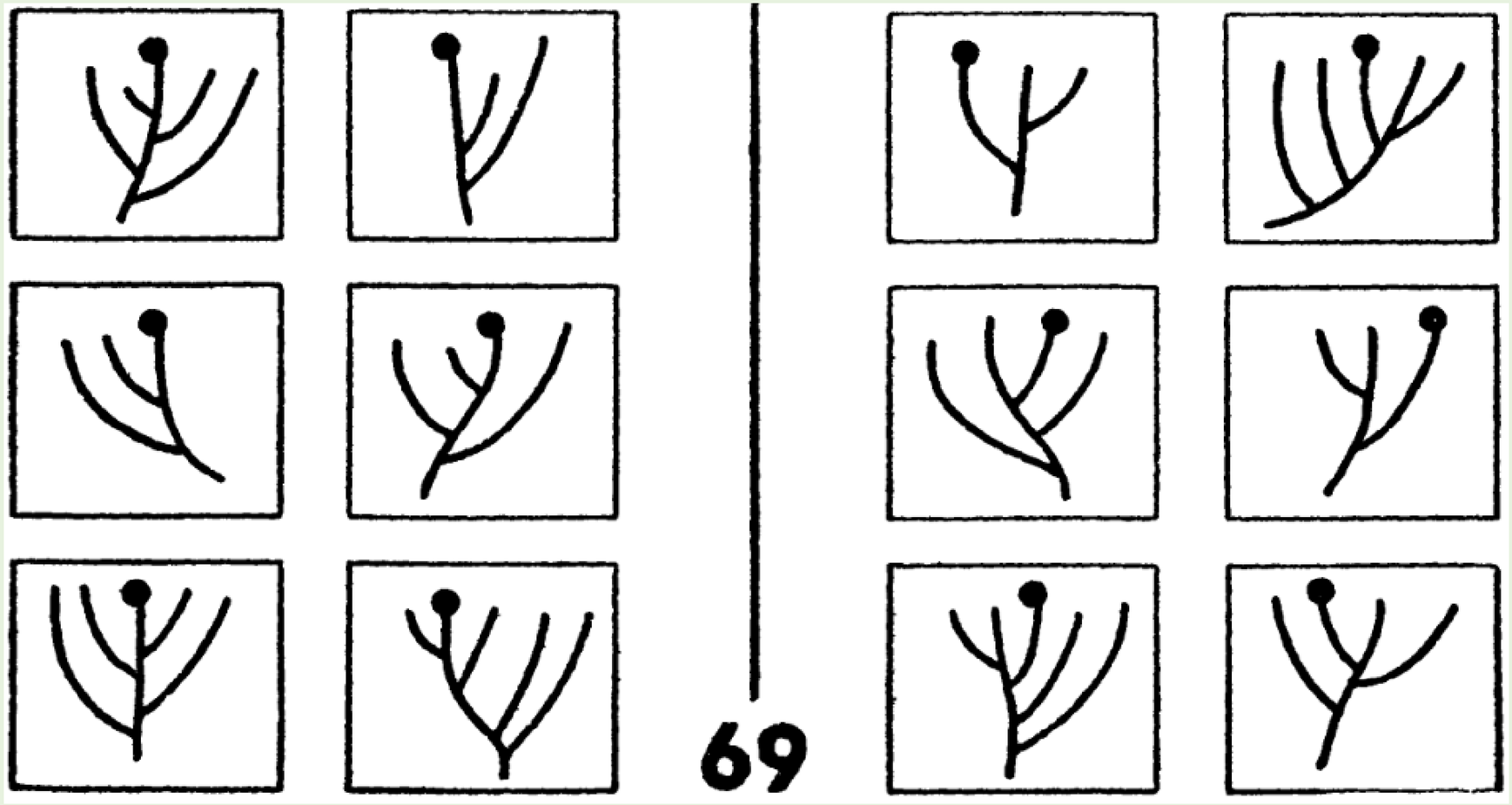
Каков риск выбрать по данным неверное правило, *предрассудок*?
Как этот риск зависит от числа примеров и сложности правил?



Достаточно ли небольшого числа примеров для выработки правила?
Что делать, если к выборке подходят сразу несколько правил?



Эти вопросы составляют проблематику машинного обучения и сегодня. М.М.Бонгард поставил все эти проблемы в середине 60-х!



