Анализ данных и машинный интеллект — профессии будущего

Воронцов Константин Вячеславович

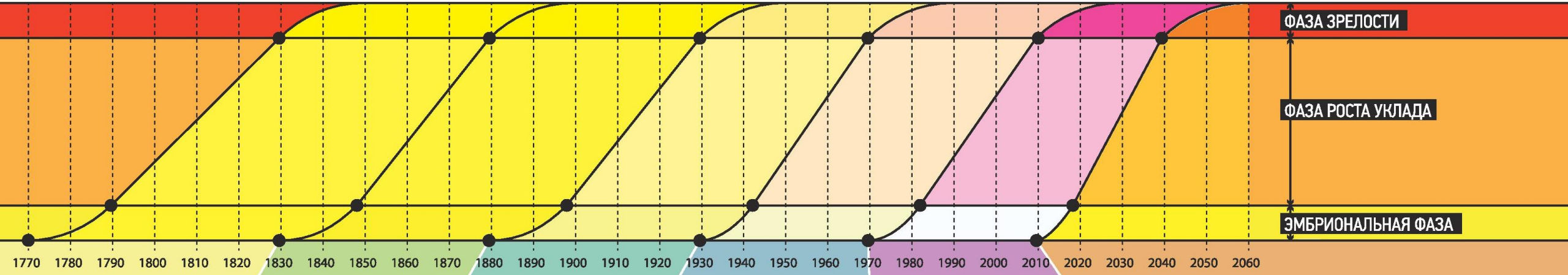
● Московский Физико-Технический Институт ●
● Вычислительный Центр им. А.А.Дородницына ФИЦ ИУ РАН ●
● ШАД Яндекс ●

voron@forecsys.ru

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

Клаус Мартин Шваб,
президент Всемирного
экономического форума

Технологические уклады по Н.Д.Кондратьеву



ПЕРВЫЙ ТЕХНОЛОГИЧЕСКИЙ УКЛАД

Основной ресурс: энергия воды

Главная отрасль: текстильная промышленность

Ключевой фактор: текстильные машины

Достижение уклада: механизация фабричного производства

ВТОРОЙ ТЕХНОЛОГИЧЕСКИЙ УКЛАД

Основной ресурс: энергия пара, уголь

Главная отрасль: транспорт, чёрная металлургия

Ключевой фактор: паровой двигатель, паровые приводы станков

Достижения уклада: рост масштабов производства, развитие транспорта

Гуманитарное преимущество: постепенное освобождение человека от тяжёлого ручного труда

ТРЕТИЙ ТЕХНОЛОГИЧЕСКИЙ УКЛАД

Основной ресурс: электрическая энергия

Главная отрасль: тяжелое машиностроение, электротехническая промышленность

Ключевой фактор: электродвигатель

Достижения уклада: концентрация банковского и финансового капитала; появление радиосвязи, телеграфа; стандартизация производства;

Гуманитарное преимущество: повышение качества жизни

ЧЕТВЕРТЫЙ ТЕХНОЛОГИЧЕСКИЙ УКЛАД

Основной ресурс: энергия углеводородов, начало ядерной энергетики

Основные отрасли: автомобилестроение, цветная металлургия, нефтепереработка, синтетические полимерные материалы

Ключевой фактор: двигатель внутреннего сгорания, нефтехимия

Достижения уклада: массовое и серийное производство

Гуманитарное преимущество: развитие связи, транснациональных отношений, рост производства продуктов народного потребления

ПЯТЫЙ ТЕХНОЛОГИЧЕСКИЙ УКЛАД

Основной ресурс: атомная энергетика

Основные отрасли: электроника и микроэлектроника, информационные технологии, генная инженерия, программное обеспечение, телекоммуникации, освоение космического пространства

Ключевой фактор: микроэлектронные компоненты

Достижения уклада: индивидуализация производства и потребления

Гуманитарное преимущество: глобализация, скорость связи и перемещения

ШЕСТОЙ ТЕХНОЛОГИЧЕСКИЙ УКЛАД

Все составляющие нового технологического уклада носят характер прогноза

Основные отрасли: нано- и биотехнологии, наноэнергетика, молекулярная, клеточная и ядерная технологии, нанобиотехнологии, биомиметика, нанобионика, нанотроника и другие наноразмерные производства; новые медицина, бытовая техника, виды транспорта и коммуникаций, использование стволовых клеток, инженерия живых тканей и органов, восстановительная хирургия и медицина

Ключевой фактор: микроэлектронные компоненты

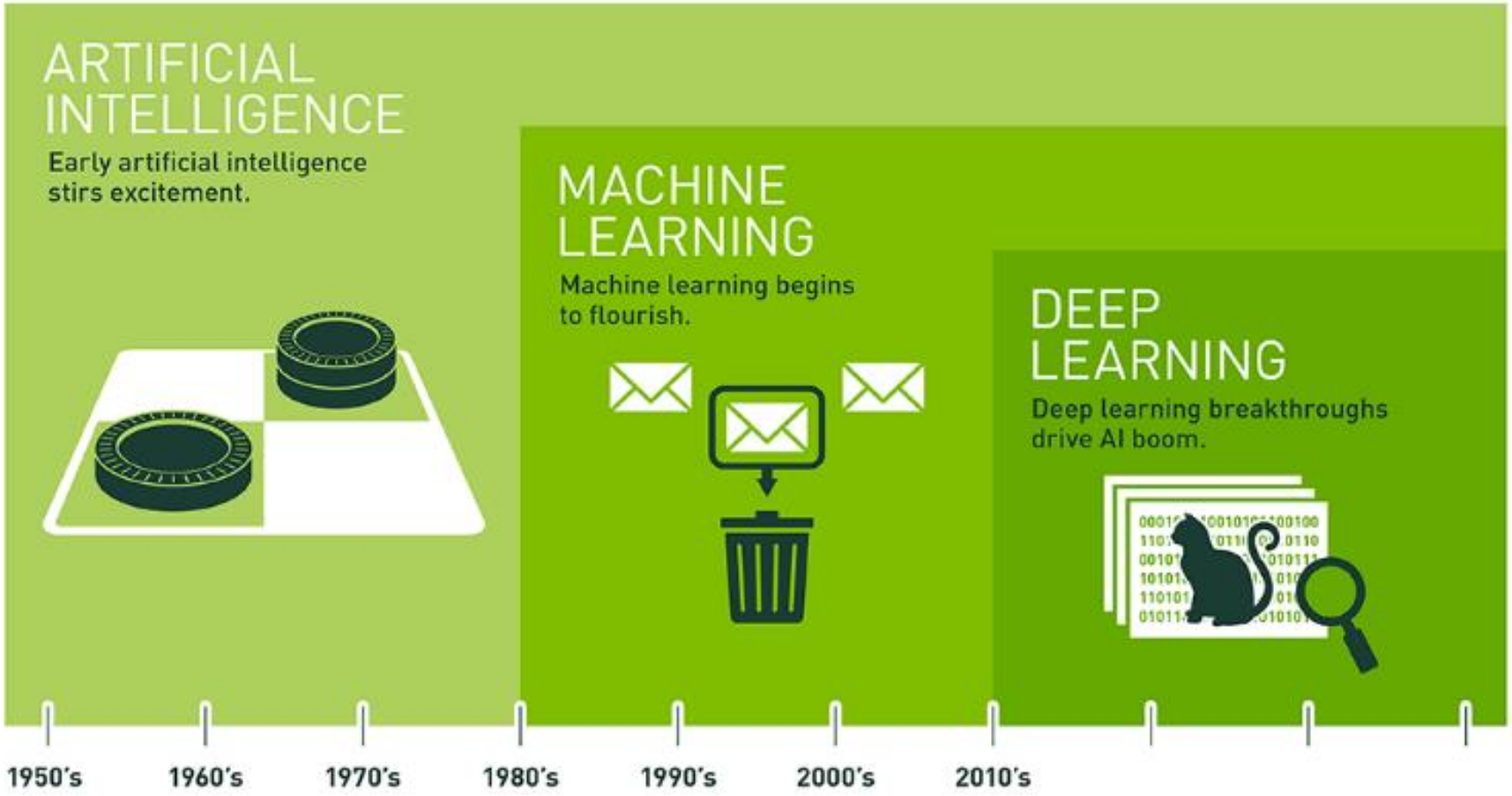
Достижения уклада: индивидуализация производства и потребления, резкое снижение энергоёмкости и материалоёмкости производства, конструирование материалов и организмов с заранее заданными свойствами

