Введение в машинное обучение

Воронцов Константин Вячеславович
(лаборатория Машинного интеллекта МФТИ)

• Прикладной анализ данных •
(кружок для старшеклассников)
4 марта 2018 • МФТИ

Машинное обучение — новый двигатель прогресса

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

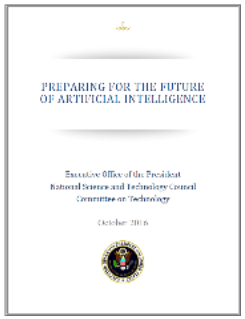
Клаус Мартин Шваб,
президент
Всемирного
экономического
форума

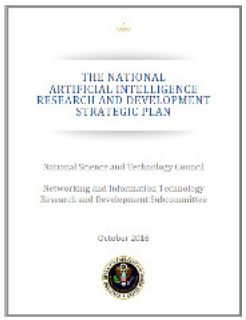
Мир наконец поверил в искусственный интеллект?...
Машинное обучение изменит мир? Или уже меняет?

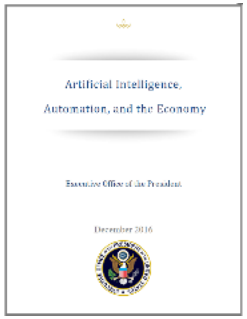
Отчёты Белого Дома США, октябрь 2016

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

- Цифровая и распределённая экономика
- Автоматизация и сокращение издержек
- Автономный транспорт и роботизация
- Оптимизация логистики и цепей поставок
- Оптимизация энергетических сетей
- Мониторинг сельского хозяйства
- Персональная медицина
- Персональные образовательные траектории
- Автономные системы вооружений







Preparing for the Future of Artificial Intelligence. NSTC. 2016.

Бум искусственного интеллекта и нейронных сетей

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

2012 Google X Lab: распознавание видеокадров с котами

2014 Facebook DeepFace распознаёт лица с точностью 97%

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

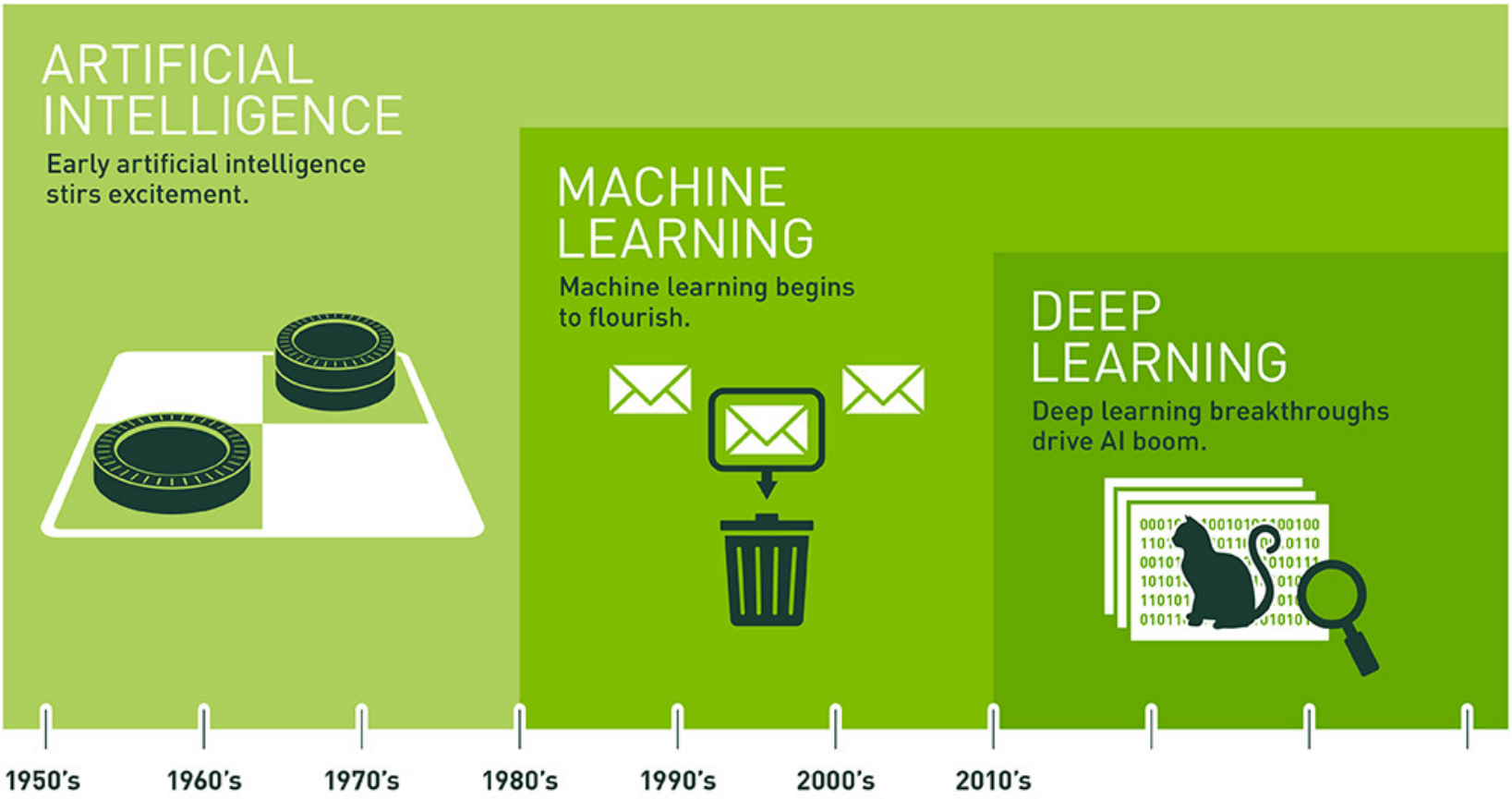
2016 DeepMind, OpenAI: динамическое обучение играм Atari

2016 Google DeepMind обыграл чемпиона мира по игре го

2017 OpenAI обыграл чемпиона мира по компьютерной игре Dota 2

Краткая история развития

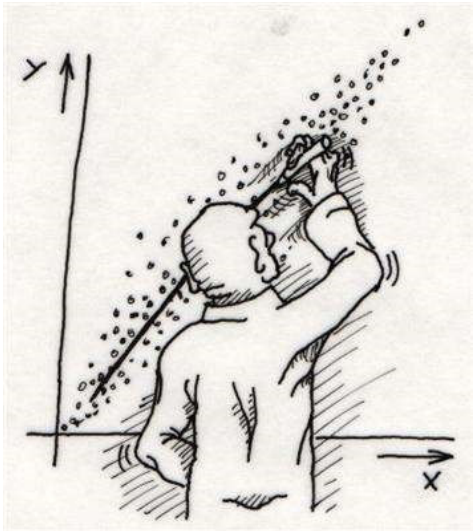
Искусственный интеллект (Artificial Intelligence)
Машинное обучение (Machine Learning)
Нейронные сети и глубокое обучение (Deep Learning)