От игры к реальной жизни

Основная задача естествознания:

- как по ограниченному числу фактов установить истину, закон природы?

Как избежать предрассудков?

- т.е. как принимать верные решения в условиях неполной информации?

Какие существуют способы умышленно сбить нас с толку?

- т.е. как устроено манипулирование, пропаганда, когнитивные войны?

Как противостоять таким попыткам?

- и чем тут может помочь искусственный интеллект?

Загадка! Летающие стержни (skyfish, rods)

«Летающие стержни — это совершенно новый вид неопознанных объектов, которые мы не в состоянии пока не только изучить, но и даже просто и банально поймать»

Небольшая голливудская телекомпания Nash Entertainment сняла в 2001 году документальный фильм «RODS!» продолжительностью 44 минуты об этом **загадочном явлении** с участием американского исследователя Хосе Эскамилья

https://www.youtube.com/watch?v=eHDQIiOQrQA (уже недоступно)

Разгадка летающих стержней

Википедия:

Скайфиш (от англ. *sky* — «небо» и *fish* — *«рыба»*), также их нередко называют ***rods*** (стержни) — артефакты, изображения в виде продолговатых тонких объектов с продольной мерцающей «бахромой», создаваемые попадающими в кадр видеосъемки быстро летящими насекомыми.

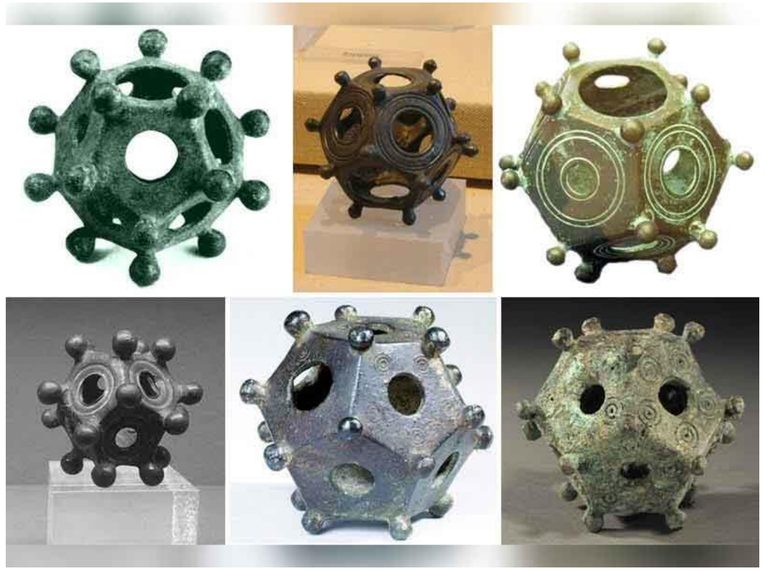
Хосе Эскамилья — уфолог ;)

Фотография летающих ночных бабочек, выполненная с большой выдержкой

Загадка! Римские додекаэдры

Каково их предназначение?

найдено более 200 штук,
II-IV вв.н.э., размер 4-11 см



Загадка! Римские додекаэдры

В интернете обсуждается масса версий. Утверждается, что предназначение до сих пор неизвестно. Тысячи комментариев с жаркими спорами и обсуждением догадок…

- подсвечник
- предмет культа
- астрономический прибор
- дальномер, в т.ч. на поле боя
- определитель фальшивых монет
- элемент крепежа шестов для сборки палаток
- инструменты для калибровки водяных труб…

Все эти версии оставляют вопросы: почему не сделать проще? почему именно додекаэдр? почему шарики на ножках?

Римские додекаэдры. Разгадка.

Приспособление для вязания, скорее всего, перчаток. Единственная версия, которая объясняет сразу всё:

- почему отверстия разного диаметра
- почему шарики в вершинах
- почему додекаэдр (изготовление 12 отдельных плоских граней потребовало бы отлить 60 шариков вместо 20)
- почему в северных провинциях
- почему это ценный предмет из бронзы

Важные выводы

Очень трудно отличить правду от выдумки (фейка) в условиях

- неполной информации,
- неустановленного достоверного ответа (известной разгадки),
- обилия взаимоисключающих версий и мнений,
- практической невозможности отыскать все версии и мнения,
- (а теперь и) появления генеративных моделей типа GPT-4, ChatGPT.