Гуманитарное преимущество: существенное увеличение продолжительности и качества жизни человека и животных

На 2010 год доля производительных сил пятого технологического уклада в наиболее развитых странах составляла примерно 60%, четвёртого — 20%, шестого — около 5%. По последним расчетам учёных, шестой технологический уклад в этих странах фактически наступит в 2014–2018 годах.

Технологии ИИ меняют мир

Эволюция искусственного интеллекта



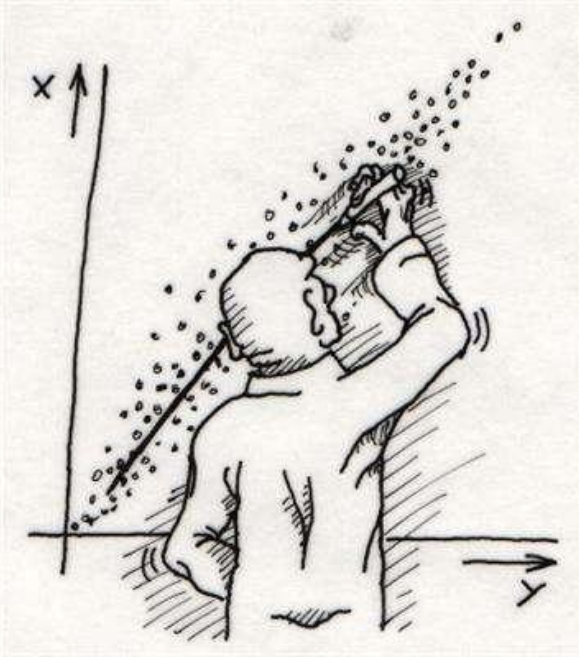
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

На полях:

Глубокое обучение – одна из новейших технологий машинного обучения

Машинное обучение – это ...

- 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, когда данных много, знаний мало
- тысячи алгоритмов
- около 100 000 научных публикаций в год

Основная задача машинного обучения

Этап №1 – обучение с учителем

- **На входе:**
  данные – выборка прецедентов «объект → ответ»
- **На выходе:**
  алгоритм, по любому объекту предсказывающий ответ

Этап №2 – применение

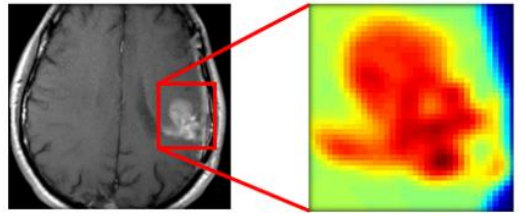
- **На входе:**
  данные – новый объект
- **На выходе:**
  предсказание ответа на новом объекте

На полях:

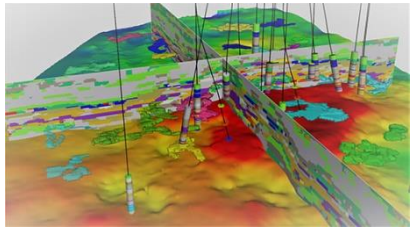
Если нет данных,
то нет
и машинного обучения

Примеры задач машинного обучения

- **Медицинская диагностика:**
  объект – данные о пациенте на текущий момент
  ответ – диагноз / лечение / риск исхода