Что такое машинное обучение

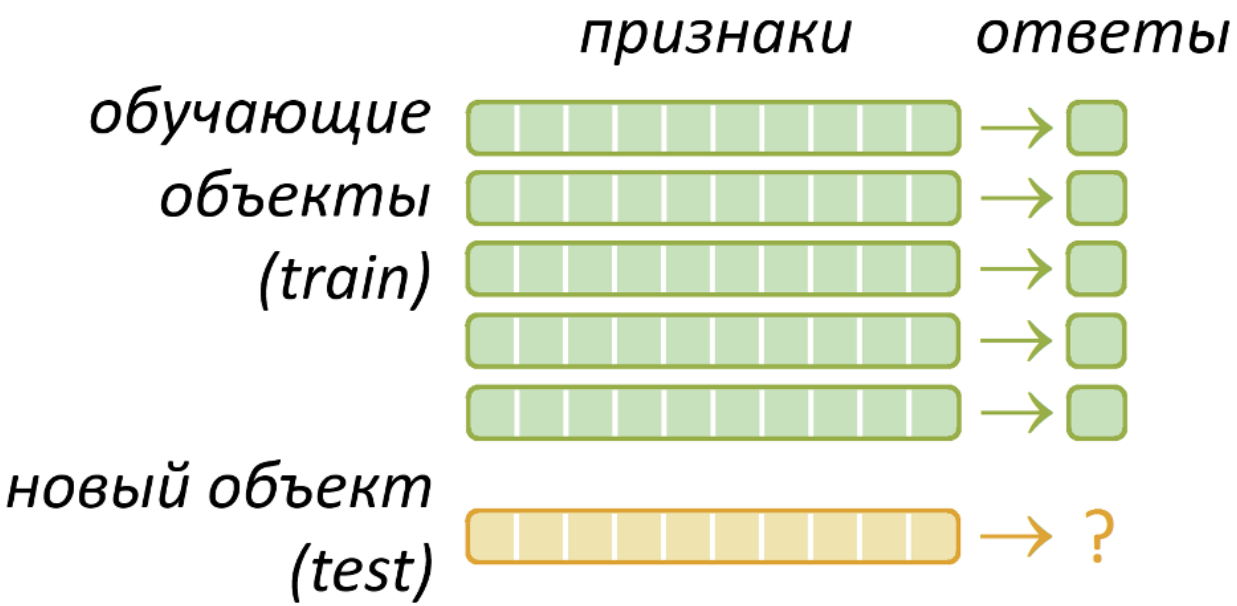
- одна из ключевых информационных технологий будущего
- наиболее успешное направление искусственного интеллекта, вытеснившее экспертные системы и инженерию знаний

- проведение функции через заданные точки в сложно устроенных пространствах
- математическое моделирование, когда знаний мало, данных много
- тысячи методов и алгоритмов
- порядка 10^5 научных публикаций в год



Задача обучения с учителем

В чём состоит задача машинного обучения с учителем?
И что есть «данные» в задачах машинного обучения?



Основной вопрос теории машинного обучения:
как гарантировать, что мы правильно восстановим «закон природы» по наблюдаемым данным?

Типы признаков и типы задач

Типы признаков, по множеству допустимых значений:

- 0 или 1 — *бинарный* признак
- A, B, C, D, E — *номинальный* признак
- 1, 2, 3, 4, 5 — *порядковый* признак
- $\mathbb{R}$ — *количественный* признак

Типы задач, по множеству допустимых ответов:

- 0 или 1 — *классификация* на 2 класса
- $\{1, \ldots, M\}$ — на M непересекающихся классов
- $\{0,1\}^M$ — на M классов, которые могут пересекаться
- $\mathbb{R}$ — задача восстановления регрессии

Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы — диагноз или способ лечения или исход заболевания.

Примеры признаков:

- **бинарные:** пол, головная боль, слабость, тошнота, и т. д.
- **порядковые:** тяжесть состояния, желтушность, и т. д.
- **количественные:** возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- обычно много «пропусков» в данных;
- как правило, недостаточный объём данных;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности (риска | успеха | исхода).

Задачи распознавания месторождений

Объект — геологический район (рудное поле).

Классы — есть или нет полезное ископаемое.

Примеры признаков:

- **бинарные:** присутствие крупных зон смятия и рассланцевания, и т. д.
- **порядковые:** минеральное разнообразие; мнения экспертов о наличии полезного ископаемого, и т. д.
- **количественные:** содержания сурьмы, присутствие в рудах антимонита, и т. д.

Особенности задачи:

- проблема «малых данных» — для редких типов месторождений объектов много меньше, чем признаков.

Задача предсказания оттока клиентов

Объект — абонент в определённый момент времени.

Классы — уйдёт или не уйдёт в следующем месяце.

Примеры признаков:

- **бинарные:** корпоративный клиент, включение услуг, и т. д.
- **номинальные:** тарифный план, регион проживания, и т. д.
- **количественные:** длительность разговоров (входящих, исходящих, СМС, и т. д.), частота оплаты, и т. д.

Особенности задачи:

- сложно идентифицировать факт ухода;
- нужно оценивать вероятность ухода;
- сверхбольшие выборки;
- не ясно, какие признаки вычислять по «сырым» данным.

Задачи биометрической идентификации личности

Идентификация по отпечаткам пальцев

Идентификация по радужной оболочке глаза

Особенности задач:

- нетривиальная предобработка для извлечения признаков;
- высочайшие требования к точности.

Задача ранжирования поисковой выдачи

Объект — пара $\langle$короткий запрос, документ$\rangle$.

Классы — асессорские оценки релевантности.

Примеры признаков:

- **количественные:**
  частота слов запроса в документе,
  число ссылок на документ,
  число кликов на документ: всего, по данному запросу,
- **номинальные:**
  ID пользователя, ID региона, язык запроса.

Особенности задачи:

- оптимизируется не число ошибок, а качество ранжирования;
- сверхбольшие выборки;
- проблема конструирования признаков по сырым данным.

Задача ранжирования в рекомендательных системах

Объект — пара $\langle$клиент, товар$\rangle$
(товары — книги, фильмы, музыка).

Предсказать: вероятность покупки или рейтинг товара.

Примеры признаков:

- **количественные:**
  рейтинг схожих товаров для данного клиента;
  рейтинг данного товара для схожих клиентов;
  вектор интересов клиента;
  вектор интересов товара;

Особенности задачи:

- сверхбольшие разреженные данные;
- интересы скрыты, их надо сначала выявить.

Задача прогнозирования стоимости недвижимости

Объект — квартира в Москве.

Примеры признаков:

- **бинарные:** наличие балкона, лифта, мусоропровода, охраны, и т. д.
- **номинальные:** район города, тип дома (кирпичный/панельный/блочный/монолит), и т. д.
- **количественные:** число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т. д.

Особенности задачи:

- выборка неоднородна, стоимость меняется со временем;
- разнотипные признаки;
- для линейной модели нужны преобразования признаков.

Задача прогнозирования объёмов продаж

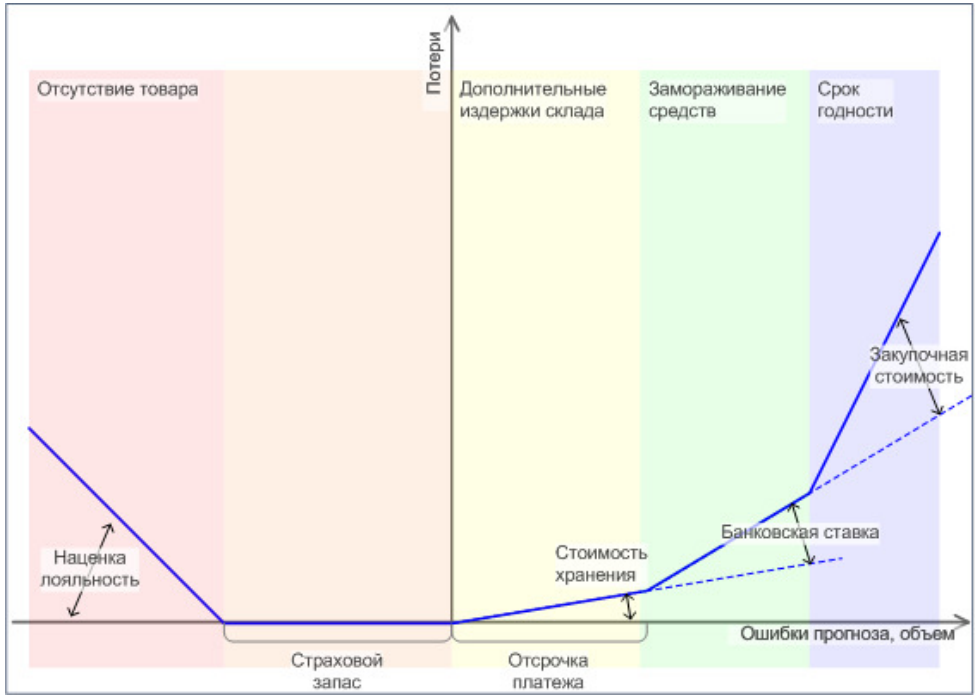
Объект — тройка $\langle$товар, магазин, день$\rangle$.