И всё это при условии, что нас не пытаются обманывать намеренно.

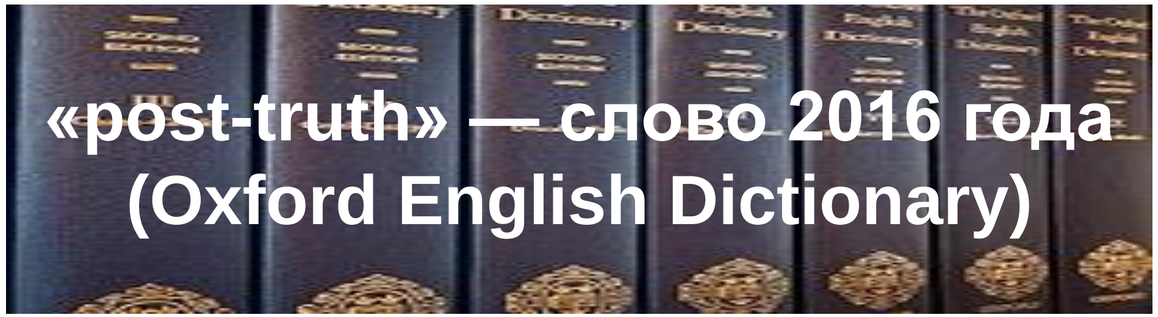
Обманывают ли нас намеренно?

«Пусть на улицах вражеской столицы шепчутся, что князь обворовывает народ, советники его предали, чиновники спились, а воины голодные и босые. Пусть жители калечат имя своего князя и произносят его неправильно... Пусть им при сытой жизни кажется, что они голодают. Пусть состоятельные жители завидуют тем, кто в княжестве Вэй пасет скот. Разжигайте внутренний пожар не огнем, а словом, и глупые начнут жаловаться и проклинать свою родину. И тогда мы пройдем через открытые ворота.»

(Сунь Цзы, Искусство войны, IV-V век до н. э.)

«Пропаганда — это инструмент тотальной политики вместе с дипломатией, экономическими мероприятиями и вооруженными силами. Ее цель заключается в экономии материальных затрат на мировое господство.»
(Гарольд Лассуэлл, 1927)

Постправда, пропаганда, когнитивные войны

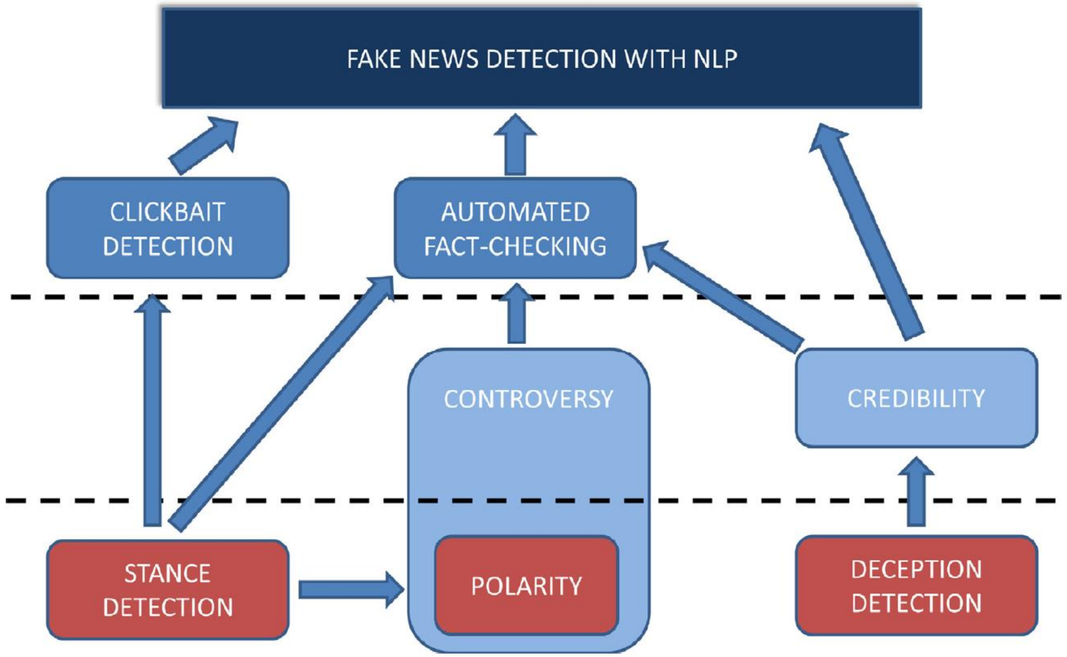


- Факты становятся менее значимы, чем эмоции и личные убеждения
- Явление «информационных пузырей»
- Явление «неопровержимой лжи»
- Постправда маскируется под «другие грани истины»
- Постправда — новая форма пропаганды и инструмент «мягкой силы»