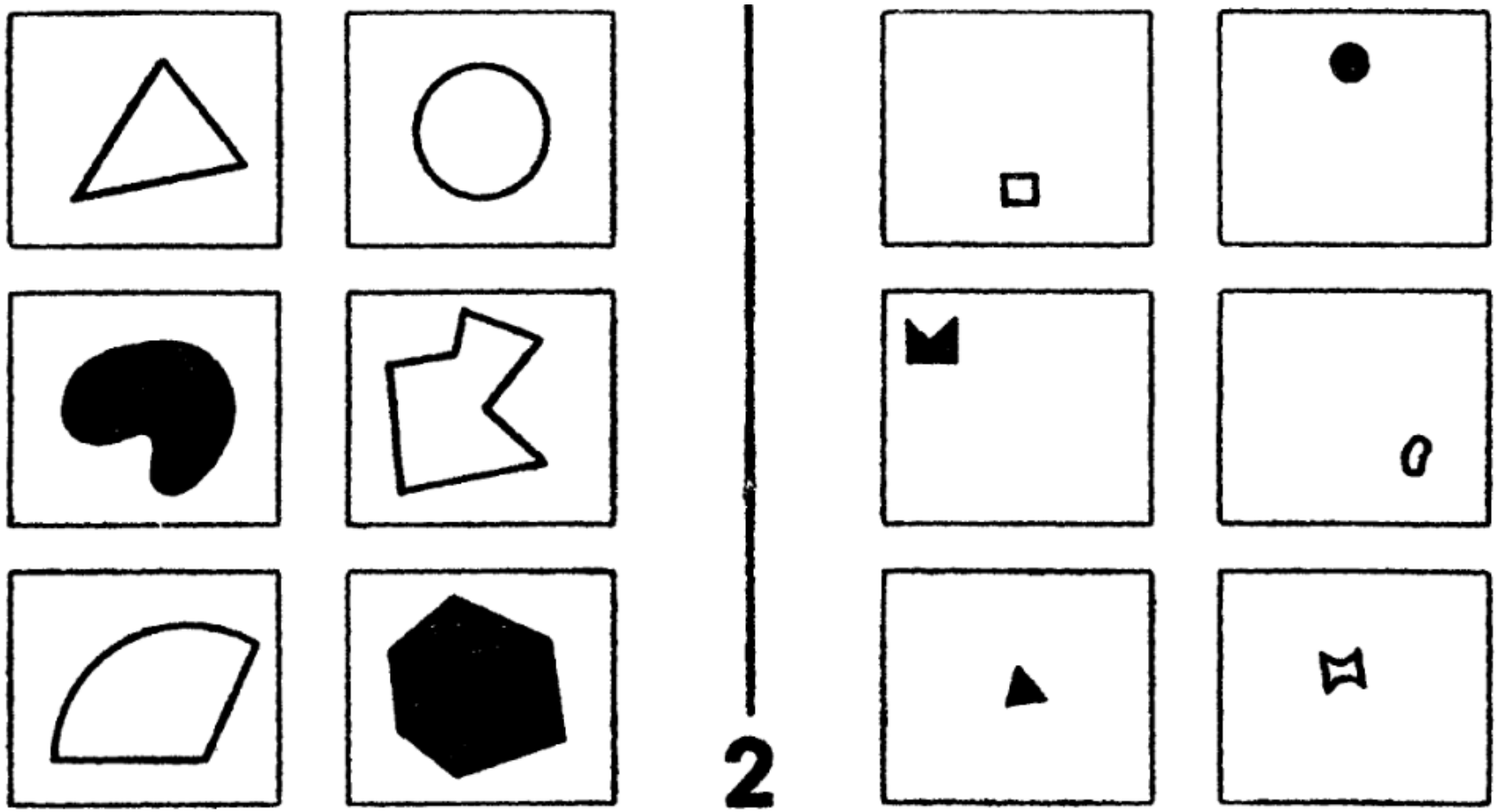


- **Поиск месторождений полезных ископаемых:**
  объект – данные о геологии района
  ответ – есть/нет месторождение

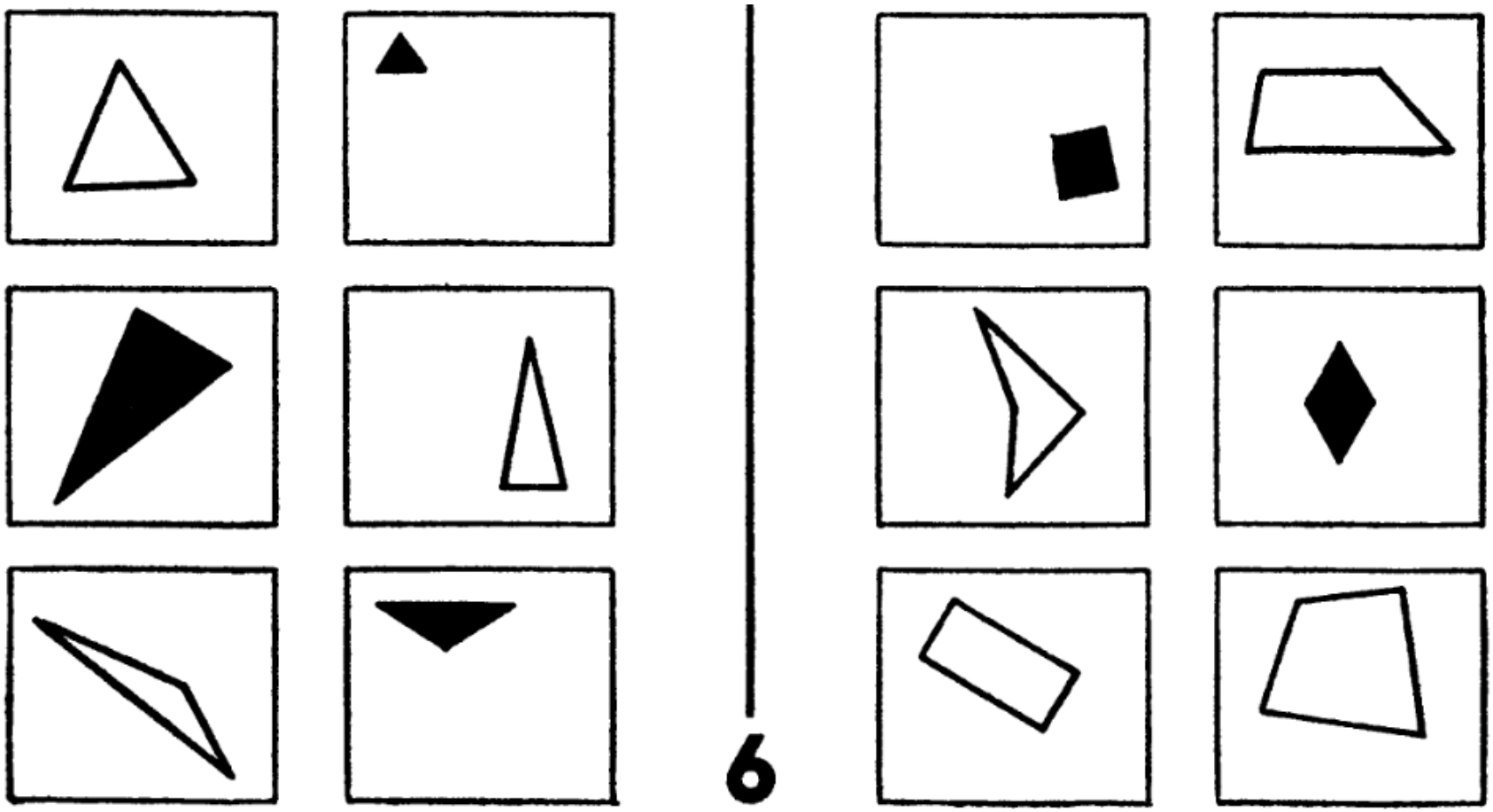


- **Управление технологическими процессами:**
  объект – данные о сырье и управляющих параметрах
  ответ – количество/качество полезного продукта

