История машинного обучения

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

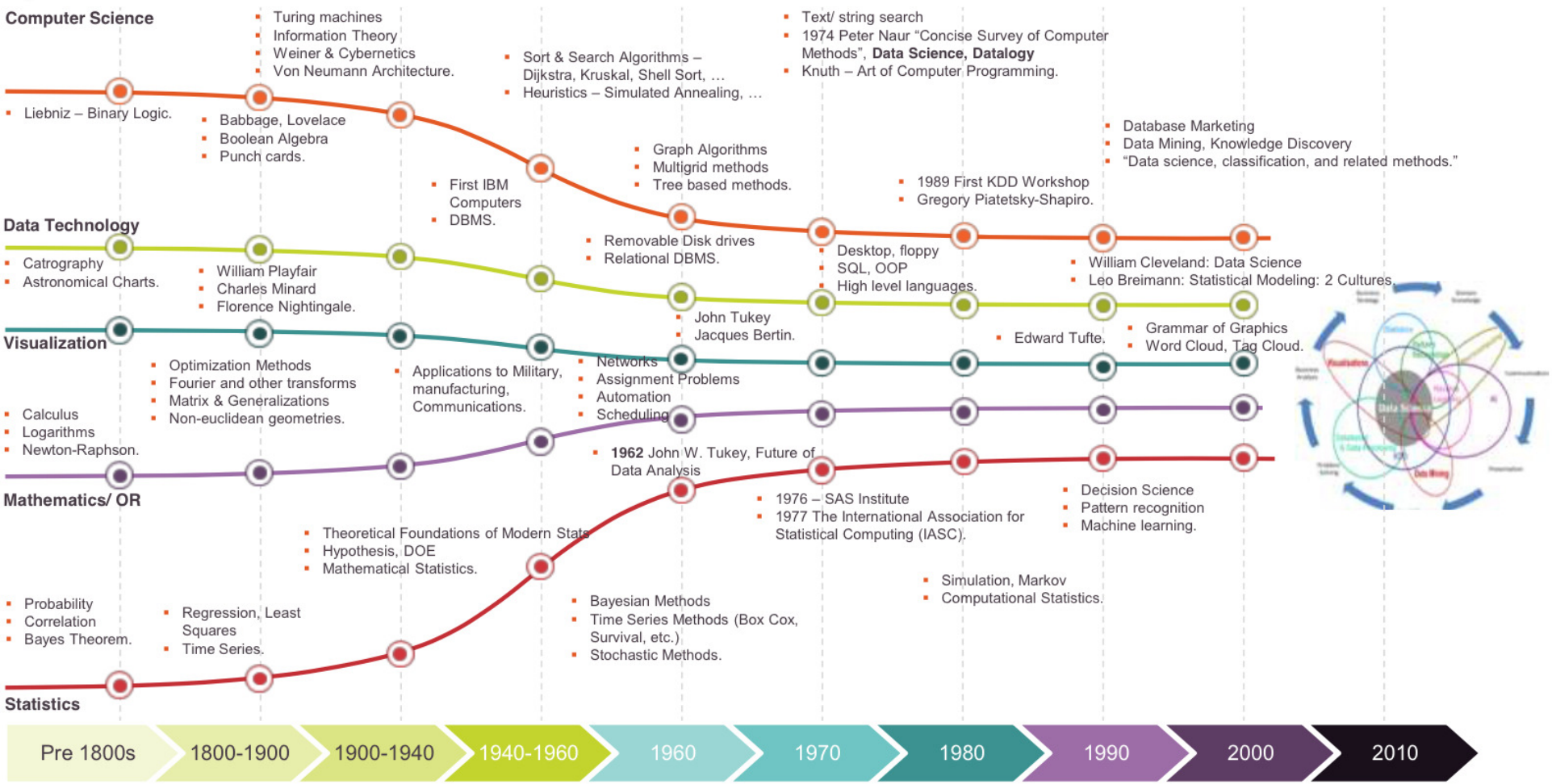
МФТИ • 19 мая 2016

Содержание

Разные названия то ли одной большой науки, то ли её частей

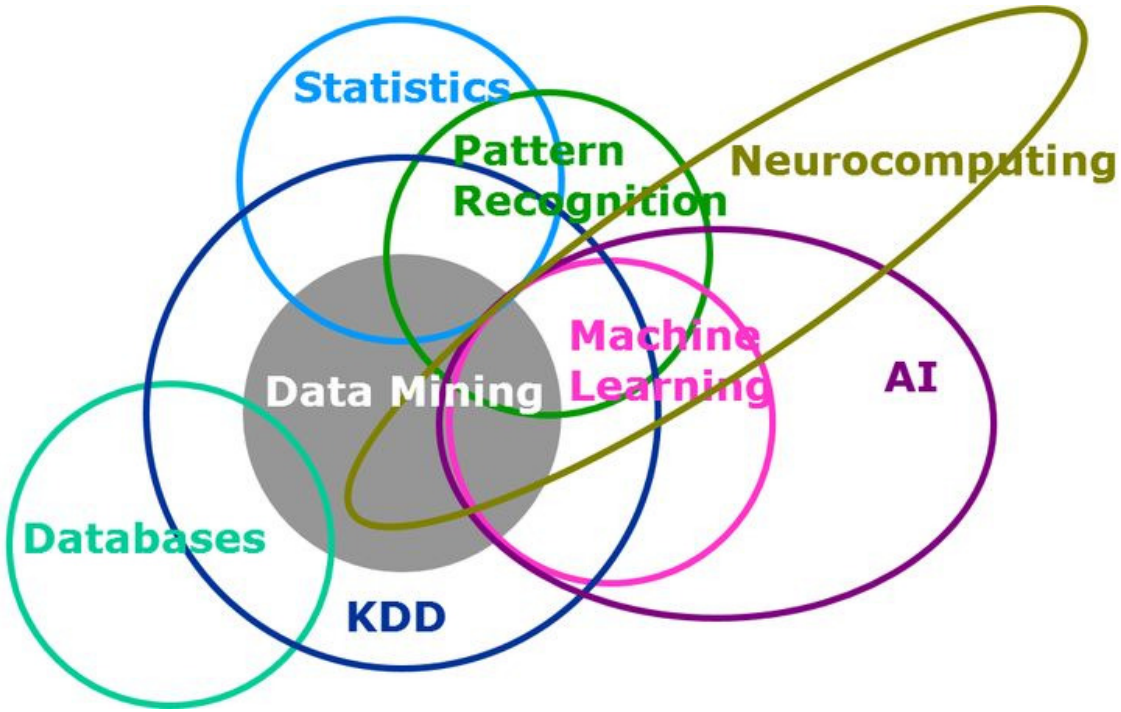
- Статистический анализ данных (Statistical Data Analysis)
- Искусственный интеллект (Artificial Intelligence) — 1955
- Распознавание образов (Pattern Recognition)
- Машинное обучение (Machine Learning) — 1959
- Статистическое обучение (Statistical Learning)
- Интеллектуальный анализ данных (Data Mining) — 1989
- Knowledge Discovery in Databases — 1989
- Науки о данных (Data Science) — 1997
- Бизнес-аналитика (Business Intelligence, Business Analytics)
- Предсказательная аналитика (Predictive Analytics) — 2007
- Большие данные (Big Data) — 2008
- Аналитика больших данных (Big Data Analytics)

Предпосылки Data Science



http://www.kdnuggets.com/2015/02/history-data-science-infographic.html

Место машинного обучения среди смежных областей



Кое-что спорно на этой диаграмме.

Более серьёзный анализ здесь:
http://insideanalysis.com/2012/04/data-mining-and-beyond

Принцип эмпирической индукции

«Не следует полагаться на сформулированные аксиомы и формальные базовые понятия, какими бы привлекательными и справедливыми они не казались. Законы природы нужно «расшифровывать» из фактов опыта. Следует искать правильный метод анализа и обобщения опытных данных; здесь логика Аристотеля не подходит в силу её абстрактности, оторванности от реальных процессов и явлений.»

Фрэнсис Бэкон (1561–1626)

Таблицы открытия: множество случаев x, когда

- свойство y *присутствовало* $y(x) = 1$
- свойство y *отсутствовало* $y(x) = 0$
- наблюдалось изменение *степени* свойства $y(x)$

Фрэнсис Бэкон. Новый органон. 1620.

Восстановление зависимостей по эмпирическим данным

Дано:
объекты $x_i = \big(f_1(x_i), \ldots, f_n(x_i)\big)$ и ответы $y_i = y(x), \;\; i = 1, \ldots, \ell$
$f_j(x)$ — признаки объекта x, $j = 1, \ldots, n$

Найти:
функцию $a(x, w)$, восстанавливающую зависимость $y(x)$

Критерий: минимум эмпирического риска

$$\sum_{i=1}^{\ell} \mathscr{L}(a(x_i, w), y_i) \to \min_w,$$

где $\mathscr{L}(a, y)$ — функция потерь от ошибки a при ответе y.

Основные типы задач обучения с учителем:

- регрессия: $y_i \in \mathbb{R}, \;\; \mathscr{L}(a, y) = (a - y)^2$
- классификация: $y_i \in \{0, 1\}, \;\; \mathscr{L}(a, y) = [a \neq y]$

Метод наименьших квадратов (Гаусс, 1795)

Линейная модель регрессии:

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

Функционал квадрата ошибки:

$$\sum_{i=1}^{\ell} \big(a(x_i, w) - y_i\big)^2 = \|Fw - y\|^2 \to \min_w.$$

Карл Фридрих Гаусс (1777–1855)

Решение системы: $w^* = (F^\intercal F)^{-1} F^\intercal y$.