Примеры признаков:

- **бинарные:** выходной день, праздник, промоакция, и т. д.
- **количественные:** объёмы продаж в предшествующие дни.

Особенности задачи:

- функция потерь не квадратична и даже не симметрична;
- разреженные данные.



Конкурс kaggle.com: TFI Restaurant Revenue Prediction

Объект — место для открытия нового ресторана.

Предсказать — прибыль от ресторана через год.

Примеры признаков:

- демографическими свойствами района;
- цены на недвижимость поблизости;
- маркетинговые данные: наличие школ, офисов и т.д.

Особенности задачи:

- мало объектов, много признаков;
- разнотипные признаки;
- есть нетипичные объекты — «выбросы»;
- разнородные объекты (возможно, имеет смысл строить разные модели для мелких и крупных городов).

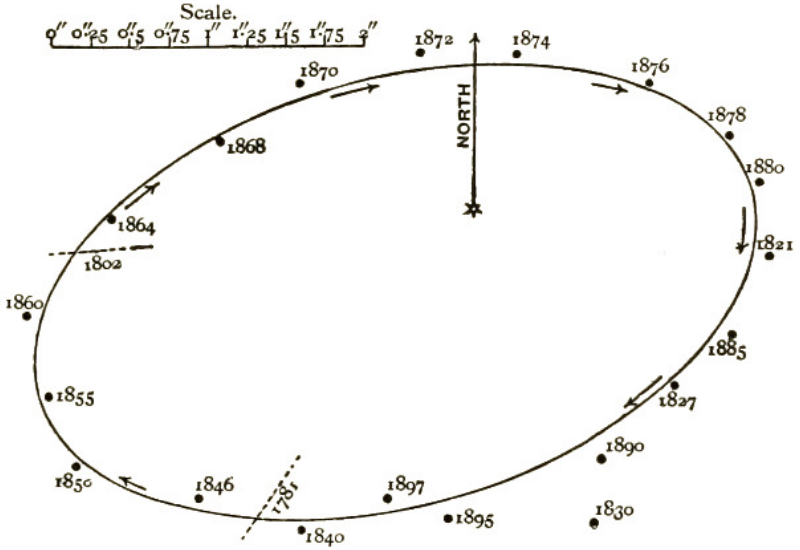
Метод наименьших квадратов (Гаусс, 1795)

Линейная модель регрессии:

$$a(x, w) = \sum_{j=1}^{n} w_j f_j(x), \qquad w \in \mathbb{R}^n.$$

Метод наименьших квадратов:

$$Q(w) = \sum^{\ell} \big(a(x_i, w) - y_i\big)^2 \to \min_w.$$



Карл Фридрих Гаусс (1777–1855)

«Our principle, which we have made use of since 1795, has lately been published by Legendre...»

C.F.Gauss. Theory of the motion of the heavenly bodies moving about the Sun in conic sections. 1809.

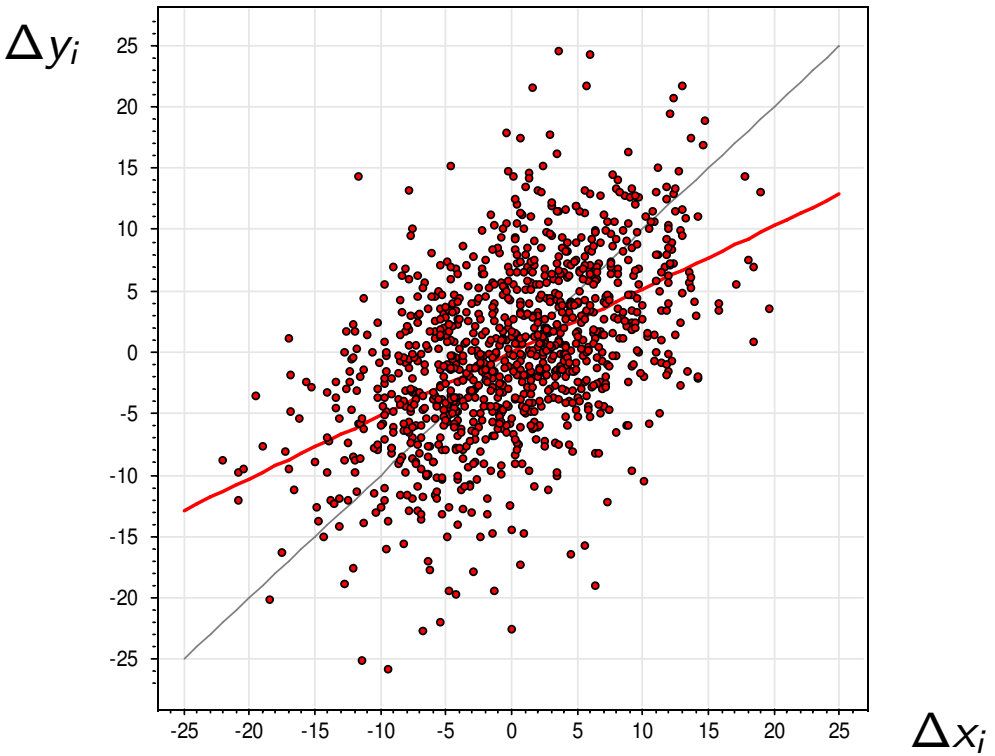
Откуда пошло название «регрессия» (Гальтон, 1886)

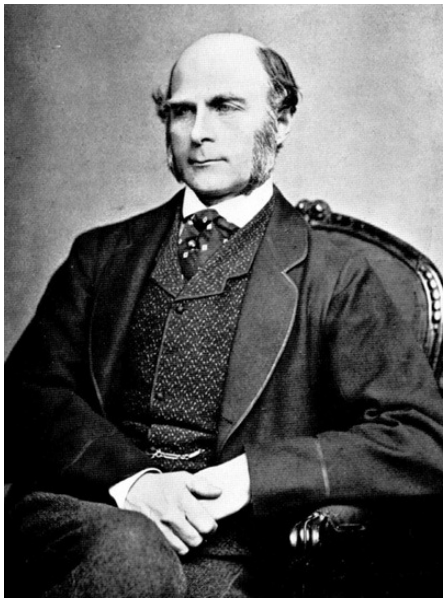
Исследование наследственности роста.