Область исследований «Fake News Detection»

Задачи ML / NLP / NLU:

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

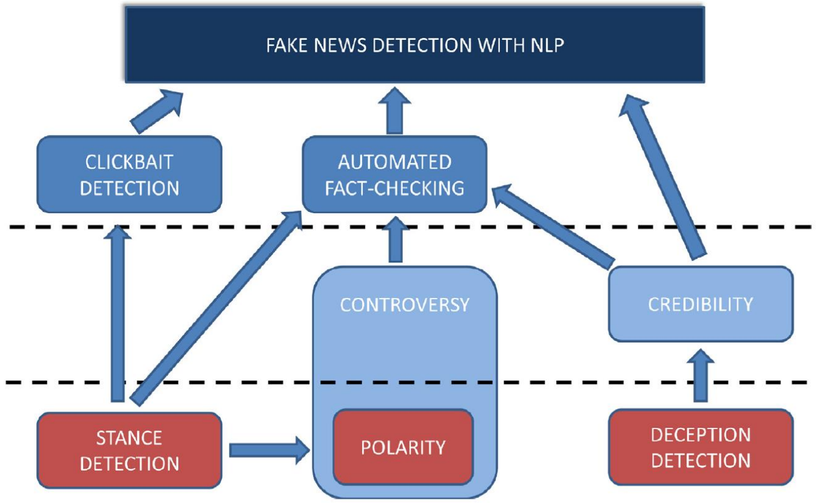


E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems With Applications, Elsevier, 2020.

Чего-то не хватает…

1. **Fake News** – не единственный, не главный не самый сильный инструмент постправды
2. **Пропаганда** использует не только фейки, но и замалчивание и подтасовку фактов, манипулятивные воздействия и т.д.
3. **Когнитивные войны** нацелены на разрушение общественного согласия, социокультурного кода, идеологии

- Как распознавать манипулятивные воздействия и идеологические атаки?
- Как находить поляризацию и замаливание?
- 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.

Концепция проекта «Новостной коллайдер»

Физики создают адронный коллайдер,
чтобы, сталкивая потоки частиц,
узнать больше о строении материи

Мы создаём новостной коллайдер,
чтобы, сталкивая потоки новостей,
защитить общество от угроз эпохи
постправды и когнитивных войн

Типология деструктивного дискурса и система задач ML /NLP /NLU для его детекции

воздействия → фейки → пропаганда → когн.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: идеологем, мифологем
8. ■□■■ оценивание возможных психо-эмоциональных реакций
9. ■□■■ выявление целевых аудиторий воздействия
10. □■■■ **оценивание виральности (virality prediction)**
11. □■■□ **оценивание достоверности источников (credibility scores)**
12. □□□■ детекция прямой агрессии (угрозы, призывы, провокации, вербовка, экстремизм)

Четыре основных типа задач ML / NLP / NLU

1. **Классификация текста (новости или предложения) целиком**
   - deception detection, fact-checking, text credibility
2. **Классификация пары текстов**
   - stance, controversy, polarization, clickbait detection
   - выявление противоречий, разногласий, замалчивания
3. **Выделение и классификация (тегирование) фрагментов текста**
   - поиск лингвистических маркеров (linguistic-based cues) в тексте
   - детекция приёмов манипулирования
   - выявление конструктов картины мира: идеологем, мифологем
   - выявление психо-эмоциональных реакций и целевых аудиторий
4. **Кластеризация или тематическое моделирование**
   - кластеризация мнений по заданной теме (controversy detection)
   - выявление устойчивых сочетаний мнений (polarization detection)
   - выявление мнений как сочетаний слов, их семантических ролей и тональностей
   - выявление «картин мира» – устойчивых сочетаний мнений и идеологем

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

Примеры приёмов манипулирования

Лозунговые слова

Также мэр **Владивостока Шестаков** рассказал, что в 2022 году **власти** займутся улучшением работы общественного транспорта.