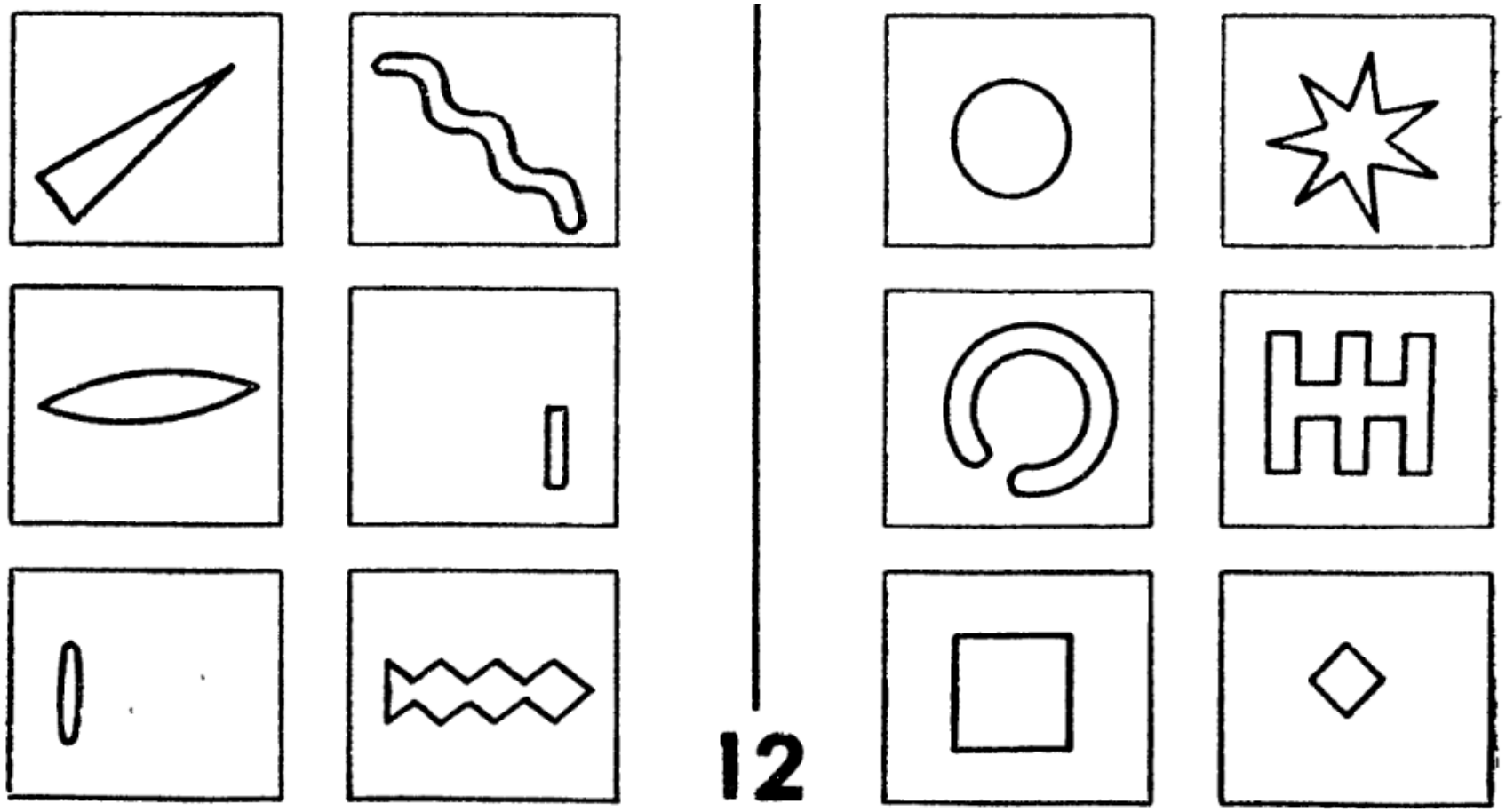
Обучающая выборка: по 6 объектов каждого из двух классов.
Требуется найти правило классификации.



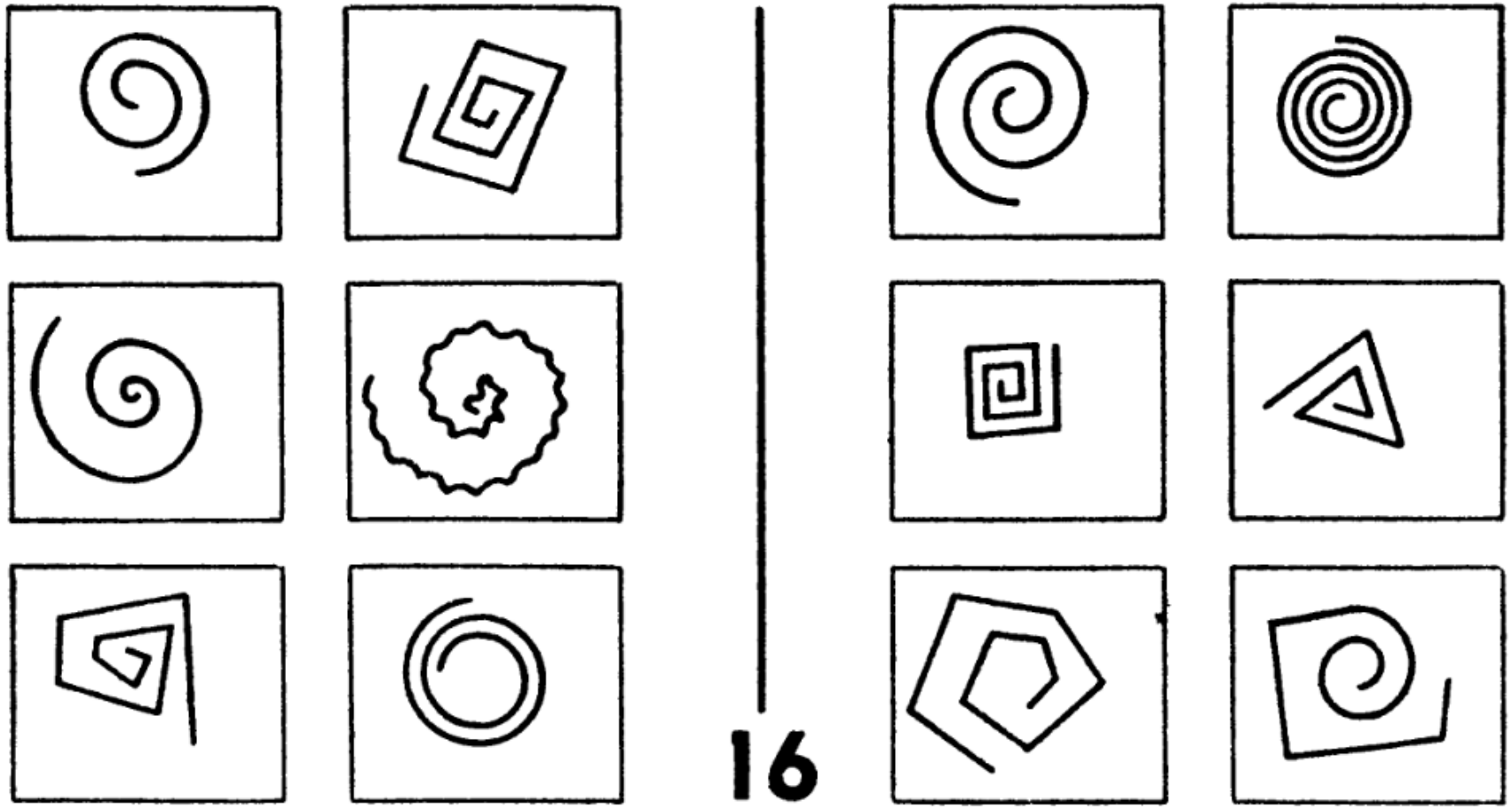
Что даёт нам уверенность, что мы нашли верное правило?
1. Безошибочная классификация примеров обучающей выборки