отклонение роста от среднего в популяции:

Δx_i — отклонение роста отца

Δy_i — отклонение роста взрослого сына

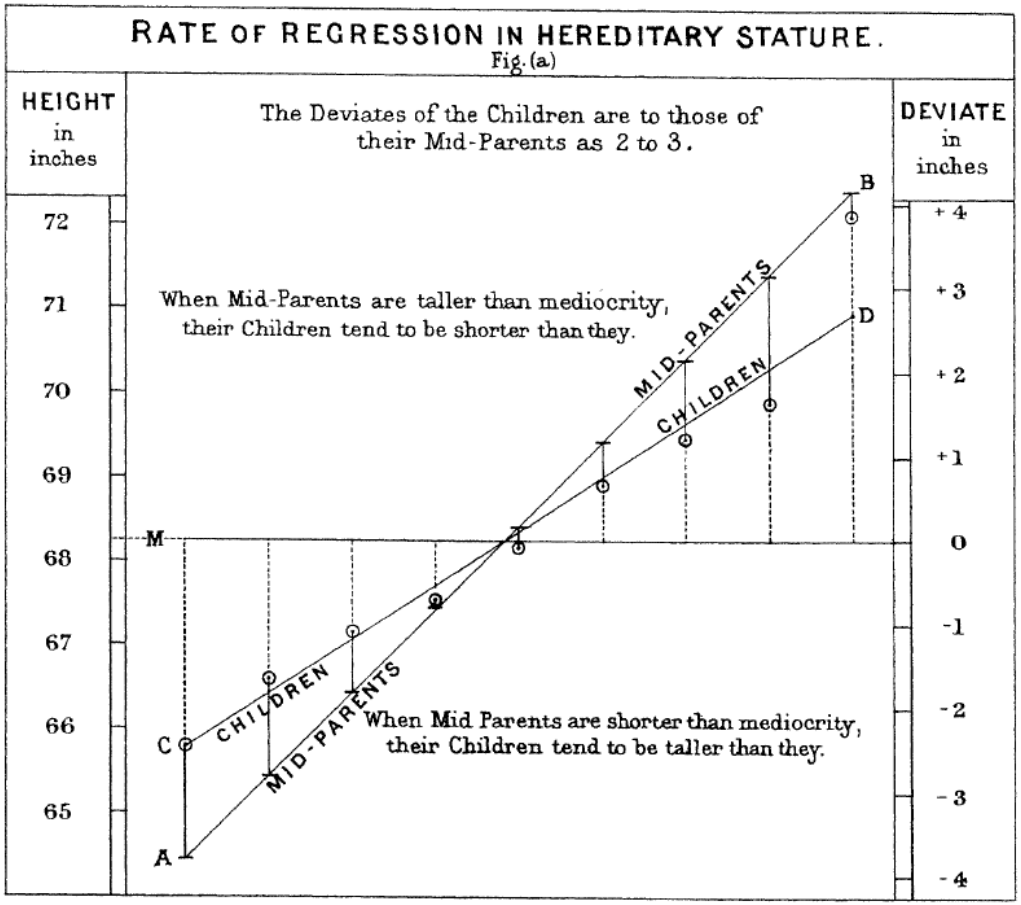


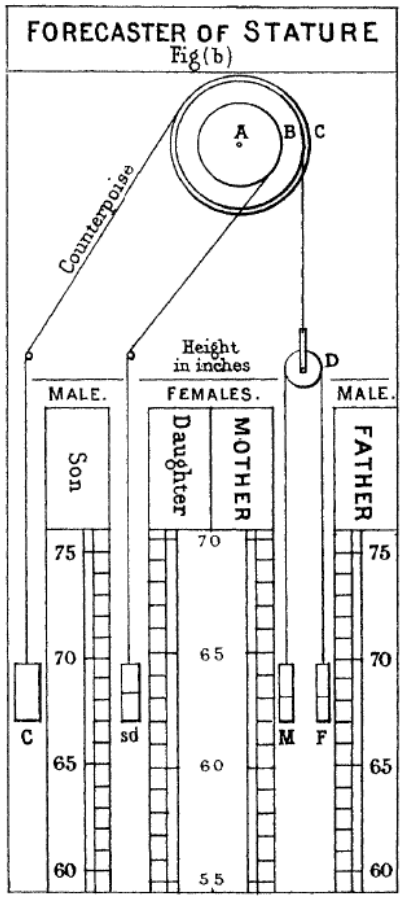


Фрэнсис Гальтон (1822–1911)

Regression to mediocrity — возвращение к посредственности

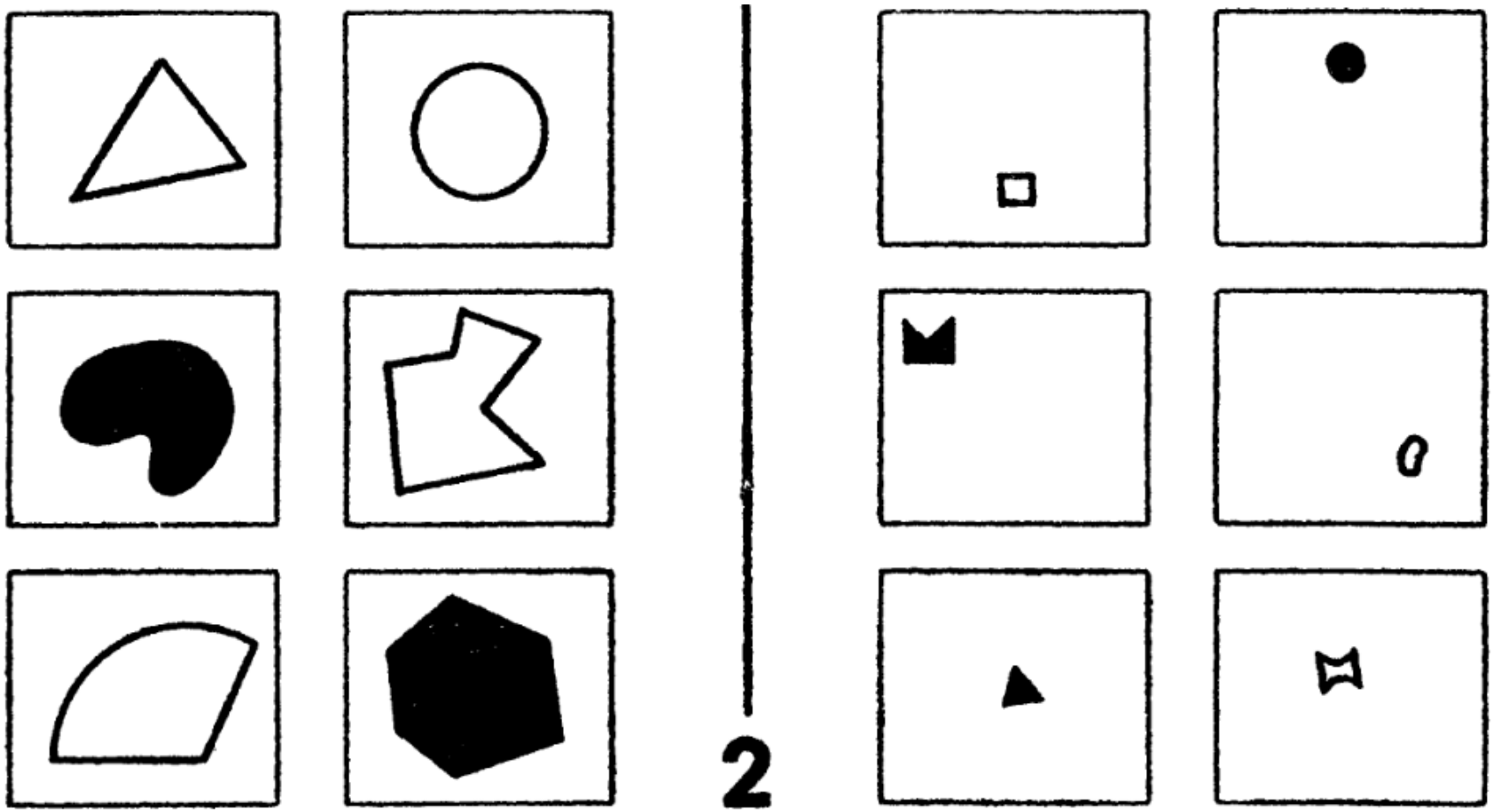
Скрытый смысл: «регрессия» — сначала данные, потом модель