Матричные обозначения:

$$\underset{\ell \times n}{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell \times 1}{y} = \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}, \quad \underset{n \times 1}{w} = \begin{pmatrix} w_1 \\ \ldots \\ w_n \end{pmatrix}.$$

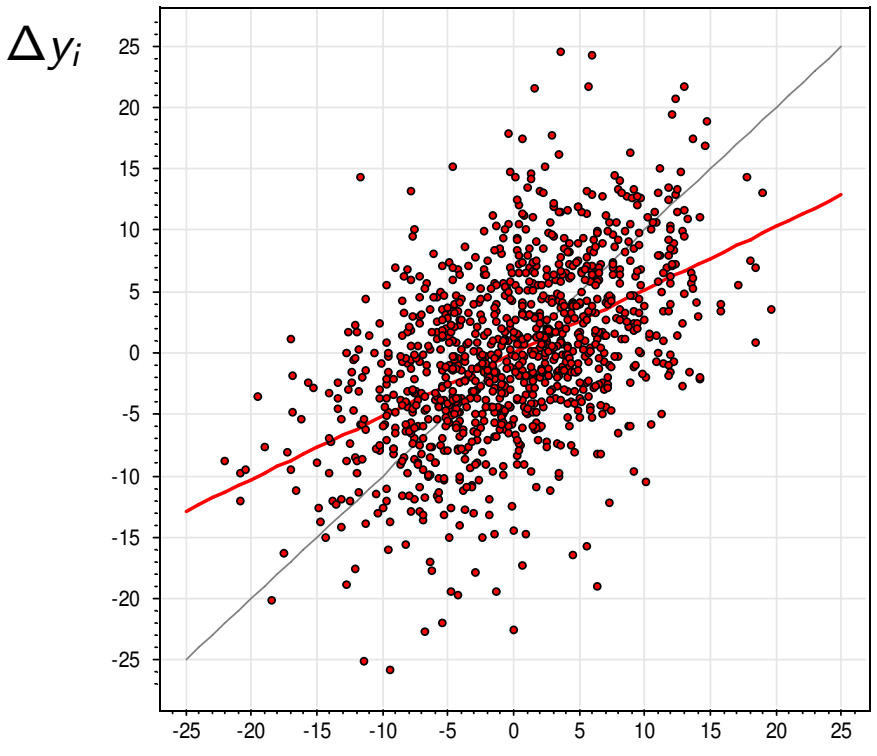
Откуда пошло название «регрессия» (Гальтон, 1886)

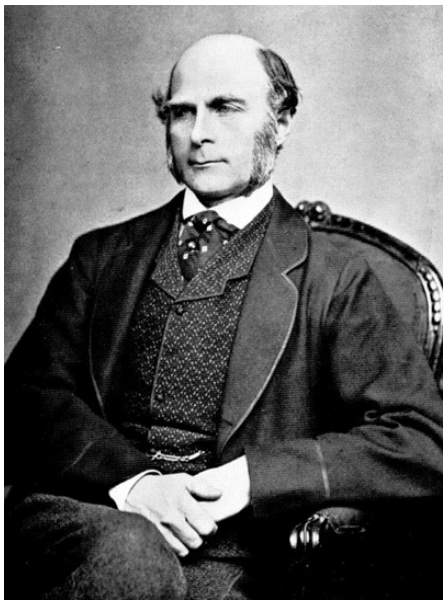
Исследование наследственности роста.

отклонение роста от среднего в популяции:

Δx_i — отклонение роста отца

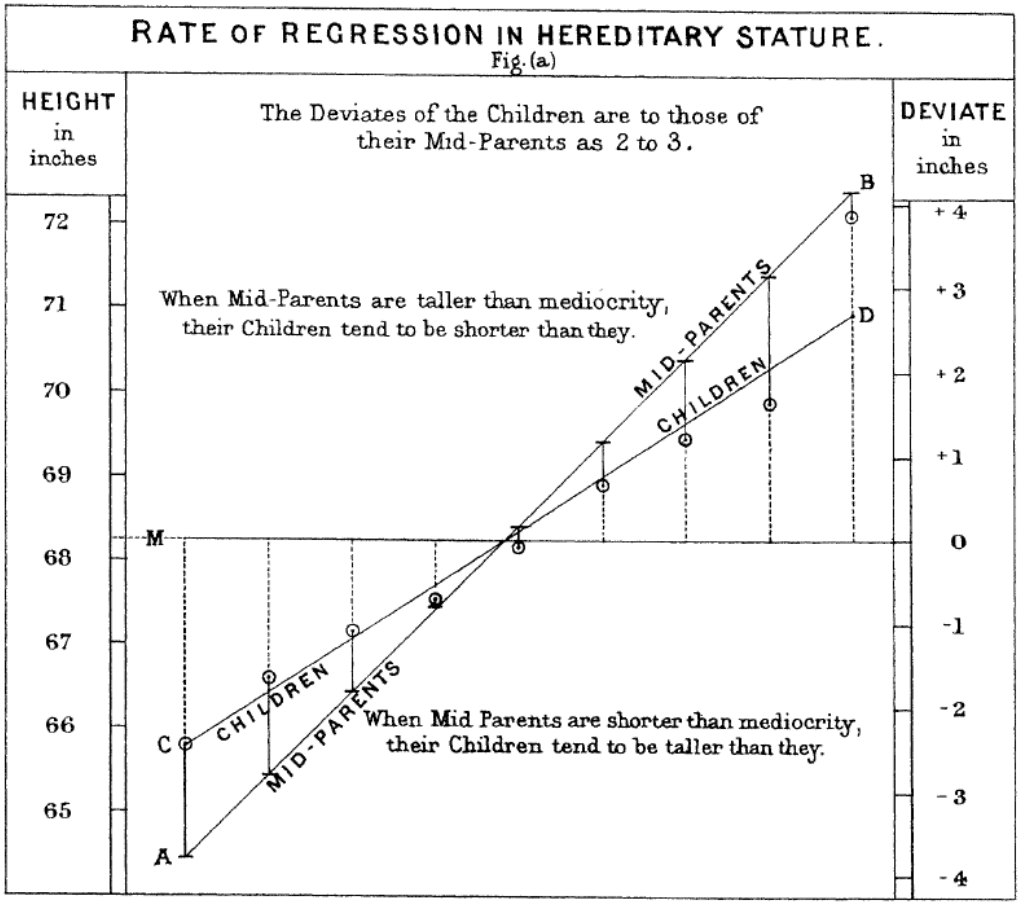
Δy_i — отклонение роста взрослого сына

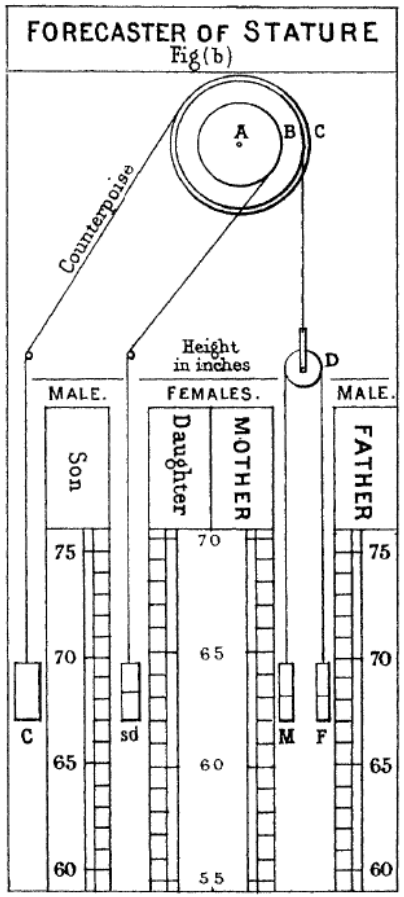




Фрэнсис Гальтон
(1822–1911)

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





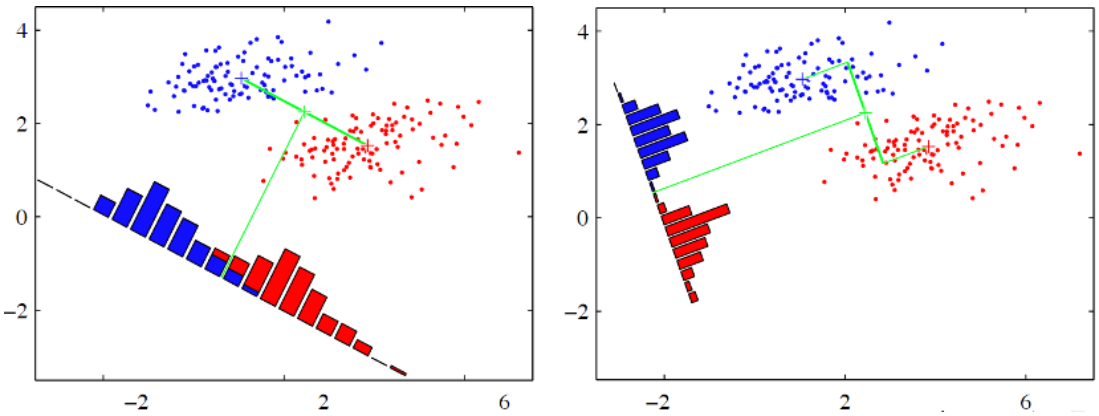
Galton F. Regression Towards Mediocrity in Hereditary Stature. 1886.