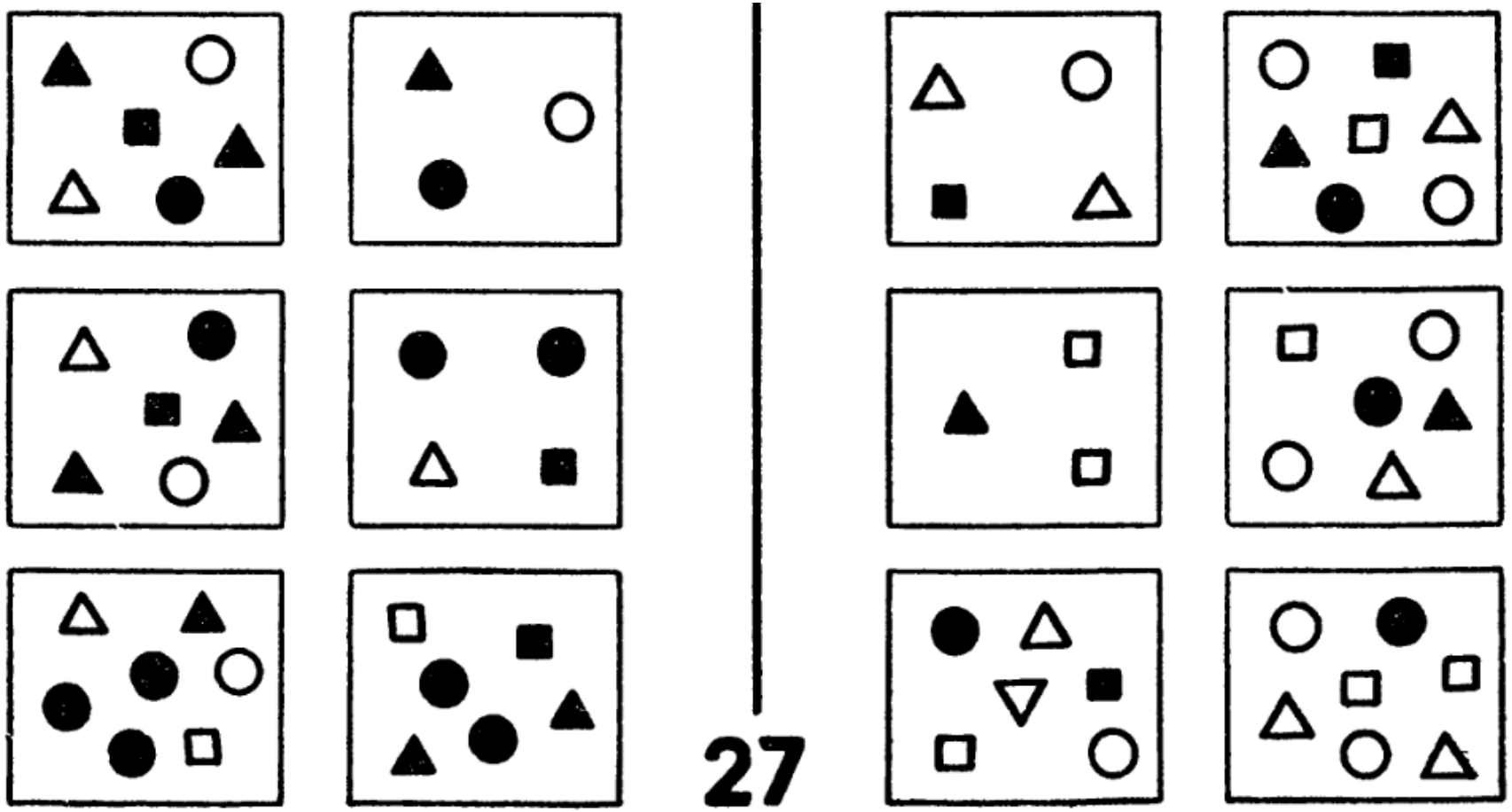
Что ещё даёт нам уверенность, что мы нашли верное правило?
2. Простота и определённое «изящество» найденного правила.



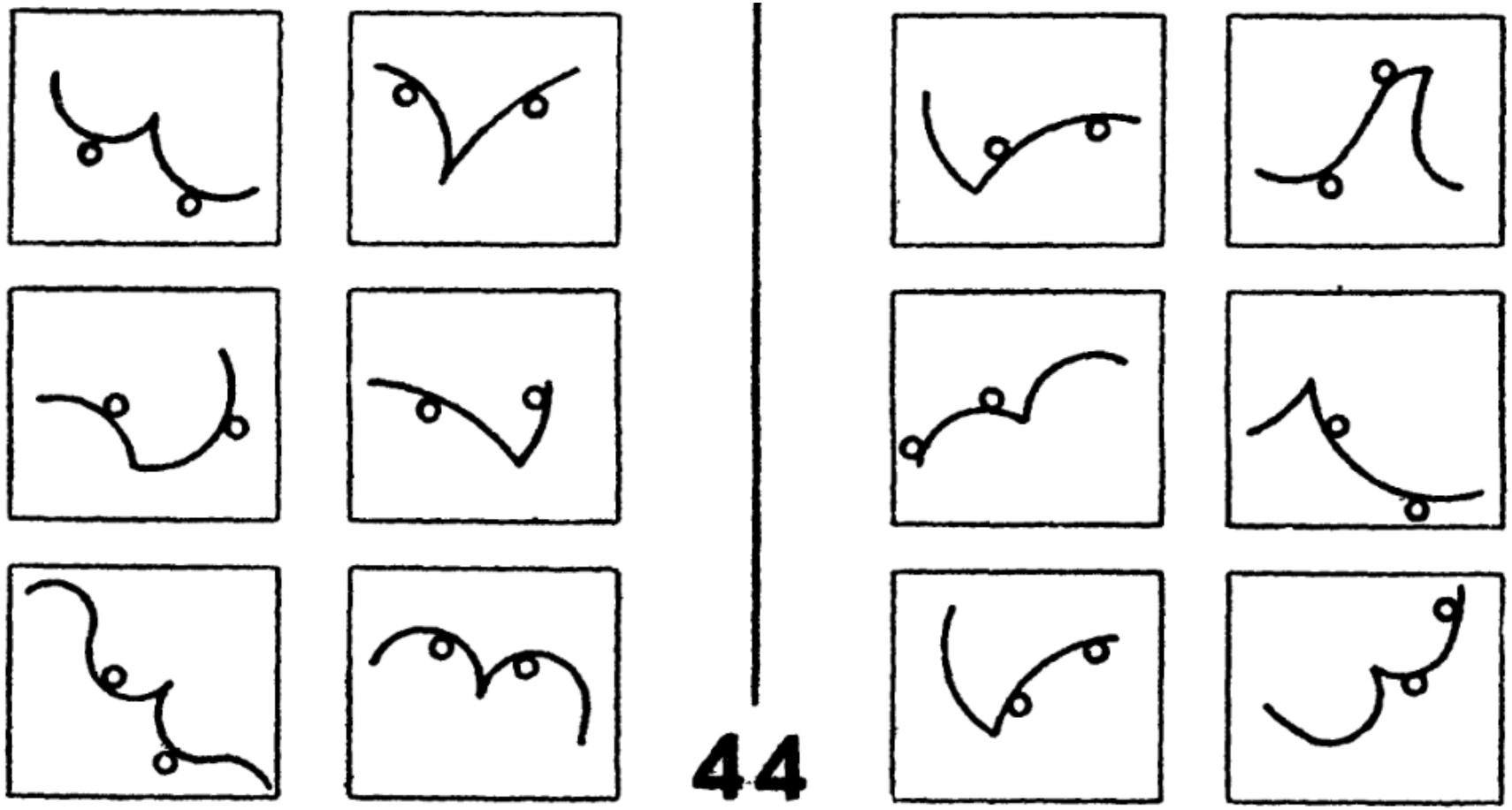
Мы решаем эти задачи почти мгновенно. Чем мы пользуемся?
Почему для компьютера они столь сложны?