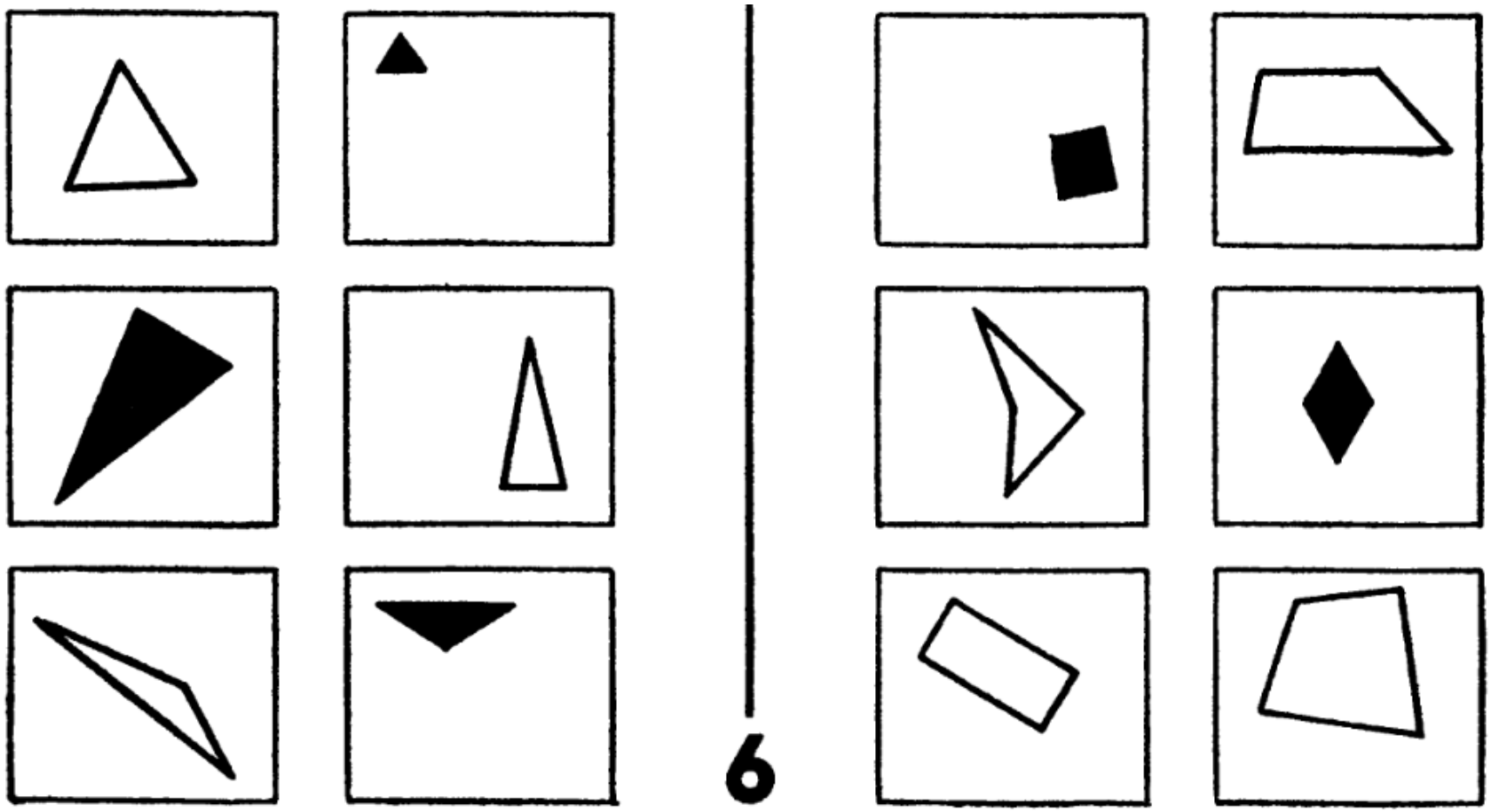
Galton F. Regression towards mediocrity in hereditary stature. 1886.

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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

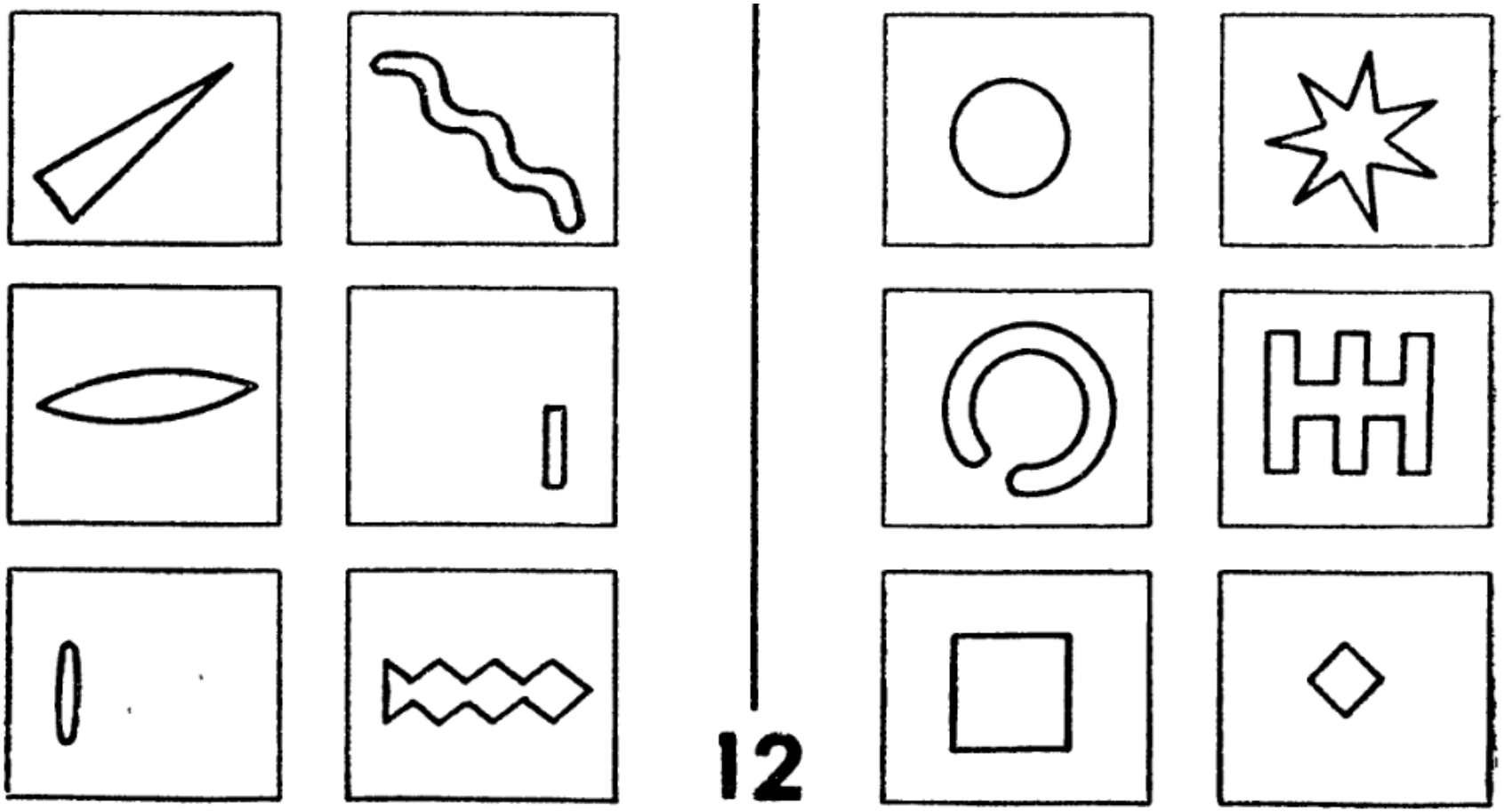
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Что даёт нам уверенность, что мы нашли верное правило?