Линейный дискриминантный анализ (Р.Фишер, 1936)

Линейная модель классификации:

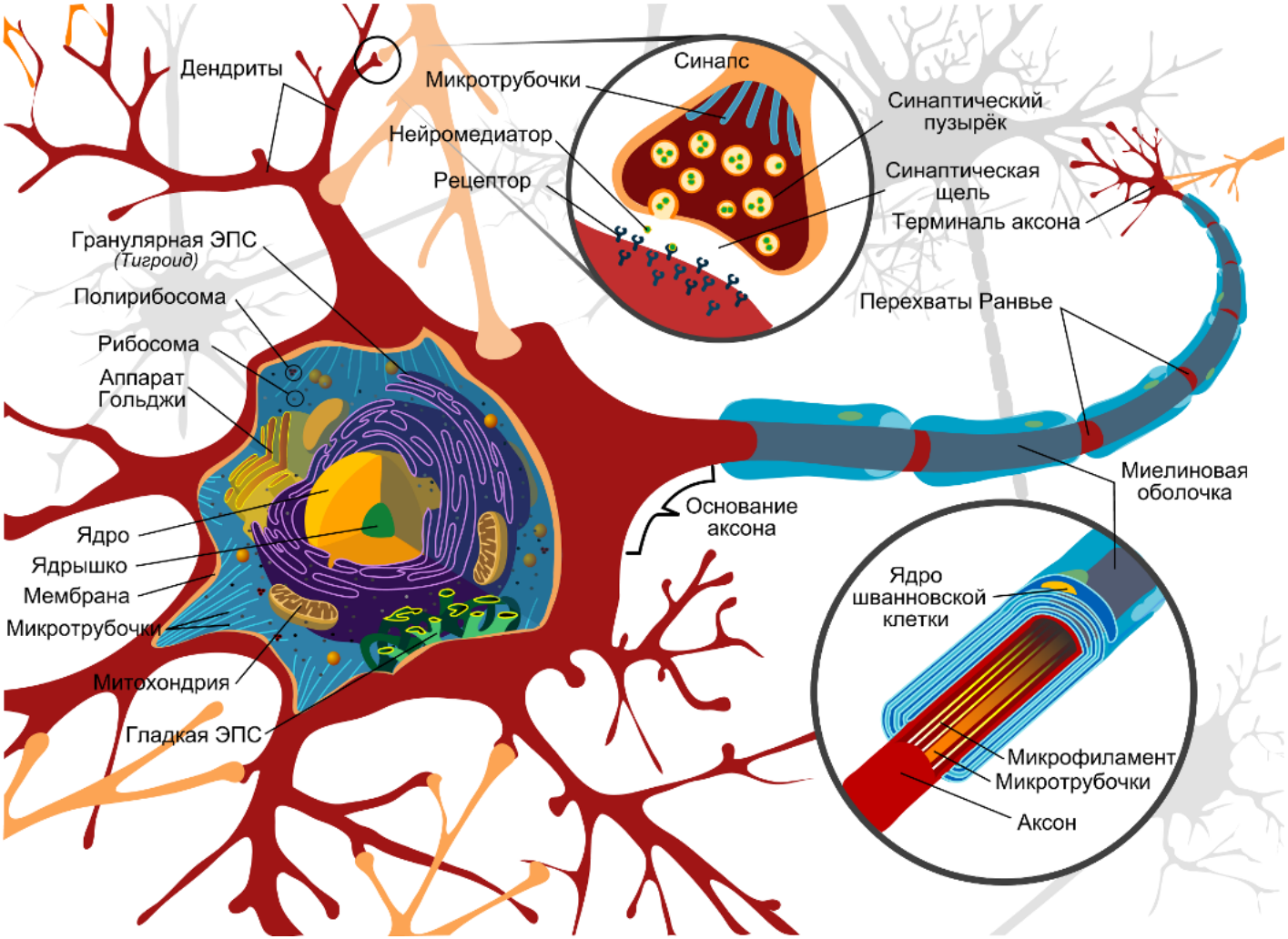
$$a(x, w) = \text{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0 \right)$$

В проекции на направляющий вектор w разделяющей гиперплоскости вероятность ошибки минимальна:



Рональд Фишер (1890–1962)

Нервная клетка — естественный нейрон

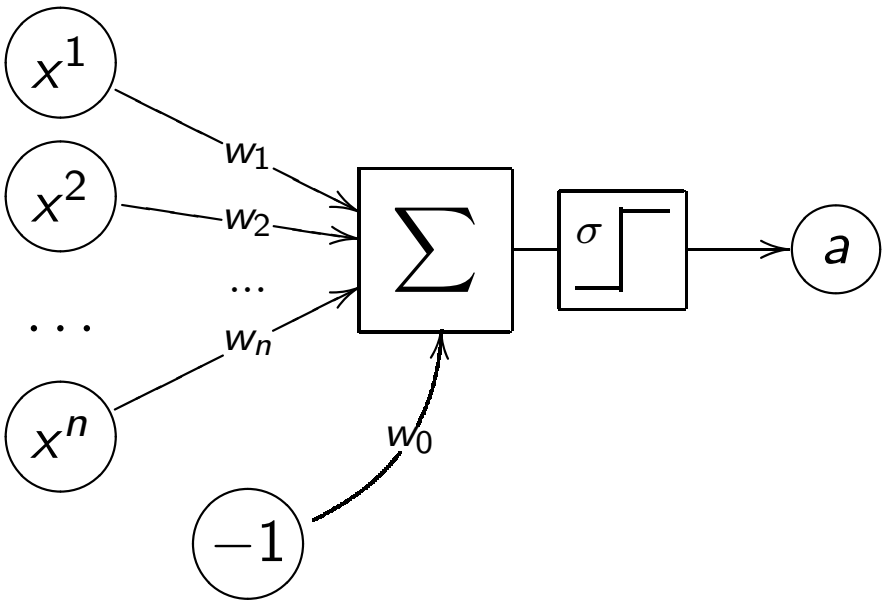


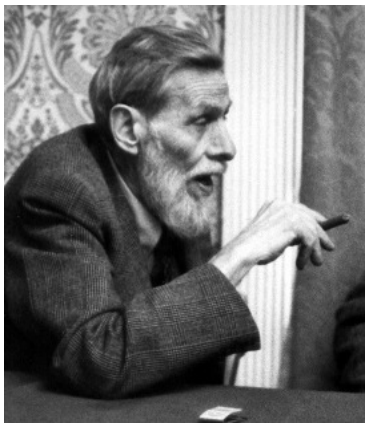
Модель МакКаллока–Питтса — искусственный нейрон

Линейная модель нейрона (1943):

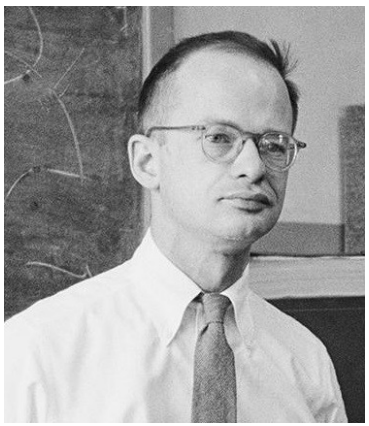
$$a(x, w) = \sigma\left(\sum_{j=1}^{n} w_j x^j - w_0\right),$$

где $\sigma(z)$ — функция активации,
например, $\text{sign}(z)$ или $\text{arcth}(z)$





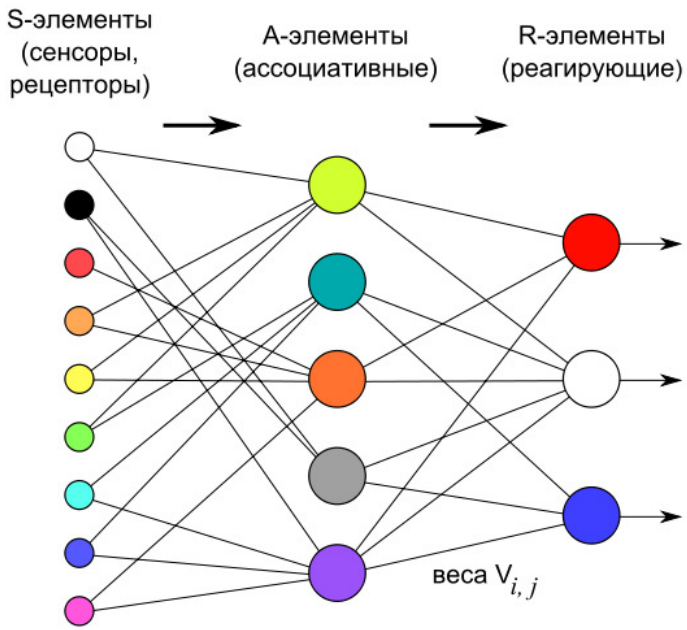
Уоррен МакКаллок
(1898–1969)



Вальтер Питтс
(1923–1969)

Персептрон Розенблатта (1957) и теорема Новикова (1960)

Mark-1 — первый нейрокомпьютер (1960)
Обучение — метод коррекции ошибки
Архитектура — двухслойная сеть



Фрэнк Розенблатт
(1928–1971)

Розенблатт Ф. Принципы нейродинамики. Перцептроны и теория механизмов мозга. 1965 (1962).

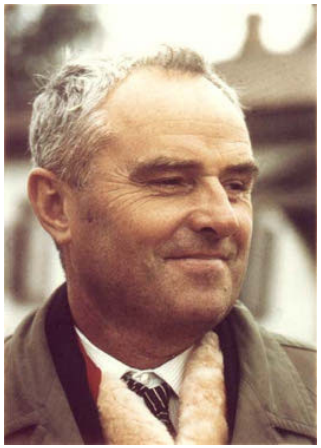
Novikoff A. B. J. On convergence proofs on perceptrons. 1962.

Основные вехи развития теории нейронных сетей

- Развитие нейронных сетей затормозилось на десятилетие
  Минский М., Пайперт С. Персепроны. 1971 (1969).
- BackProp — метод обратного распространения ошибок
  Галушкин А. И. Синтез многослойных систем распознавания образов. 1974.
  Werbos P. J. Beyond regression: new tools for prediction and analysis in the behavioral sciences. 1974.
  LeCun Y. Une procedure d'aprentissage pour reseau a seuil assymetrique. 1985.
  Parker D. B. Learning-logic: Casting the cortex of the human brain in silicon. 1985.
  Rummelhart D., Hinton G., Williams R. Learning Internal Representations by Error Propagation. 1986.
- Сети радиальных базисных функций
  Bashkirov O. A., Braverman E. M., Muchnik I. B. Potential function algorithms for pattern recognition learning machines. 1964.
  Broomhead D. S., Lowe. D. Multivariable functional interpolation and adaptive networks complex systems. 1988.
- Самоорганизующиеся сети Кохонена
  Kohonen T. Self-organized formation of topologically correct feature maps. 1982.
- Рекуррентные сети Хопфилда
  Hopfield. Neural networks and physical systems with emergent collective computational abilities. 1982.
- Свёрточные сети
  LeCun, Bottou, Bengio, Haffner. Gradient-based learning applied to document recognition. 1998.
- Глубокие сети (сети с большим числом слоёв)
  Ivakhnenko A.G., Lapa V.G. Cybernetic Predicting Devices. 1965.
  Rina Dechter. Learning while searching in constraint-satisfaction problems. 1986.
  Hinton G.E. Learning multiple layers of representation. 2007.

Метод группового учёта аргументов, МГУА (GMDH)

1. *Самоорганизация моделей* — подбор оптимальной структуры модели из огромного числа вариантов
2. Качество моделей оценивается в процессе перебора по совокупности разнообразных *внешних критериев*
3. Сотни применений, около 300 диссертаций в 70-80-е гг.



Алексей Григорьевич Ива́хненко (1913–2007)

Ивахненко А. Г., Лапа В. Г. Кибернетические предсказывающие устройства. 1965.
Ивахненко А. Г., Зайченко Ю. П., Димитров В. Д. Принятие решений на основе самоорганизации. 1976.
Ивахненко А. Г. Индуктивный метод самоорганизации моделей сложных систем. 1982.

Алгоритмические композиции

- Простое и взвешенное голосование
  Мазуров В. Д. Комитеты системы неравенств и задача распознавания. 1971.
  Журавлёв Ю. И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. 1977.
  Freund Y., Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting. 1995.
  Friedman G. Greedy Function Approximation: A Gradient Boosting Machine. 1999.
- Случайный лес
  Breiman L. Random Forests. 2001.
- Восстановление смесей распределений
  Шлезингер М. И. О самопроизвольном различении образов. 1965.
  Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM-algorithm. 1977.
- Смеси классификаторов с областями компетентности
  Растригин Л. А., Эренштейн Р. Х. Коллективные правила распознавания. 1981.
  Jacobs R. A., Jordan M. I., Nowlan S. J., Hinton G. E. Adaptive mixtures of local experts. 1991.

Градиентный бустинг и *случайный лес* — универсальные и наиболее успешные методы классификации.

Яндекс.MatrixNet — параллельная распределённая реализация Gradient Boosting над ODT (Oblivious Decision Tree).

Научная школа М. М. Бонгарда

1. 1958: Программа «Открой закон» восстанавливала зависимость полным перебором формул
2. 1959: Программа «Арифметика» для сокращения перебора использовала оценки информативности
3. 1961: Программа «КоРа» перебирала информативные тройки признаков

Михаил Моисеевич Бонгард (1924–1971)

«КоРа-3»: первое применение распознавания незрительных образов для распознавания в скважине границы нефть-вода.

Впервые применено *голосование* и *скользящий контроль*.

Бонгард, Вайнцвайг, Губерман, Извекова, Смирнов. Использование обучающейся программы для выявления нефтеносных пластов. 1966.

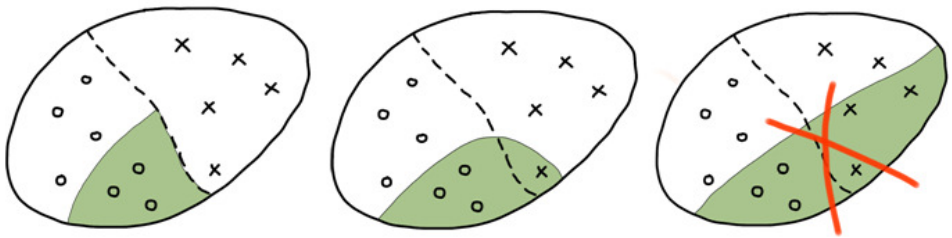
Понятие закономерности

Воплощение принципа эмпирической индукции Ф.Бэкона:

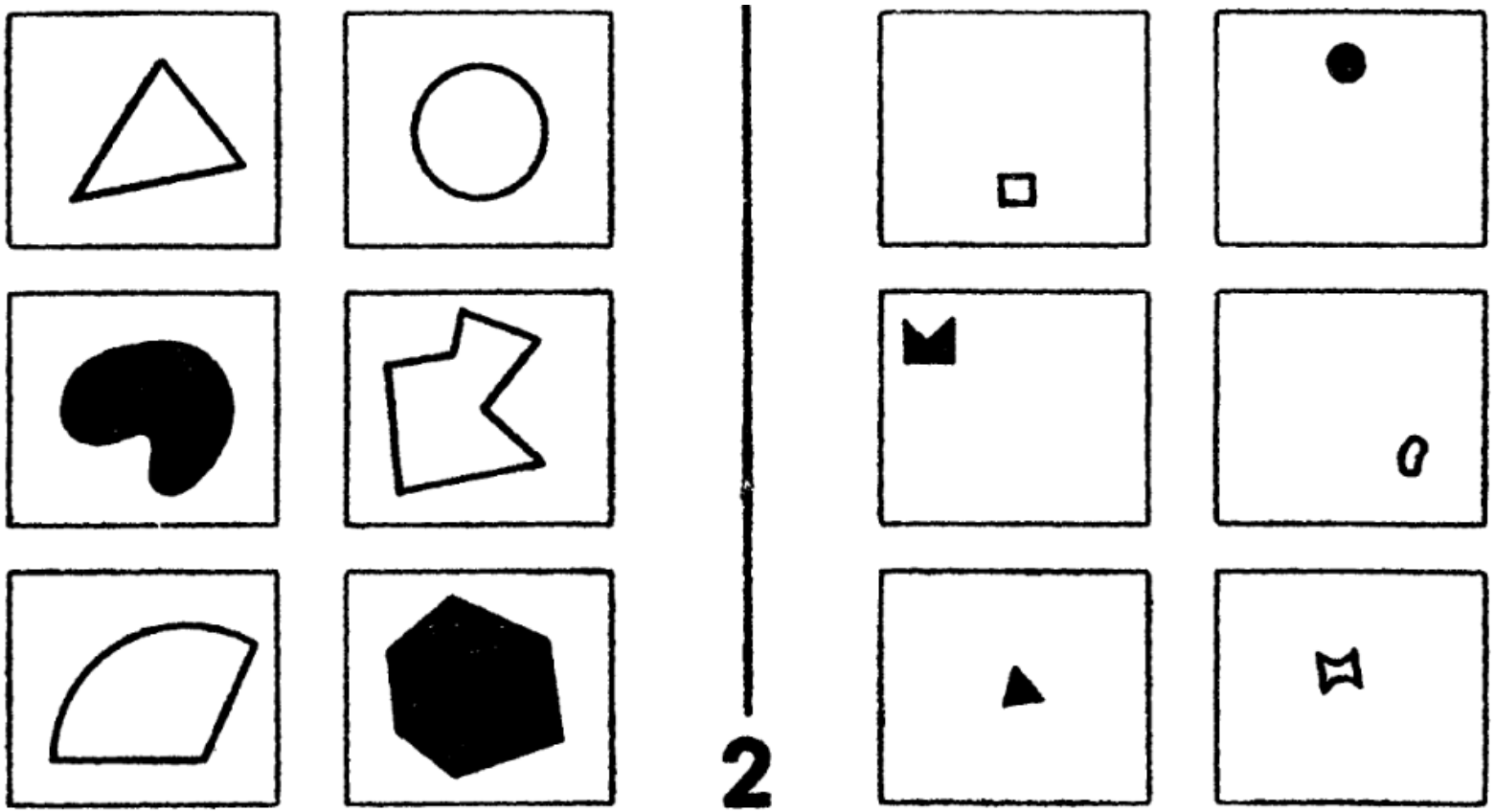
Логическая закономерность (правило, rule) — предикат $R(x)$, удовлетворяющий двум требованиям:

1. *информативность* относительно класса $y \in Y$:
   $p(R) = \#\{x_i : R(x_i)=1 \text{ и } y_i=y\} \to \max$
   $n(R) = \#\{x_i : R(x_i)=1 \text{ и } y_i \neq y\} \to \min$
2. *интерпретируемость*:
   1) R записывается на естественном языке
   2) R зависит от небольшого числа признаков (1–7)

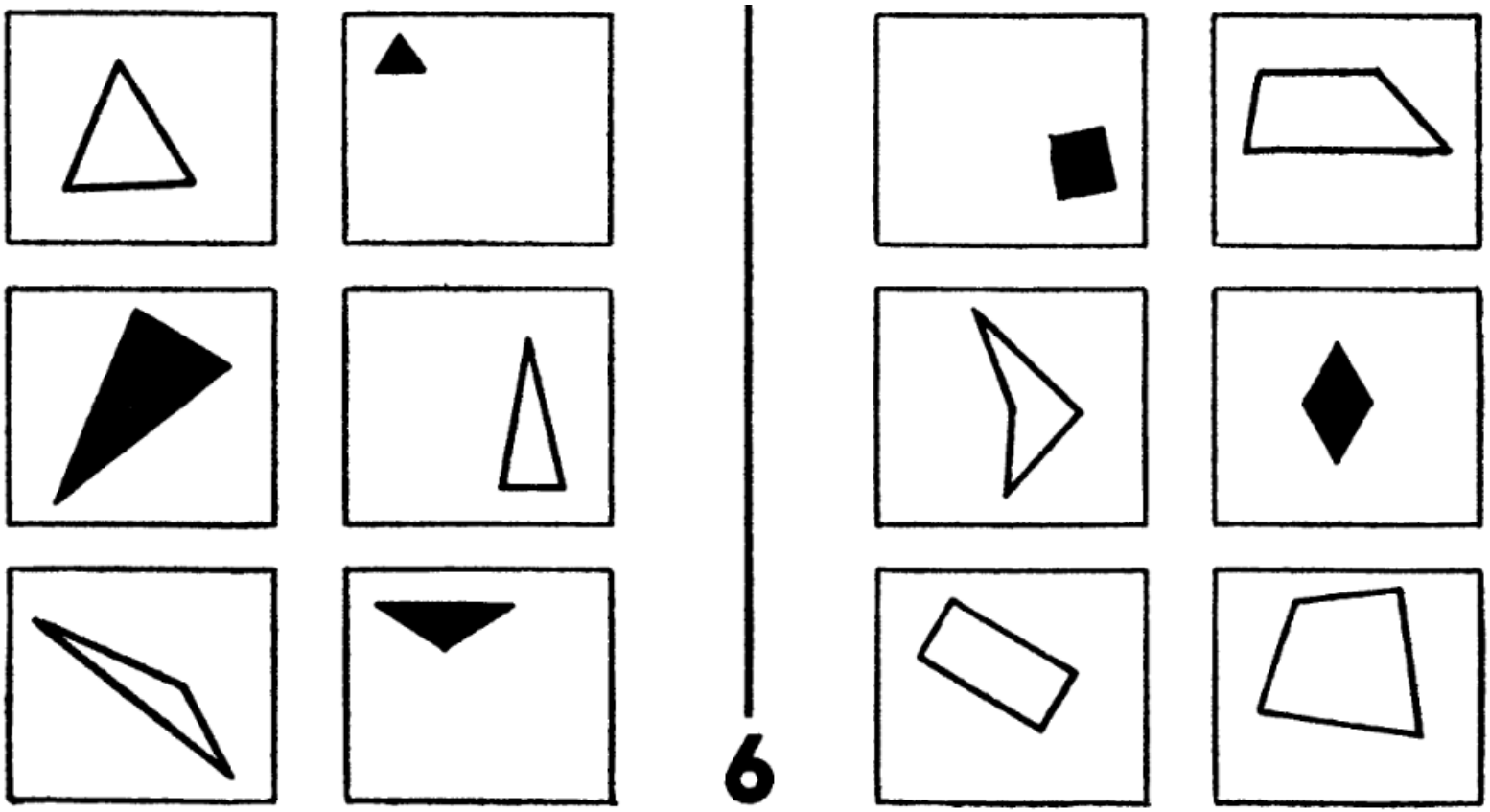
Если $R(x) = 1$, то говорят «R выделяет x» (R covers x).



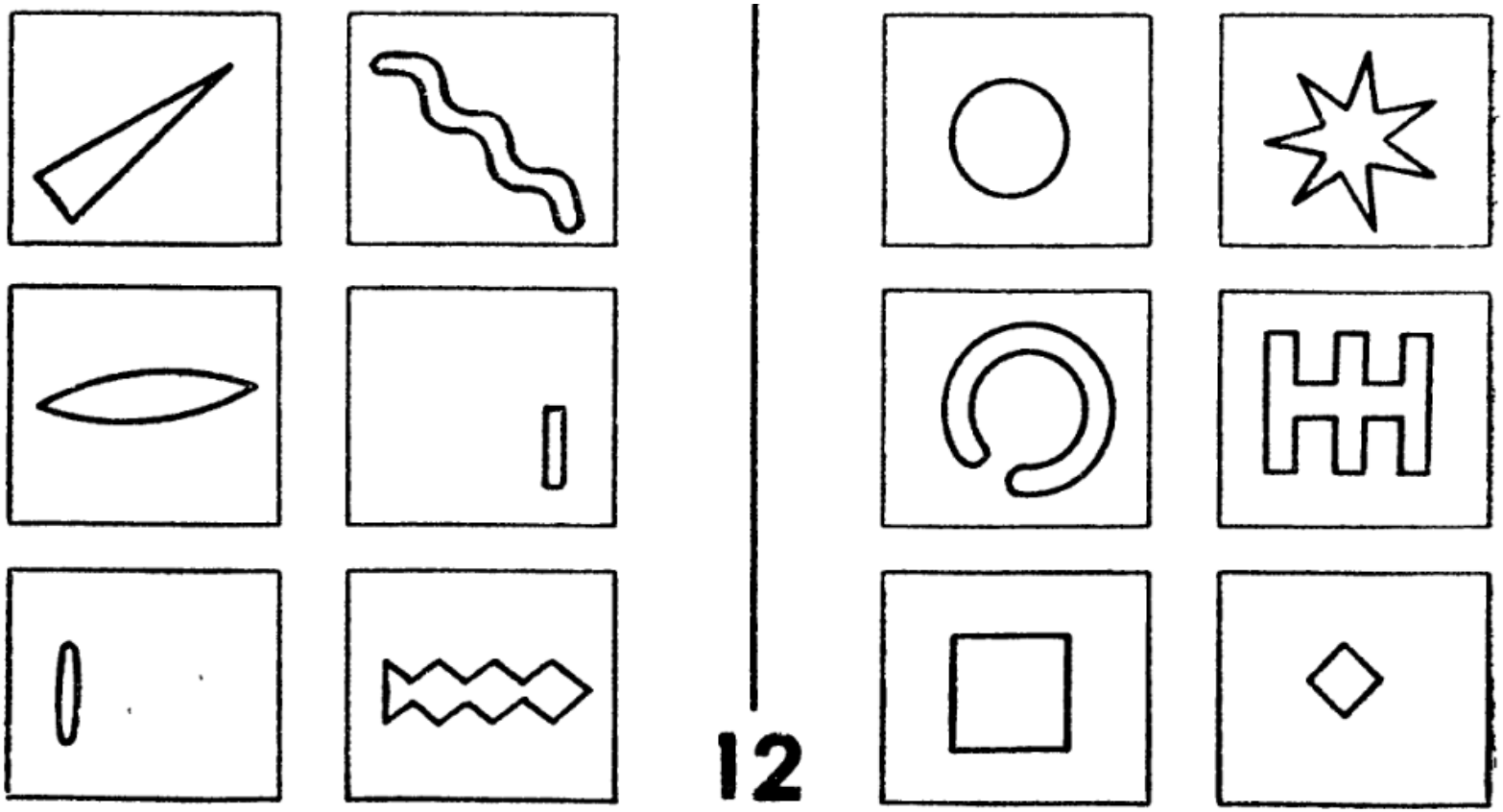
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



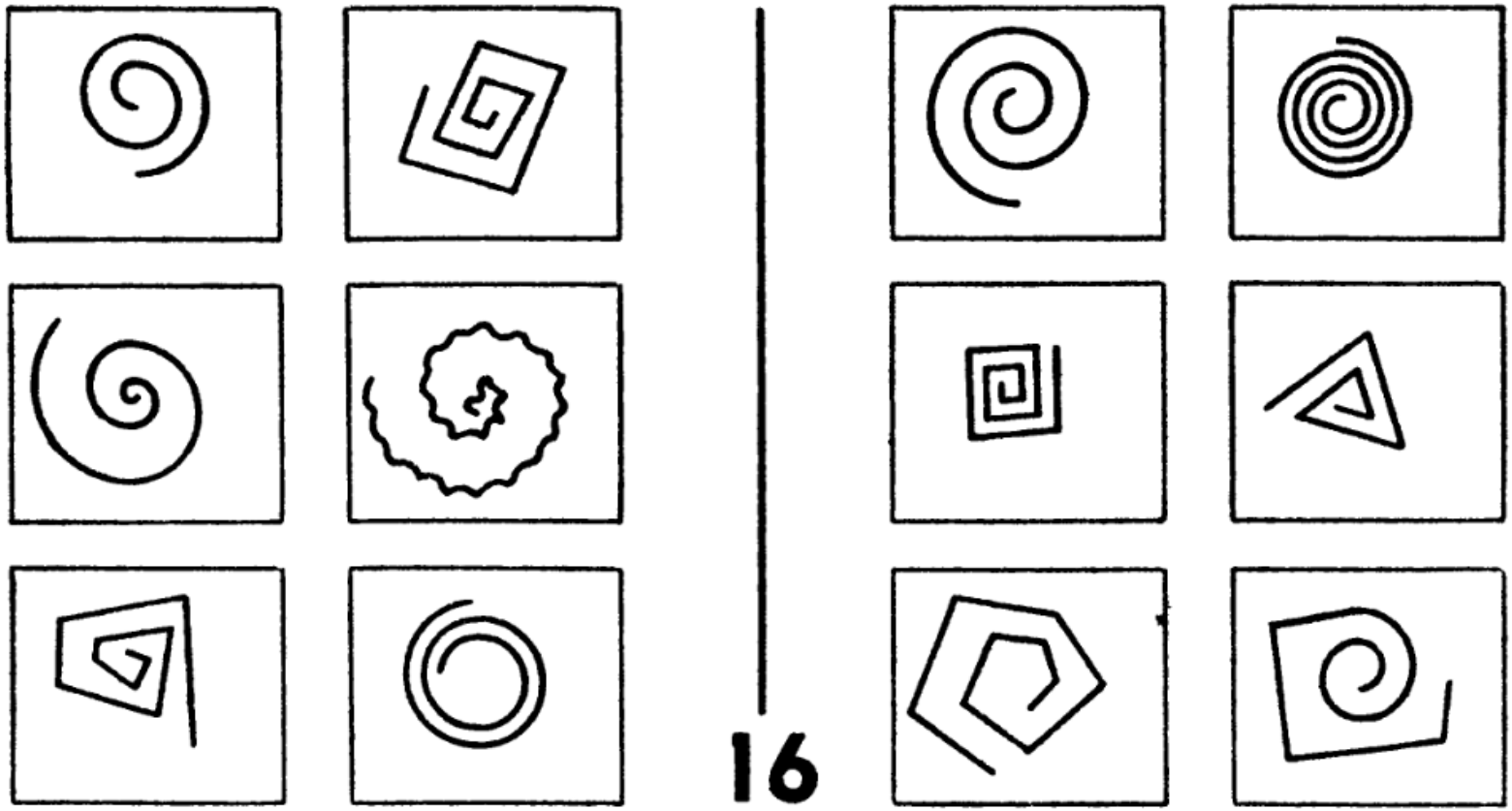
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



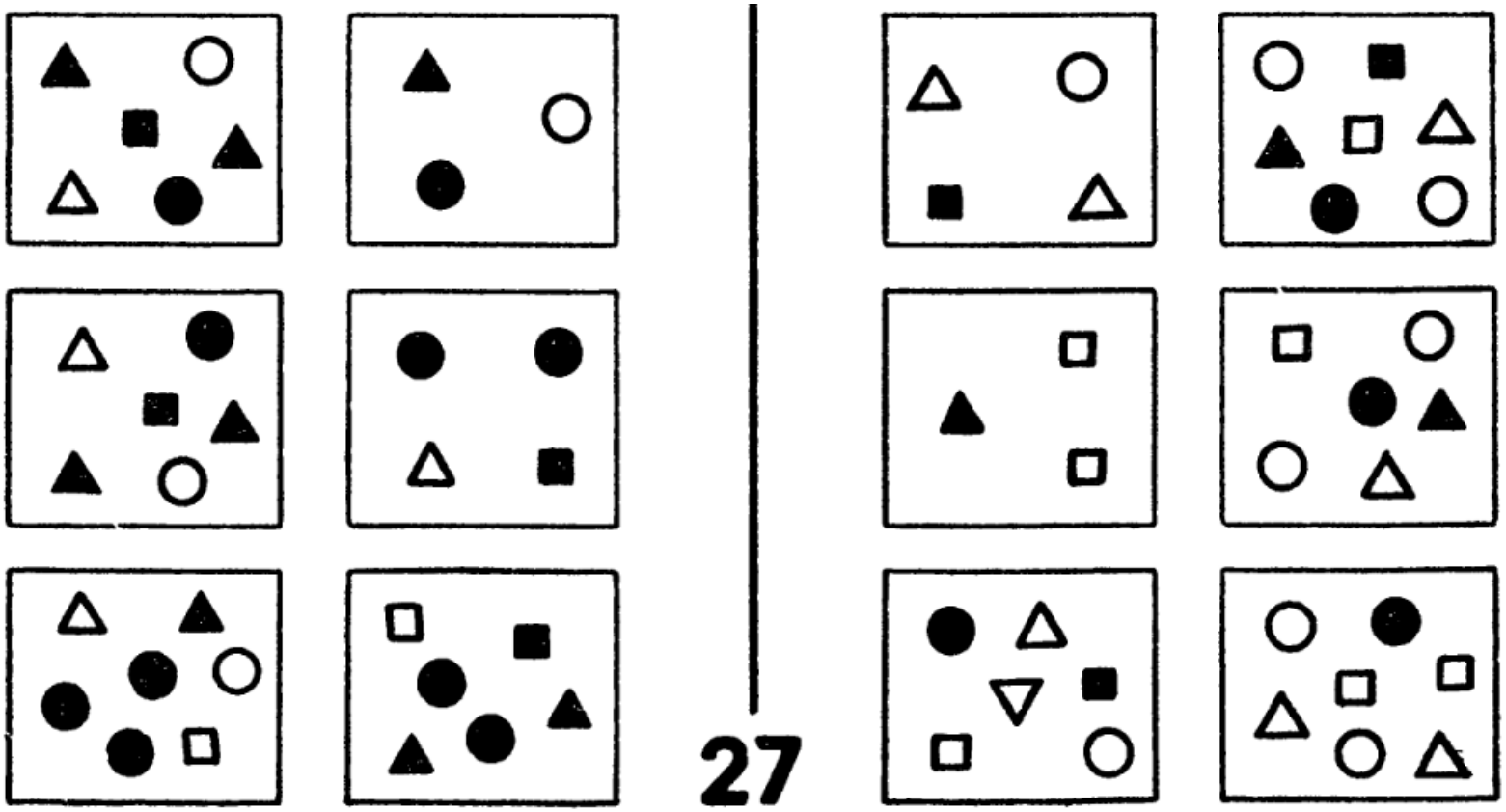
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



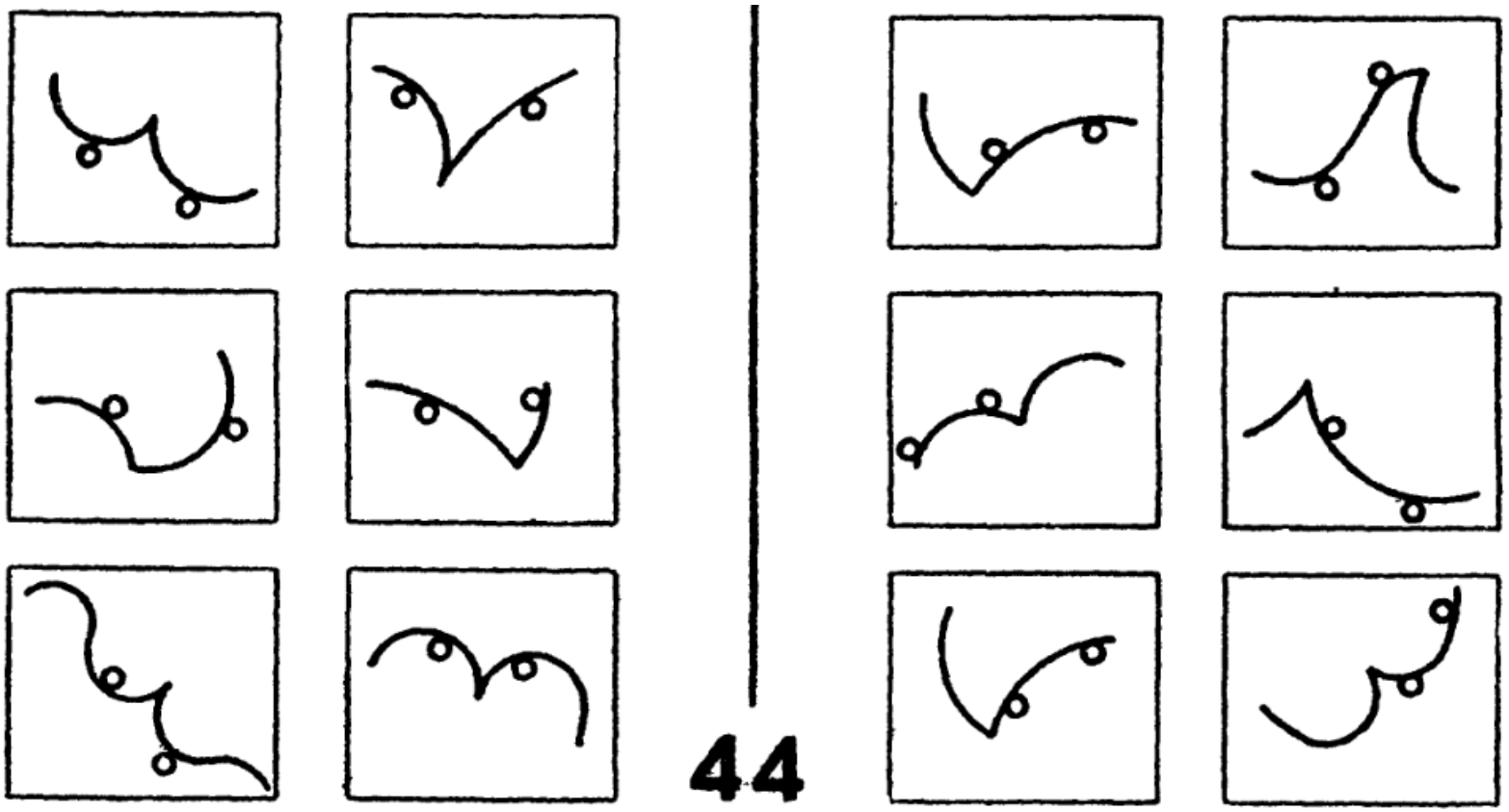
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



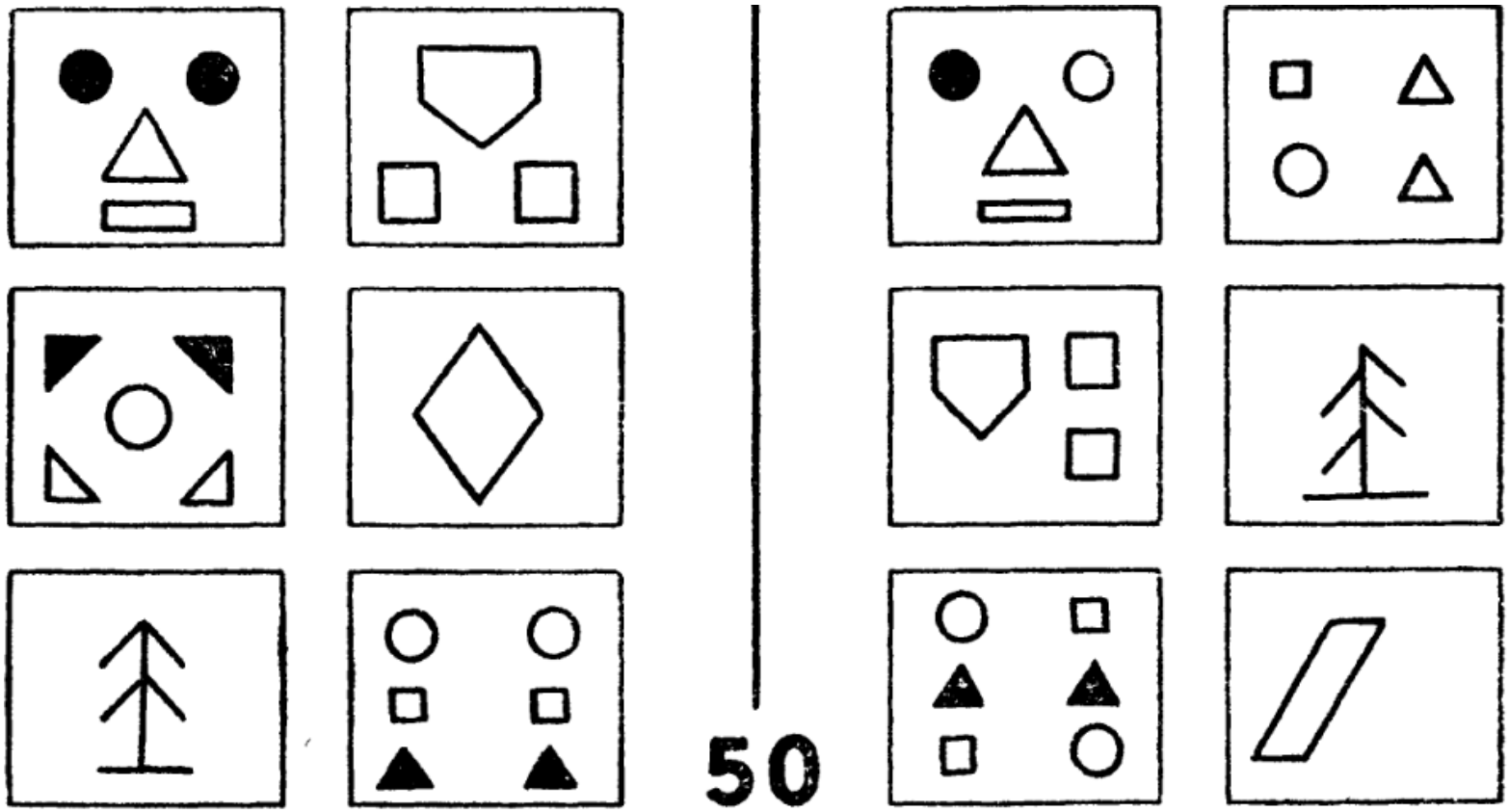
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



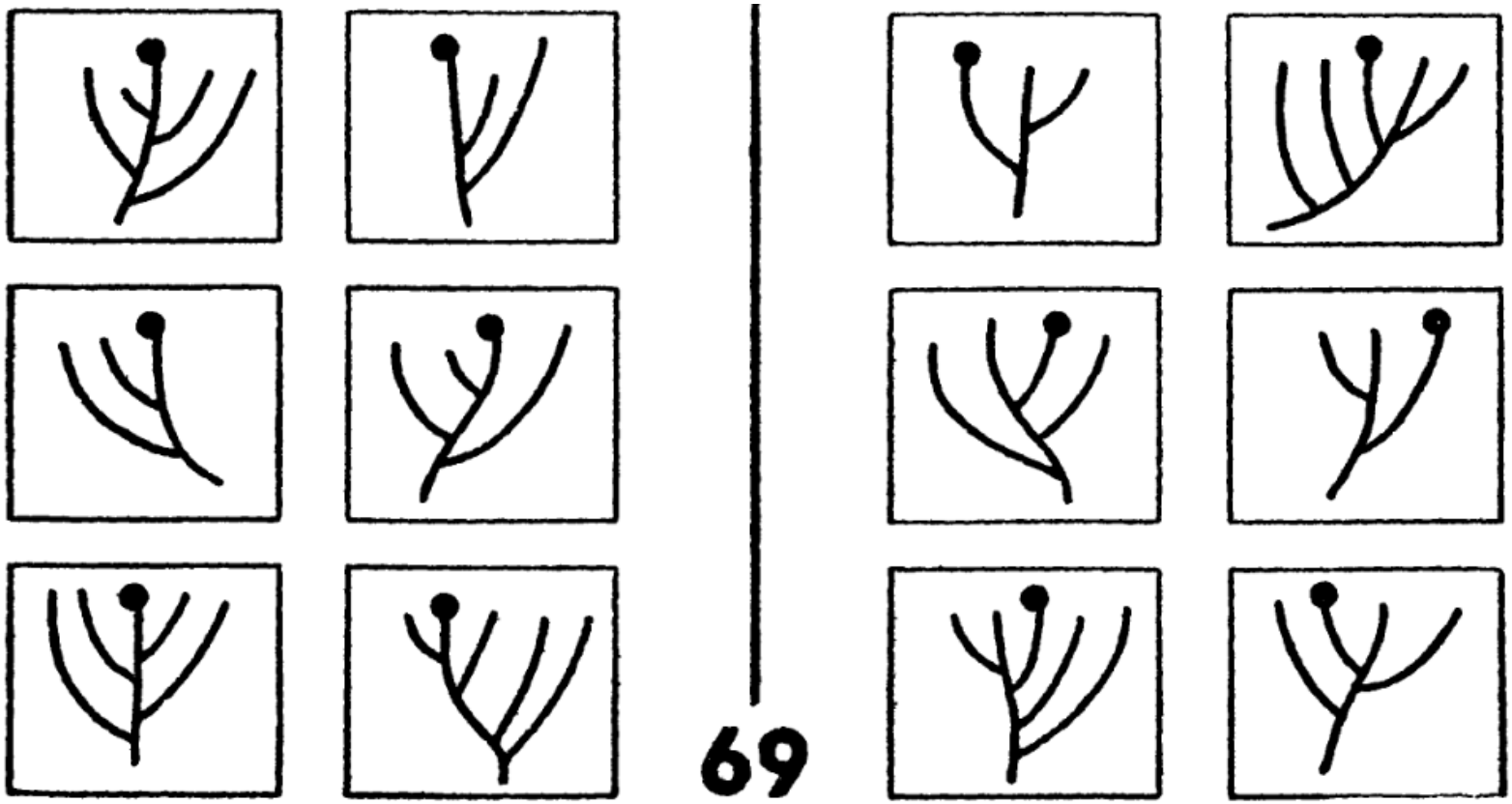
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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



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



Научная школа М. А. Айзермана

- *Гипотеза компактности*: близкие объекты, как правило, находятся в одном классе
- Идея *потенциальных функций* заимствована из физики
- Линейные модели в пространстве близостей $f_i(x) = K(x, x_i)$ объекта x до обучающих объектов x_i

Марк Аронович Айзерман (1913–1992)

М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. 1964.

М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр. Метод потенциальных функций в теории обучения машин. 1970.

А. Г. Аркадьев, Э. М. Браверман. Обучение машин распознаванию образов. 1964.

Определение обучаемости в статистической теории

Семейство классификаторов A обучаемо:

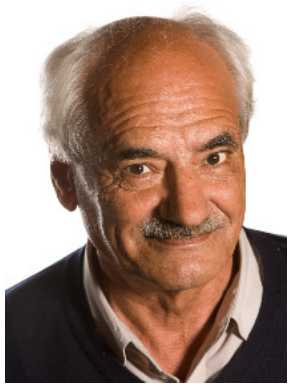
$$\mathsf{P}\left\{\sup_{a\in A}\left|P(a)-\nu(a,X^\ell)\right|>\varepsilon\right\}\leqslant\eta,$$

$P(a)$ — вероятность ошибки классификатора,
$\nu(a,X^\ell)$ — эмпирический риск (частота ошибки классификатора a на выборке).

Основные результаты VC-теории:

- Обосновано ограничение сложности A
- Введена мера сложности VCdim
- Метод структурной минимизации риска

Владимир Наумович Вапник



Алексей Яковлевич Червоненкис (1938–2014)

Вапник В. Н., Червоненкис А. Я.
Теория распознавания образов. М.: Наука, 1974.

Метод опорных векторов (Support Vector Machine, SVM)

Метод обобщённого портрета (1963) $\to$ SVM (1992)

- Линейный классификатор с зазором максимальной ширины
- Аппроксимация и регуляризация эмпирического риска

$$\sum_{i=1}^{\ell}\big(1-\langle x_i, w\rangle y_i\big)_+ + \frac{1}{2C}\|w\|^2 \;\to\; \min_w$$

- Изящный переход в спрямляющее пространство

$$\big\langle x, x'\big\rangle \to K(x,x') = \big\langle \psi(x), \psi(x')\big\rangle$$

- В результате — двухслойная нейронная сеть
  с высокой обобщающей способностью
  и автоматическим выбором числа нейронов скрытого слоя

Задачи, некорректно поставленные по Адамару

Корректно поставленная задача:

- решение существует,
- решение единственно,
- решение устойчиво (непрерывно зависит от данных в некоторой разумной топологии).

Задачи восстановления зависимостей по эмпирическим данным — всегда некорректно поставленные.

Жак Саломон Адамар (1865–1963)

Регуляризация — это введение дополнительных ограничений.

Hadamard J. Sur les problèmes aux dérivées partielles et leur signification physique. 1902.

Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. 1974.

Регуляризация линейных моделей

Регуляризатор — добавка к внутреннему критерию $Q(a, X^\ell)$, штраф за сложность (complexity penalty) модели $a \in A$:

$$Q_{\text{рег}}(a, X^\ell) = Q(a, X^\ell) + \text{штраф}(A) \;\to\; \min_{a \in A}$$

Линейные модели: $A = \{a(x) = \text{sign}\langle w, x\rangle\}$ — классификация,
$A = \{a(x) = \langle w, x\rangle\}$ — регрессия.

L_2-регуляризация (ридж-регрессия): $\text{штраф}(w) = \tau \sum_{j=1}^{n} w_j^2$

L_1-регуляризация (LASSO): $\text{штраф}(w) = \tau \sum_{j=1}^{n} |w_j|$

L_0-регуляризация (AIC, BIC): $\text{штраф}(w) = \tau \sum_{j=1}^{n} \left[w_j \neq 0\right]$

Регуляризаторы для отбора признаков

- Критерий Акаике

  Akaike H. Information theory and an extension of the maximum likelihood principle. 1973.

- Байесовский информационный критерий

  Schwarz G. E. Estimating the dimension of a model. 1978

- Структурная минимизация риска

  Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. 1974.

- LASSO (least absolute shrinkage and selection operator)

  Tibshirani R. Regression Shrinkage and Selection via the lasso. 1996

- LARS (least angle regression)

  Efron B., Hastie T., Johnstone I., Tibshirani R. Least Angle Regression. 2004

- ElasticNet (сумма L_0 и L_1 регуляризаторов)

  Hui Zou, Hastie T. Regularization and Variable Selection via the Elastic Net. 2005

- Негладкие регуряризаторы для отбора признаков

  Tatarchuk A., Mottl V., Eliseyev A., Windridge D. Selectivity supervision in combining pattern recognition modalities by feature- and kernel-selective Support Vector Machines. 2008.