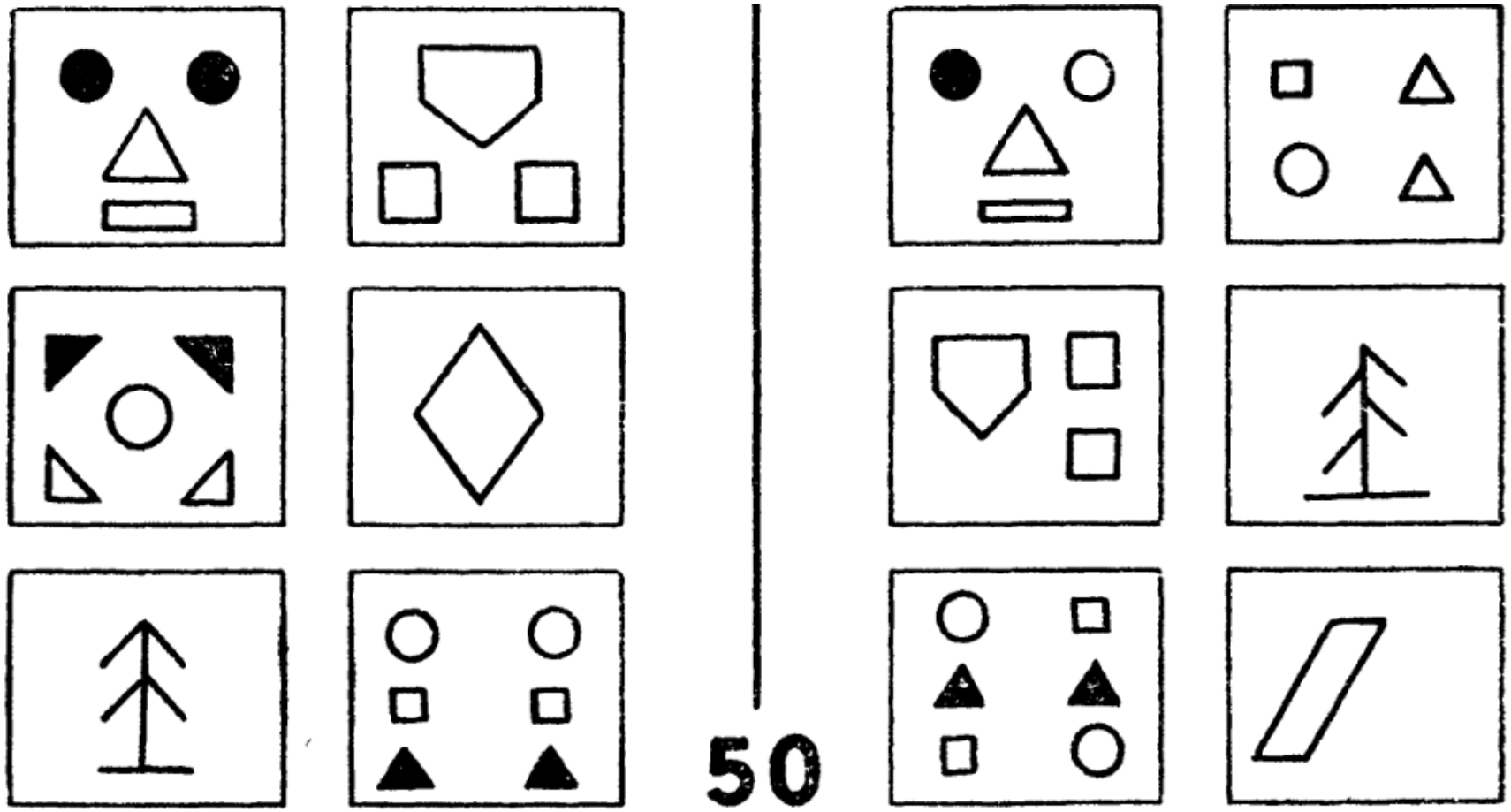
Нужно ли закладывать знания геометрии в явном виде?
Или возможно выучить геометрические понятия на примерах?



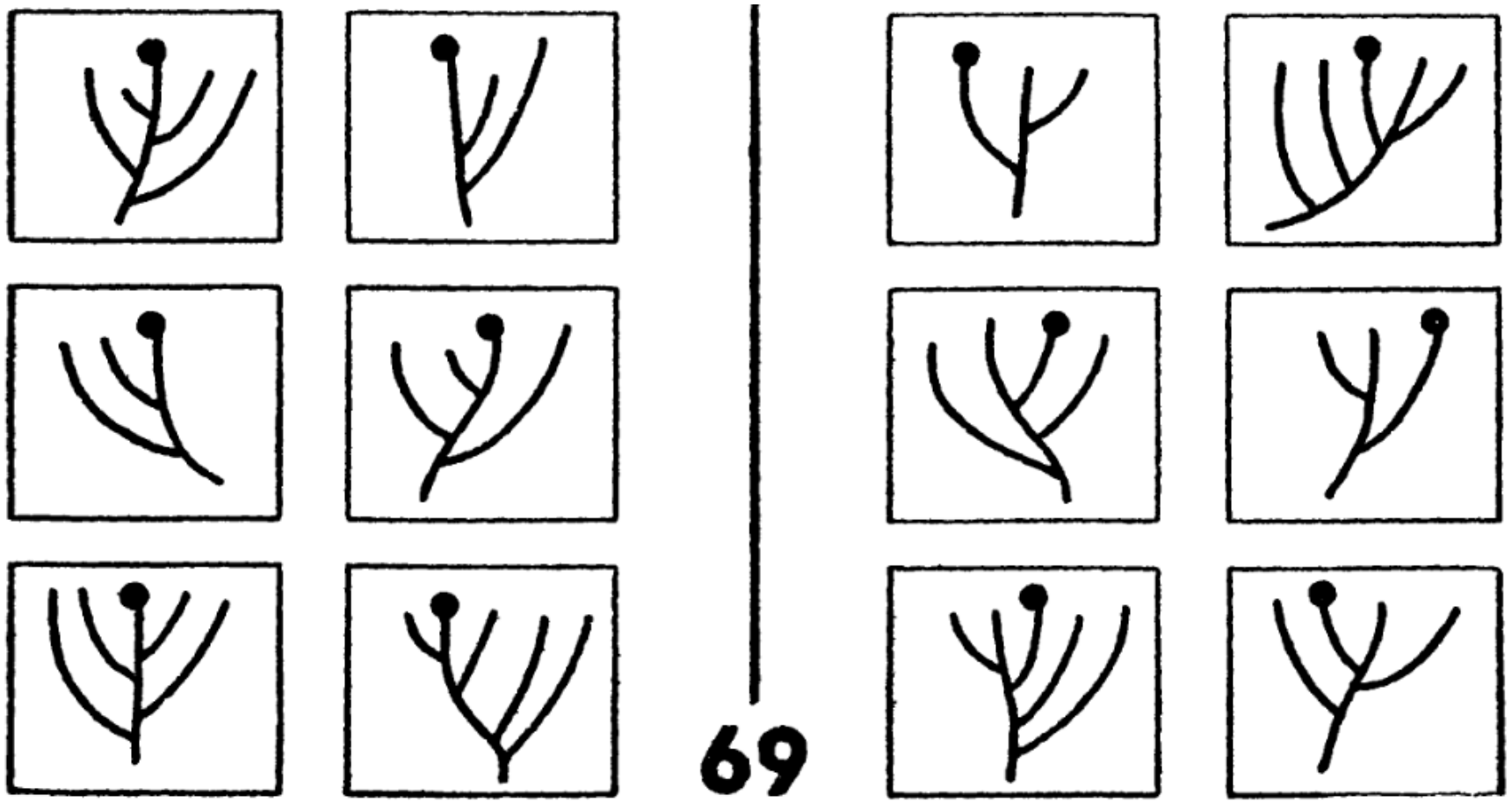
Как вычислять полезные признаки по сложным сырым данным?
Возможно ли поручить перебор признаков и моделей машине?