1. Безошибочная классификация примеров обучающей выборки.

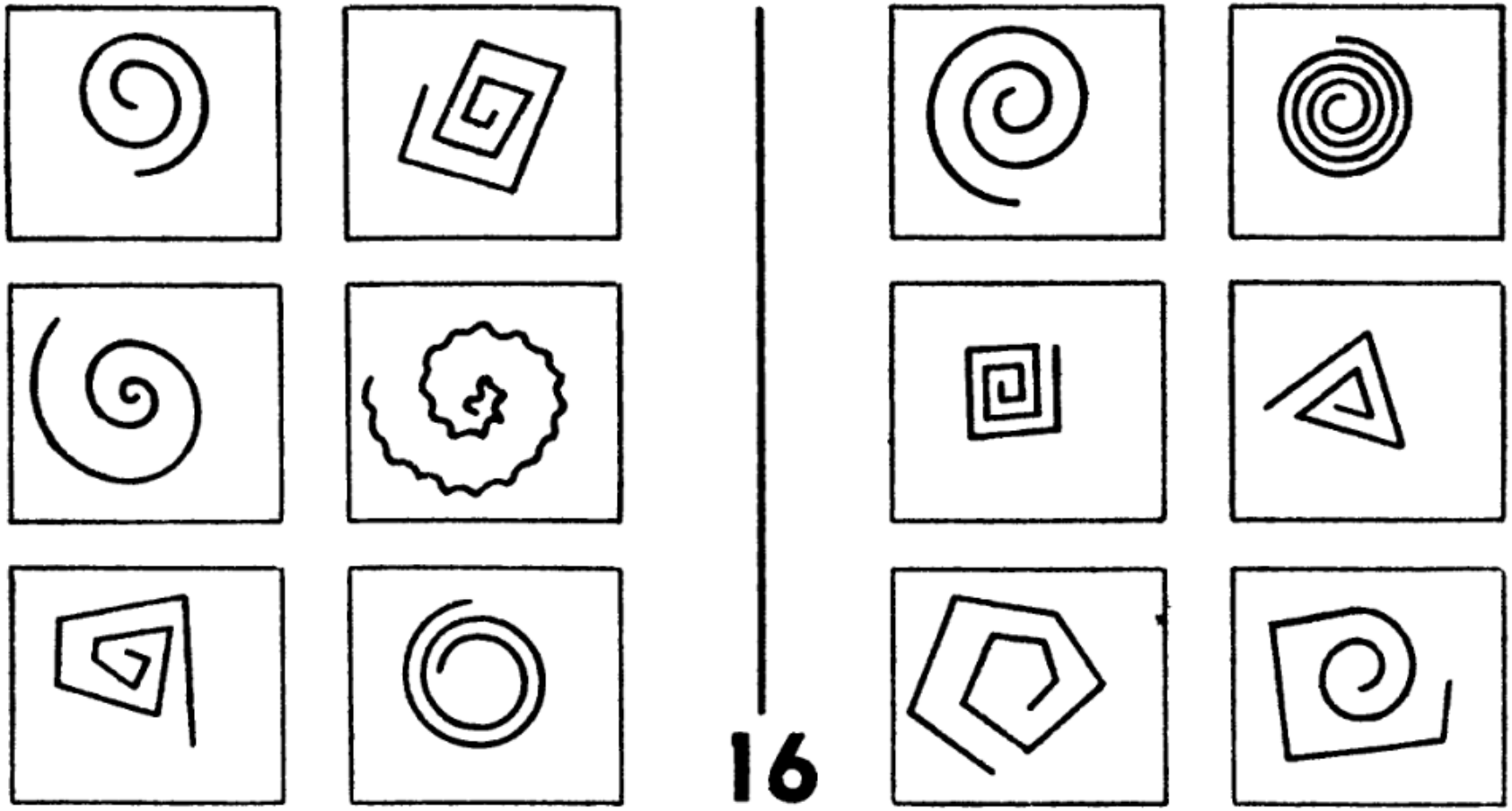
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Что даёт нам уверенность, что мы нашли верное правило?

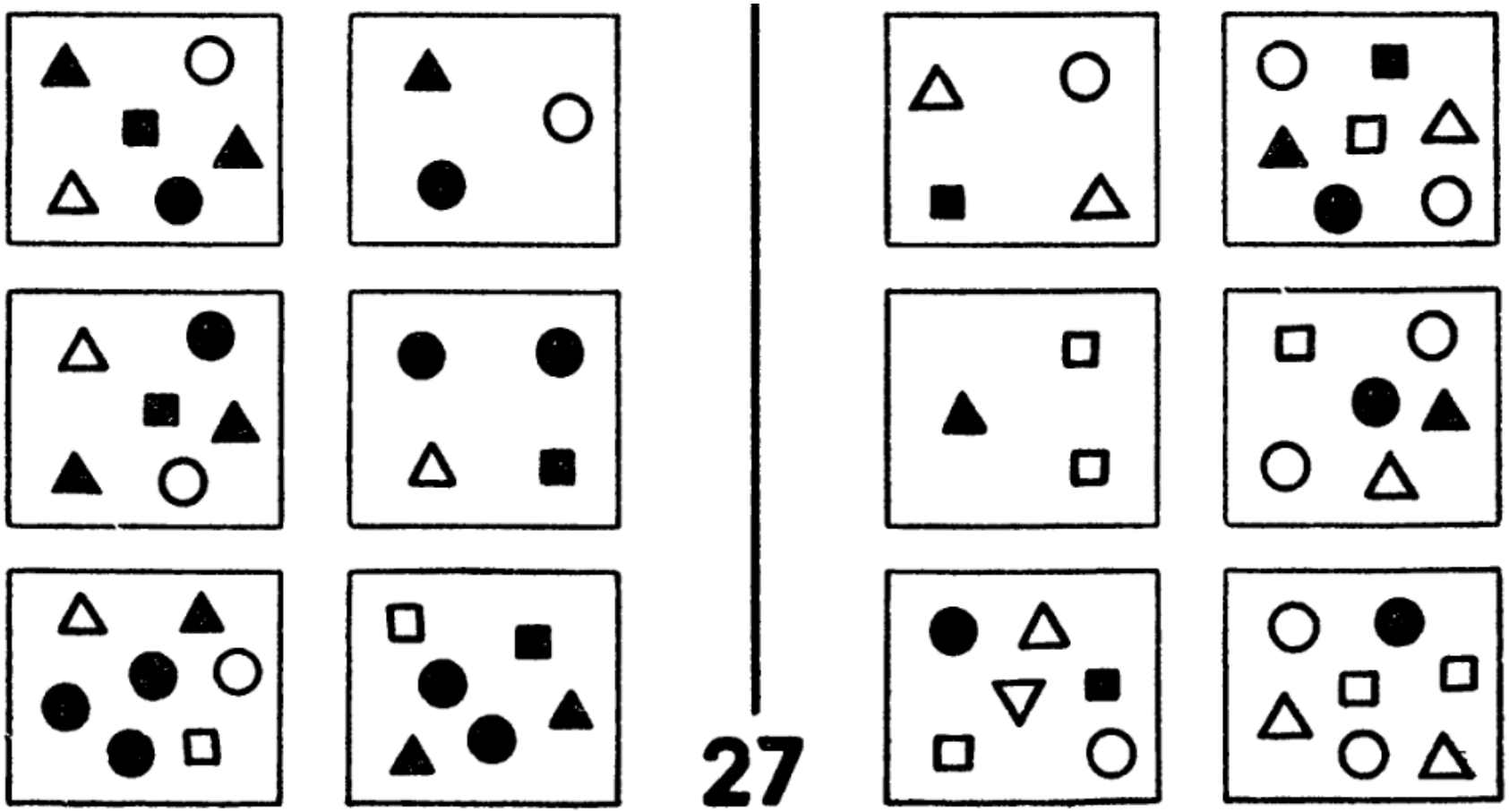
2. Простота и определённое «изящество» найденного правила.