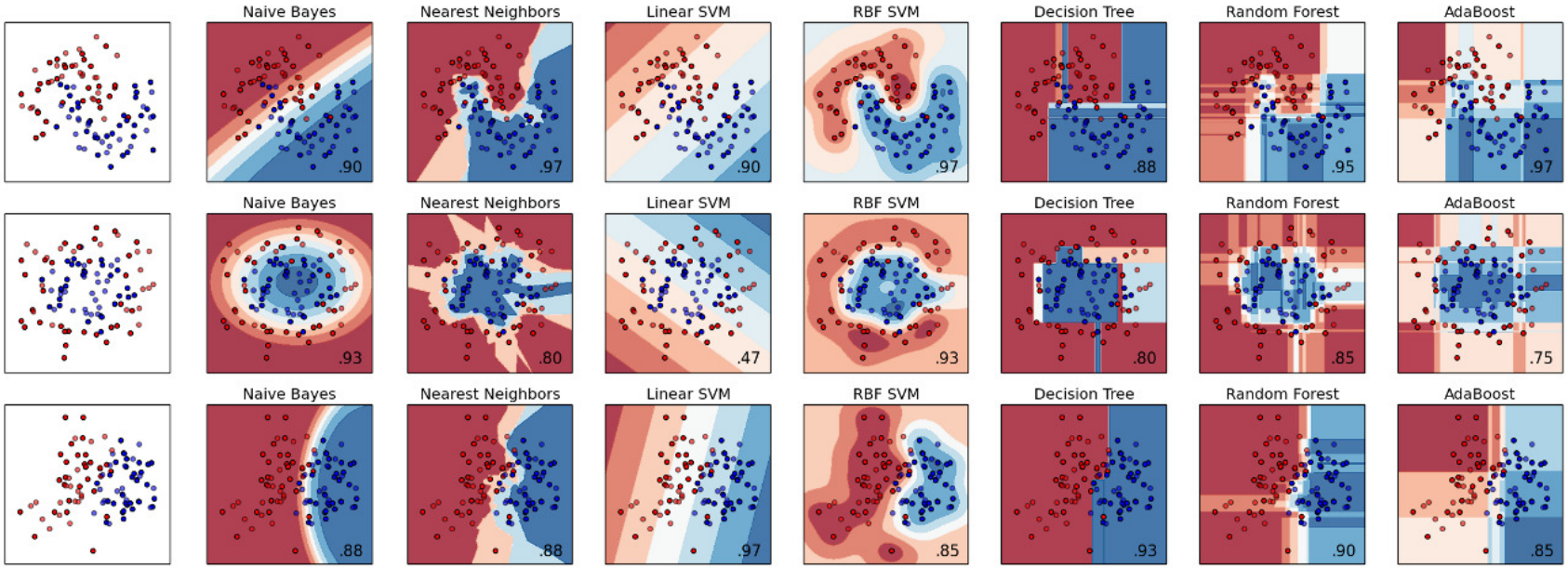
  Tatarchuk A., Urlov E., Mottl V., Windridge D. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. 2010.

Какие методы классификации наиболее универсальны?

Метода, подходящего для решения любых задач, не существует!

Тем не менее, для многих задач оказались успешными:

- градиентный бустинг деревьев решений и случайный лес
- SVM с ядрами
- нейронные сети при тщательном подборе архитектуры



Основные типы методов машинного обучения

- минимизация эмпирического риска
  MVR, Linear Regression, Logistic Regression
- регуляризация эмпирического риска
  SVM, RLR, ElasticNet, LASSO, Least Angle Regression
- метрические методы
  kNN, RBF, Kernel Regression, Kernel Density Estimation
- логические методы
  Decision Tree, Decision Forest, Rule Induction
- байесовские методы
  Naive Bayes, Linear Discriminant, Bayesian Networks
- нейросетевые методы
  BackPropagation, Deep Belief Nets, Deep Learning
- композиционные методы
  Boosting, Bagging, Stacking, Яндекс.MatrixNet

Основные типы задач машинного обучения

1. Предварительная обработка (data preparation)
   - извлечение признаков (feature extraction)
   - отбор признаков (feature selection)
   - восстановление пропусков (missing values)
2. Обучение с учителем (supervised learning)
   - классификация (classification)
   - регрессия (regression)
   - ранжирование (learning to rank)
   - прогнозирование (forecasting)
   - одноклассовая классификация (one-class classification, outlier/anomaly/novelty detection)
3. Обучение без учителя (unsupervised learning)
   - кластеризация (clustering)
   - восстановление плотности (density estimation)
   - поиск ассоциативных правил (association rule learning)

Основные типы задач машинного обучения

4. Частичное обучение (semi-supervised learning)
   - трансдуктивное обучение (transductive learning)
5. Привилегированное обучение (privilege learning)
6. Обучение представлений (representation learning)
   - обучение признаков (feature learning)
   - обучение многообразий (manifold learning)
   - анализ главных компонент (principal component analysis)
   - матричные разложения (matrix factorization)
   - коллаборативная фильтрация (collaborative filtering)
   - тематическое моделирование (topic modeling)
7. Динамическое обучение (online/incremental learning)
8. Обучение с подкреплением (reinforcement learning)
9. Активное обучение (active learning)
10. Мета-обучение (meta-learning)

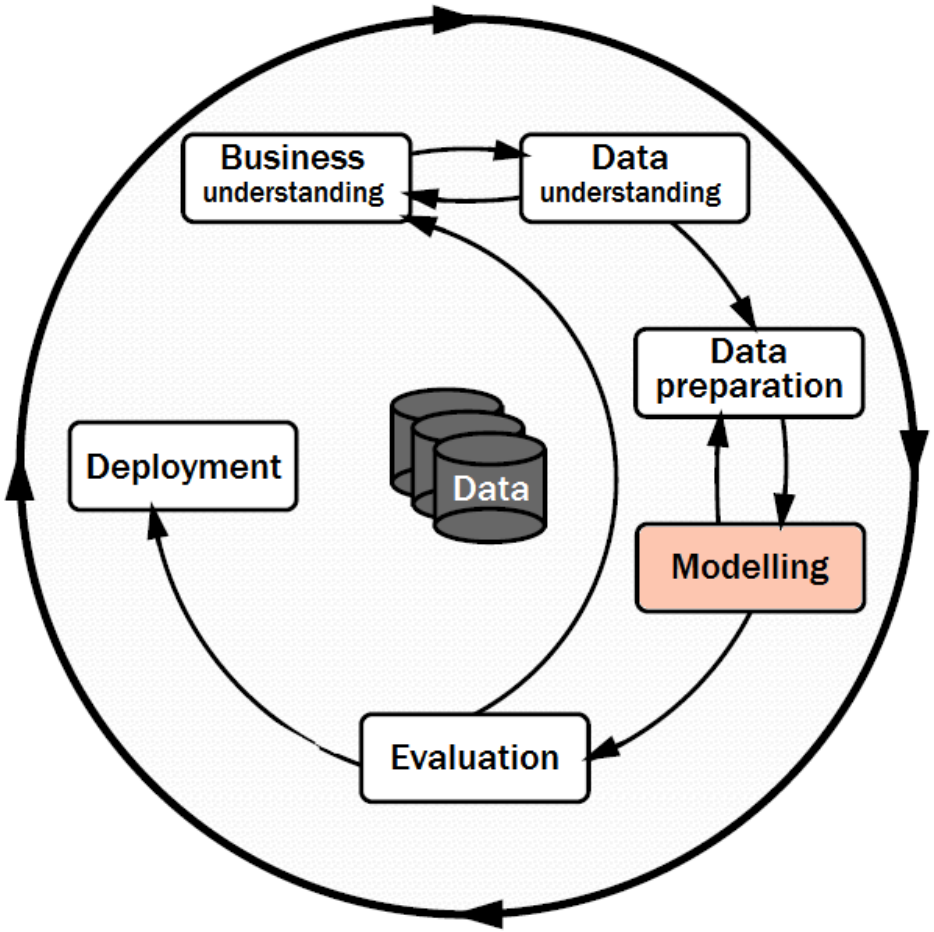
Особенности данных и постановок прикладных задач

- разнородные (признаки измерены в разных шкалах)
- неполные (измерены не все, имеются пропуски)
- неточные (измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- неструктурированные (нет признаковых описаний)

- заказчик не знает точно, чего хочет
- критерии качества нетривиальны или неясны
- заказчик не заботится о качестве своих данных

CRISP-DM: CRoss Industry Standard Process for Data Mining

CRISP-DM — межотраслевой стандарт интеллектуального анализа данных (1999)



Компании -инициаторы:

- SPSS
- Teradata
- Daimler AG
- NCR Corp.
- OHRA

Этапы решения задач машинного обучения

- понимание задачи и данных
- предобработка данных и изобретение признаков
- построение модели
- сведение обучения к оптимизации
- решение проблем оптимизации и переобучения
- оценивание качества решения
- внедрение и эксплуатация.

Арсенал технологий

Экосистемы машинного обучения:

- Python + SciPy + SciKit-Learn
- Java + Weka
- R
- Deductor — аналитическая платформа BaseGroup Labs

Инструменты для хранения и обработки больших данных:

- Hadoop — распределённое хранение данных
- Spark — распределённые вычисления

Инструменты для обучения нейронных сетей:

- TensorFlow
- Theano
- Torch

Полезные ссылки

- www.MachineLearning.ru — русскоязычная вики
- www.kdnuggets.com — главный сайт датамайнеров
- www.datasciencecentral.com — 72 000 датамайнеров
- www.kaggle.com — конкурсы анализа данных
- 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

Литература

- *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.
- *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.
- *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. 2004–2016.