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



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



Примеры задач машинного обучения

- **Информационный поиск в Интернете:**
  объект – данные о паре «запрос и документ»
  ответ – оценка релевантности документа запросу
- **Продажа рекламы в Интернете:**
  объект – данные о тройке «пользователь, страница, баннер»
  ответ – оценка вероятности клика
- **Рекомендательные системы в Интернете:**
  объект – данные о паре «пользователь, товар»
  ответ – оценка вероятности, что пользователь купит товар

Примеры задач машинного обучения

- **Статистический машинный перевод:**
  объект – предложение на естественном языке
  ответ – его перевод на другой язык
- **Перевод речи в текст:**
  объект – аудиозапись речи человека
  ответ – текстовая запись речи
- **Компьютерное зрение:**
  объект – изображение предмета в видеопоследовательности
  ответ – решение (объехать, остановиться, игнорировать)

На полях:

*Прогресс в этих областях связан с «**Большими данными**», англ. «Big Data»*

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% $\rightarrow$ 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

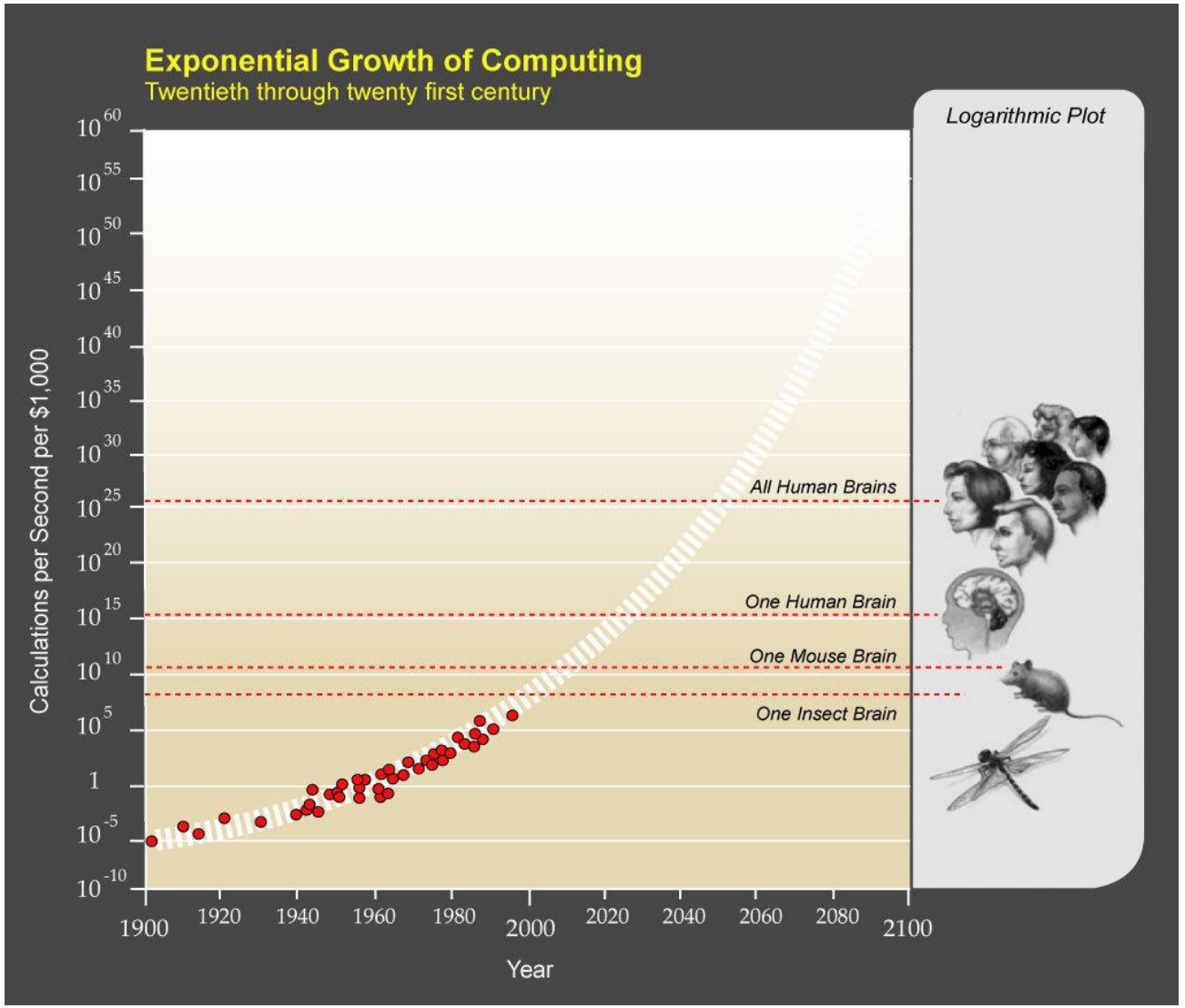
Три предпосылки этого бума

– три перехода количества в качество:

- Повсеместность и доступность компьютерных технологий
  $\rightarrow$ *Накопление больших выборок данных*
- Постепенное развитие математических методов и эвристик
  $\rightarrow$ *Накопление критической массы опыта*
- Достижения микроэлектроники
  $\rightarrow$ *Рост вычислительных мощностей по закону Мура*

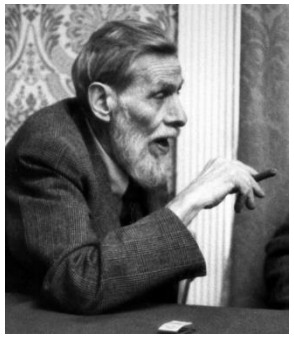
Закон Мура

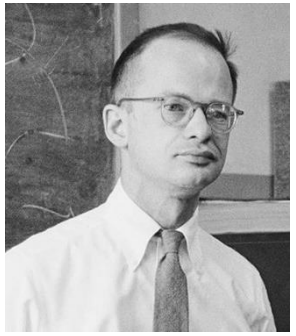
Закон ускоряющейся отдачи
(Рэймонд Курцвейл)

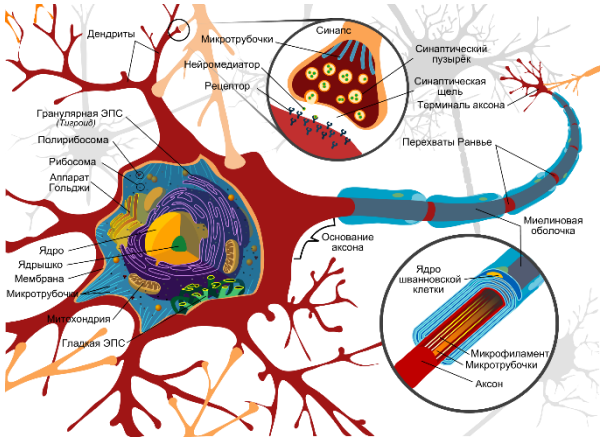


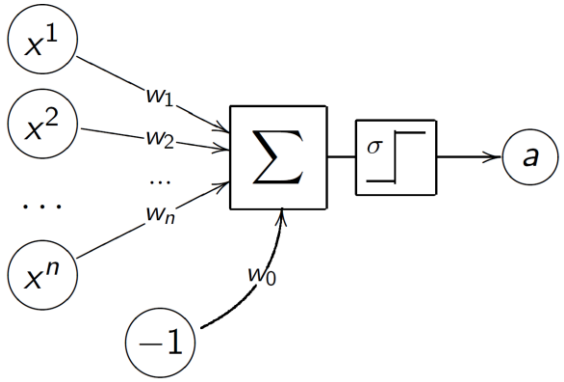
Что такое «искусственные нейронные сети»

Математическая модель нейрона (МакКаллок и Питтс, 1943)



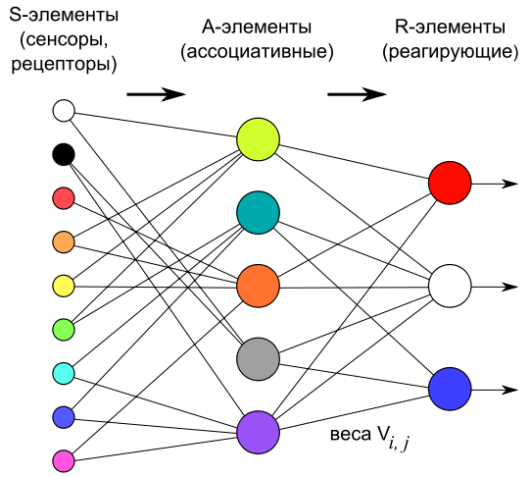




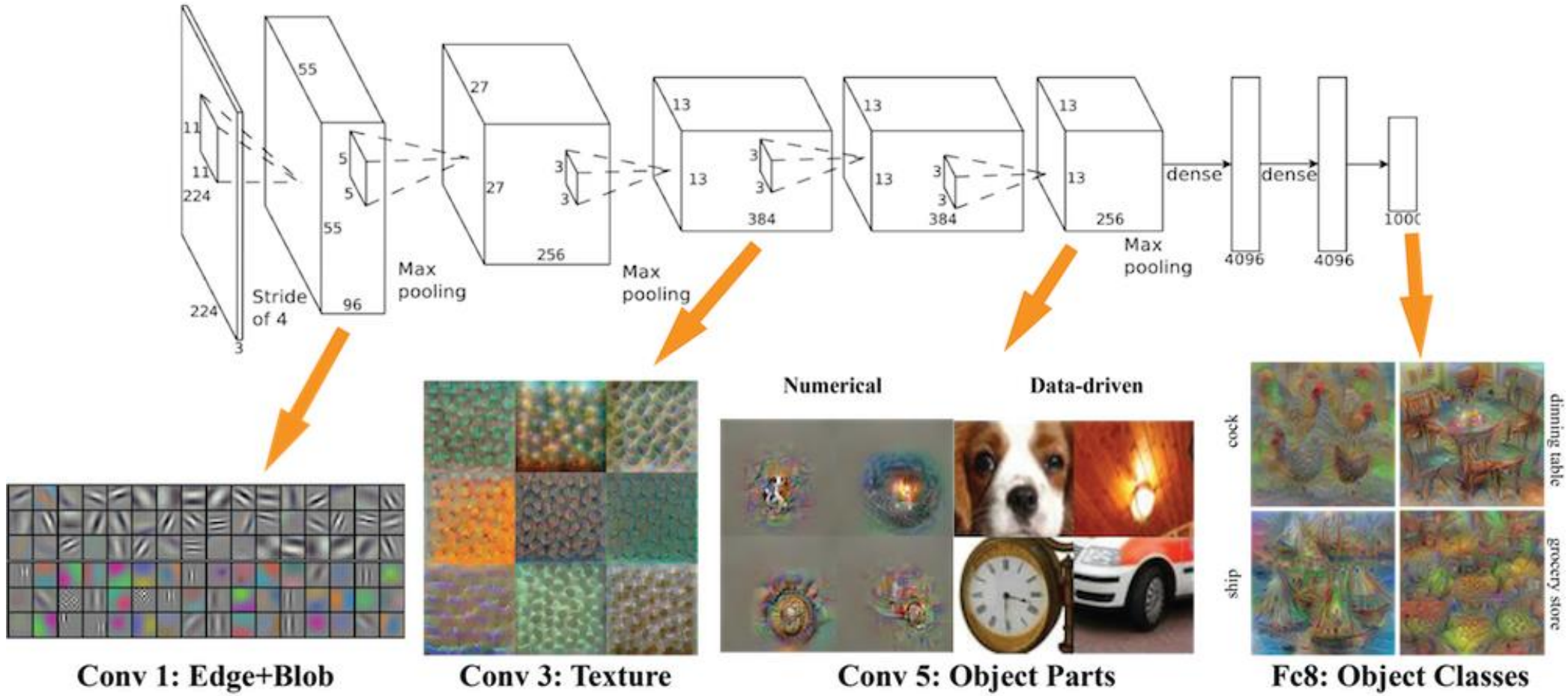


$$a(x, w) = \sigma\left(\sum_{j=1}^{n} w_j x^j - w_0\right)$$

Первый нейрокомпьютер Mark-1 (Фрэнк Розенблатт, 1960)



Что такое «глубокие нейронные сети»



Что такое «цифровая экономика»

- Автоматизация и сокращение издержек повсеместно
- Автономный транспорт и роботизация
- Оптимизация логистики и цепей поставок
- Оптимизация энергетических и транспортных сетей
- Сенсорные сети, мониторинг сельского хозяйства
- Информационные сервисы, распределённая экономика
- Персональная медицина
- Персональные образовательные траектории и карьерный рост
- Автономные системы вооружений

Анализ данных - профессия будущего

- Правда ли, что машины оставят людей без работы
- Какие профессии будут исчезать
- Какие профессии появятся: разработчики, инженеры, обучатели машин, пользователи
- Краудсорсинг – низкоквалифицированный труд будущего
- Что должен знать и уметь data scientist

С чего начать

- **Увлечься!** (математикой, программированием, роботами, конкурсами, проектной деятельностью, …)
- Популярные лекции Андрея Себранта (Яндекс)
- *Педро Домингос.* «Верховный алгоритм». 2016.
- Язык программирования Python
- *Коэльо Л. П., Ричарт В.* Построение систем машинного обучения на языке Python. 2016.
- www.kaggle.com – конкурсы анализа данных