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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

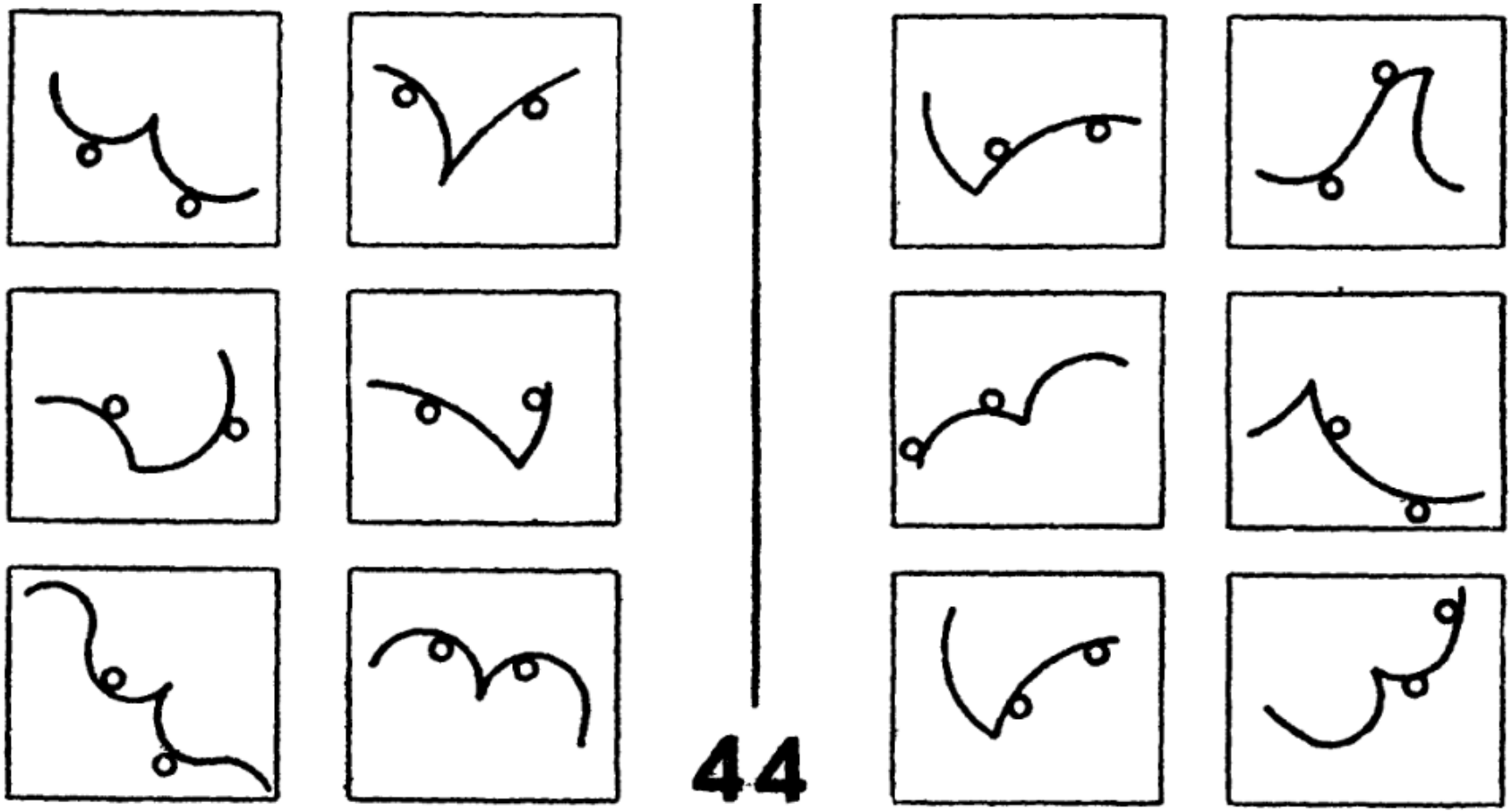
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Нужно ли закладывать знания геометрии в явном виде?

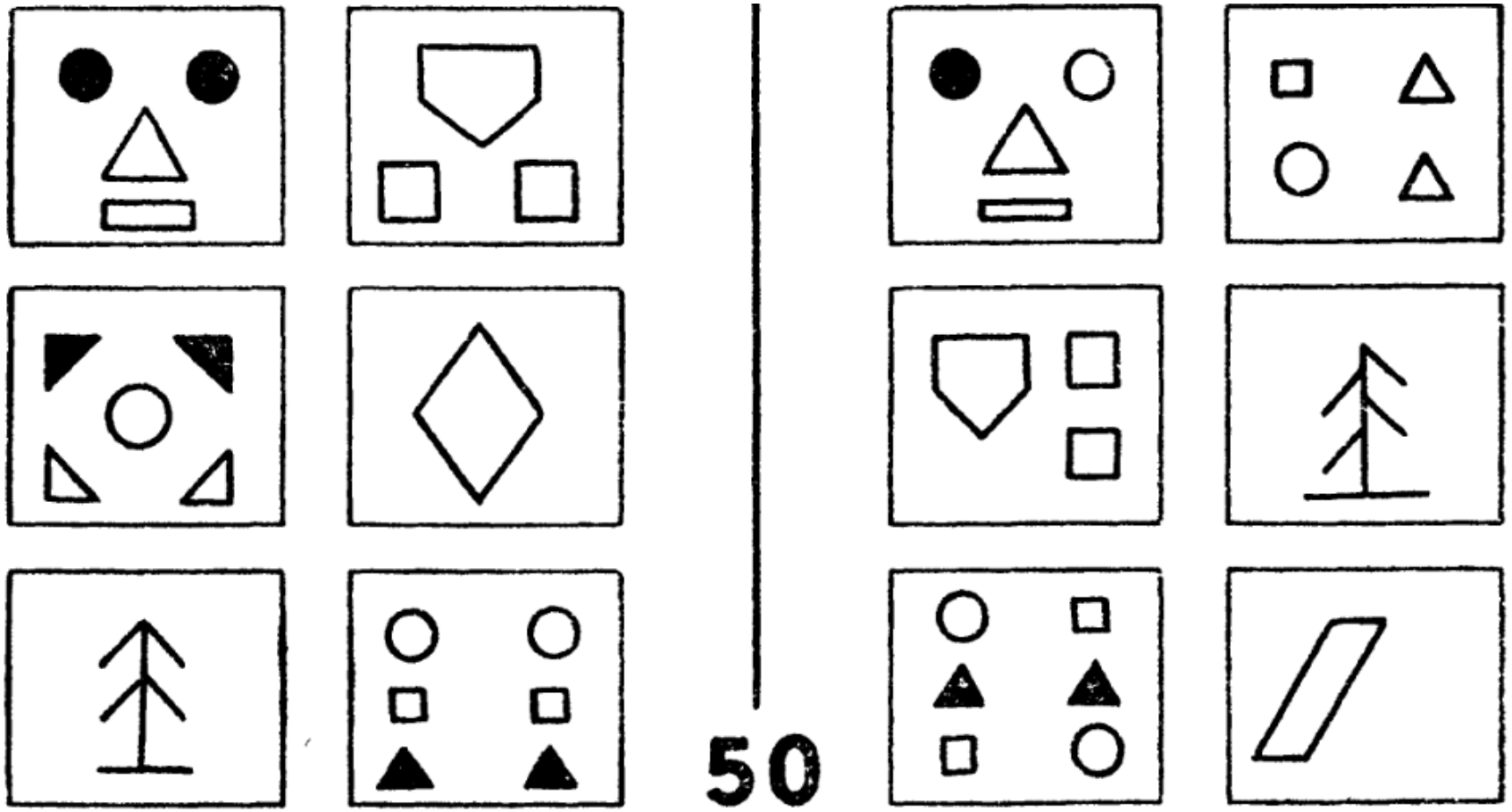
Или возможно выучить геометрические понятия на примерах?