Навешивание ярлыков

11 октября 2021 года Талибы назвали провалом **политику НАТО в Афганистане**

Эвфемизация

Кардашьян приложила немало усилий, чтобы помочь **супругу** в том числе и с его психическими проблемами, однако со временем ей стало это в тягость.

Негативирующая гиперболизация

Сергей Нетесов в беседе с «Известиями» указал на то, что **Россия** стоит на пороге самой мощной волны пандемии за всё время существования COVID-19**.**

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

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

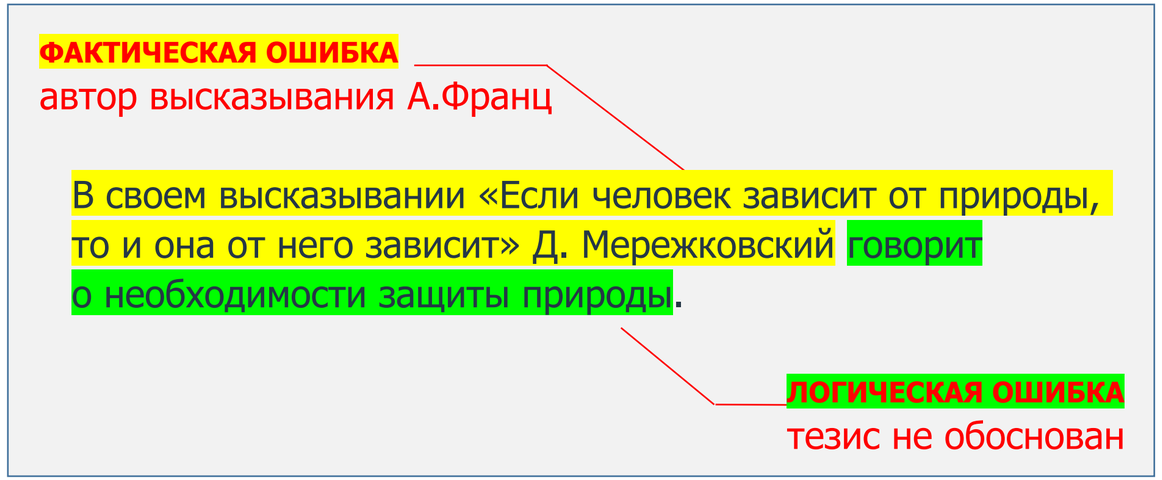
Период: декабрь 2019 — июнь 2022, три цикла испытаний.

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)

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

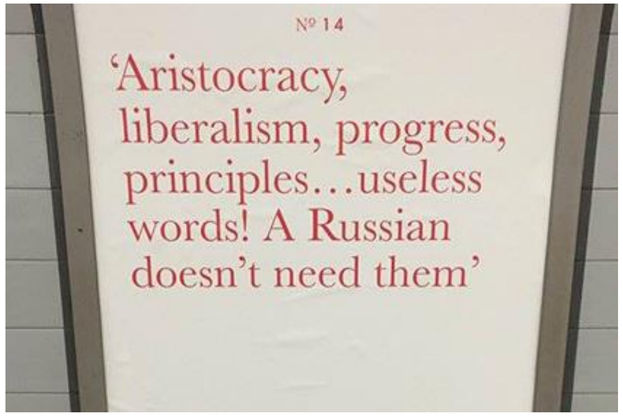
Обучающие выборки:

- 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.

Выводы

1. ML — это оптимизация параметров предсказательных моделей
2. AI (ИИ) — не интеллект, а «Имитация Интеллекта», обучаемая векторизация сложно структурированных данных
3. Модели трансформеры (BERT, GPT) позволяют теперь решать сложные задачи понимания естественного языка
4. В том числе стоить модели для мониторинга и детекции угроз в медийном информационном пространстве
5. Разметка текстовых данных — магистральный путь формализации гуманитарных знаний в таких задачах
6. Методология разметки и оценивания идёт к стандартизации

Спасибо за внимание!

Цивилизационная идеология

ДЗЕН-канал

https://dzen.ru/civideology

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