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



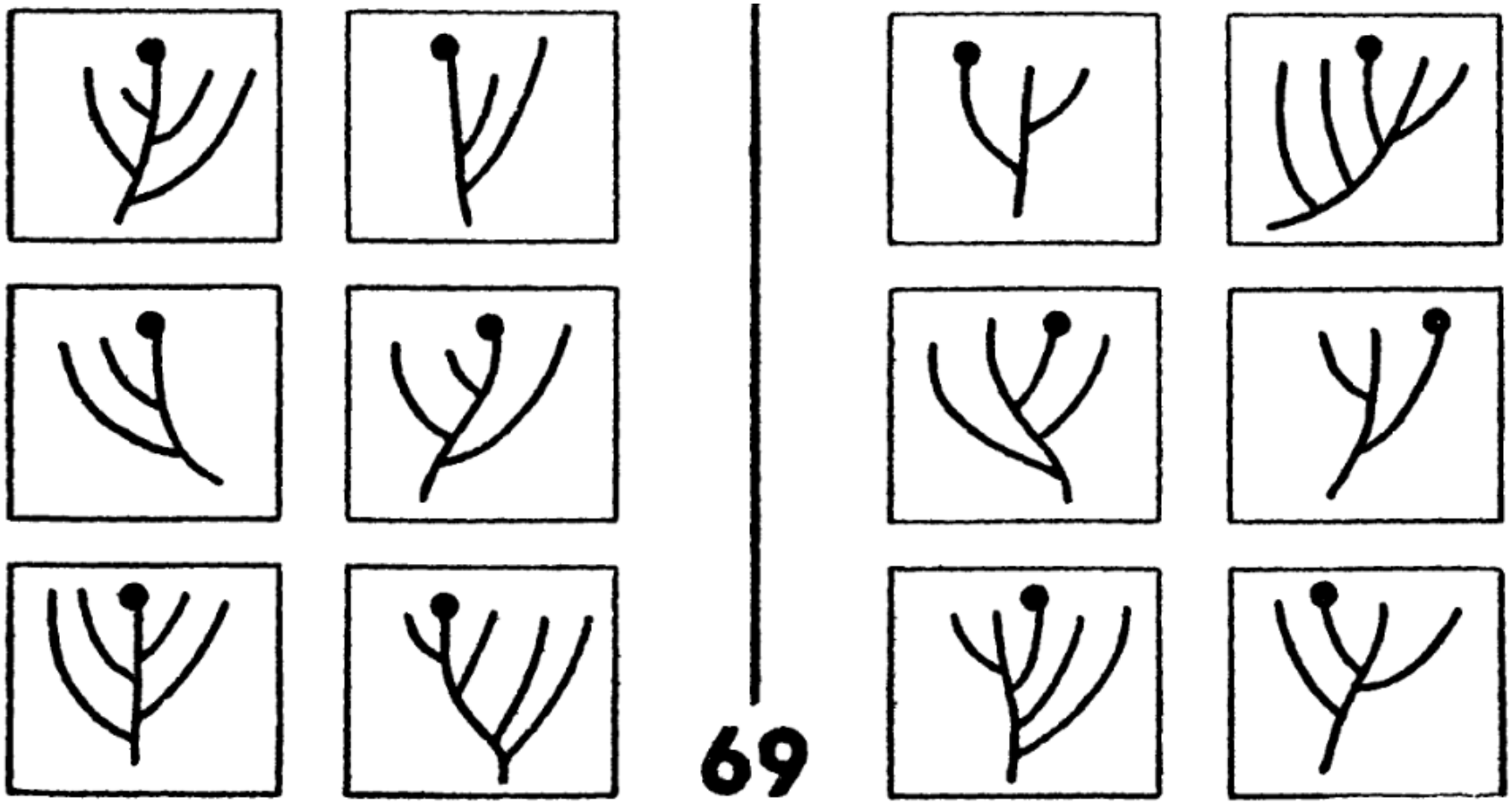
Как вычислять полезные признаки по сложным сырым данным?
Возможно ли поручить перебор признаков и моделей машине?

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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

Основные школы машинного обучения

1. *символизм* – поиск логических закономерностей
   - Decision Tree, Rule Induction
2. *коннекционизм* – обучаемые нейронные сети
   - BackPropagation, Deep Belief Nets, Deep Learning
3. *эволюционизм* – саморазвитие сложных моделей
   - Genetic Algorithms, Genetic Programming
4. *байесионизм* – оценивание распределений параметров
   - Naive Bayes, Bayesian Networks, Graphical Models
5. *аналогизм* – «близким объектам близкие ответы»
   - kNN, RBF, SVM, Kernel Smoothing

⊕ *композиционизм* – кооперация моделей
   - Weighted Voting, Boosting, Bagging, Stacking, Random Forest, Яндекс.MatrixNet

Домингос П. Верховный алгоритм. 2016. 336 с.

Полезные ссылки

- www.kaggle.com — конкурсы анализа данных
- www.kdnuggets.com — главный сайт датамайнеров
- www.MachineLearning.ru — русскоязычная вики
- www.datasciencecentral.com — 72 000 датамайнеров
- archive.ics.uci.edu/ml — UCI ML Repository (349 datasets)
- ru.coursera.org/learn/machine-learning — курс Эндрю Ына
- ru.coursera.org/learn/vvedenie-mashinnoe-obuchenie — курс Воронцова от ВШЭ и ШАД Яндекс
- ru.coursera.org/specializations/machine-learning-data-analysis — специализация от МФТИ и ШАД Яндекс

Рекомендуемая литература

- *Домингос П.* Верховный алгоритм. 2016. 336 с.
- *Коэльо Л. П., Ричарт В.* Построение систем машинного обучения на языке Python. 2016. 302 с.
- Машинное обучение (курс лекций, К. В. Воронцов). www.MachineLearning.ru. 2004–2017.
- *Мерков А. Б.* Распознавание образов. Введение в методы статистического обучения. 2011. 256 с.
- *Мерков А. Б.* Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
- *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning. Springer, 2014. 739 p.
- *Bishop C. M.* Pattern recognition and machine learning. Springer, 2006. 738 p.

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov