

# Машинное обучение: вводная лекция

Воронцов Константин Вячеславович

vokov@forecsys.ru

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

Этот курс доступен на странице вики-ресурса

<http://www.MachineLearning.ru/wiki>

«Машинное обучение (курс лекций, К.В.Воронцов)»

Видеолекции: <http://shad.yandex.ru/lectures>

ШАД Яндекс • 13 февраля 2019

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

*Клаус Мартин Шваб,*

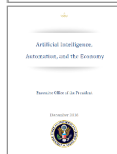
президент Всемирного экономического форума



Мир наконец поверил в искусственный интеллект.  
Машинное обучение — новый двигатель прогресса.  
Машинное обучение изменит мир. И уже меняет!

«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
- Автоматизация банковских услуг (FinTech)
- Автоматизация юридических услуг (LegalTech)
- Мониторинг сельского хозяйства
- Персональная медицина
- Персональные образовательные траектории
- Автономные системы вооружений



- 1 Основные понятия и обозначения**
  - Данные в задачах обучения по прецедентам
  - Модели и методы обучения
  - Обучение и переобучение
- 2 Примеры прикладных задач**
  - Задачи классификации
  - Задачи регрессии
  - Задачи ранжирования
- 3 Методология машинного обучения**
  - Особенности данных
  - Межотраслевой стандарт CRISP-DM
  - Эксперименты на синтетических и реальных данных

## Задача обучения по прецедентам

$X$  — множество *объектов*;

$Y$  — множество *ответов*;

$y: X \rightarrow Y$  — неизвестная зависимость (target function).

**Дано:**

$\{x_1, \dots, x_\ell\} \subset X$  — *обучающая выборка* (training sample);

$y_i = y(x_i)$ ,  $i = 1, \dots, \ell$  — известные ответы.

**Найти:**

$a: X \rightarrow Y$  — алгоритм, решающую функцию (decision function), приближающую  $y$  на всём множестве  $X$ .

Весь курс машинного обучения — это конкретизация:

- как задаются объекты и какими могут быть ответы;
- в каком смысле « $a$  приближает  $y$ »;
- как строить функцию  $a$ .

## Как задаются объекты. Признаковое описание

$f_j: X \rightarrow D_j, j = 1, \dots, n$  — признаки объектов (features).

Типы признаков:

- $D_j = \{0, 1\}$  — *бинарный* признак  $f_j$ ;
- $|D_j| < \infty$  — *номинальный* признак  $f_j$ ;
- $|D_j| < \infty, D_j$  упорядочено — *порядковый* признак  $f_j$ ;
- $D_j = \mathbb{R}$  — *количественный* признак  $f_j$ .

Вектор  $(f_1(x), \dots, f_n(x))$  — *признаковое описание* объекта  $x$ .

Матрица «объекты–признаки» (feature data)

$$F = \parallel f_j(x_i) \parallel_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

## Как задаются ответы. Типы задач

### Задачи классификации (classification):

- $Y = \{-1, +1\}$  — классификация на 2 класса.
- $Y = \{1, \dots, M\}$  — на  $M$  непересекающихся классов.
- $Y = \{0, 1\}^M$  — на  $M$  классов, которые могут пересекаться.

### Задачи восстановления регрессии (regression):

- $Y = \mathbb{R}$  или  $Y = \mathbb{R}^m$ .

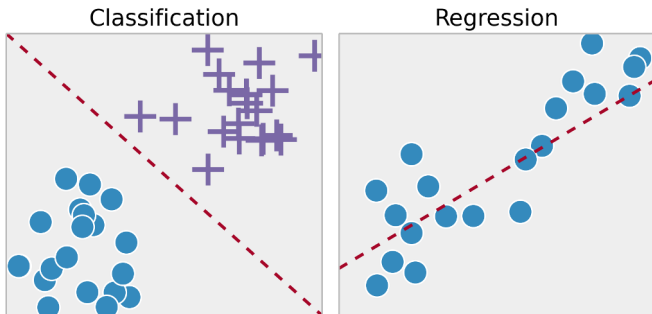
### Задачи ранжирования (ranking, learning to rank):

- $Y$  — конечное упорядоченное множество.

## Статистическое (машинное) обучение с учителем

- = обучение по прецедентам
- = восстановление зависимостей по эмпирическим данным
- = предсказательное моделирование
- = проведение функции через заданные точки

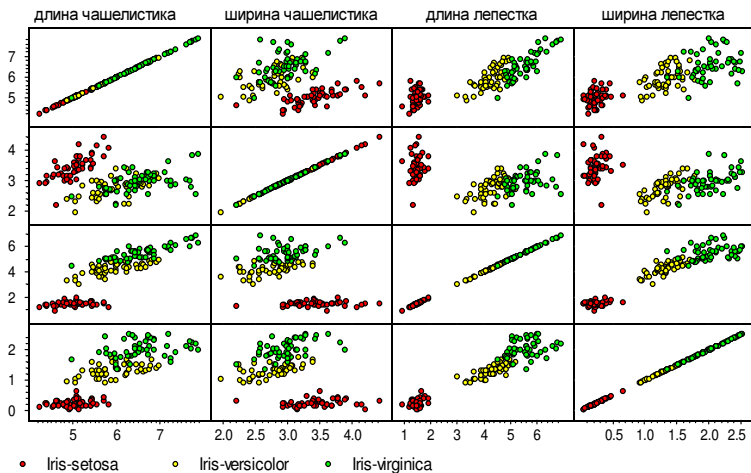
Два основных типа задач — *классификация* и *регрессия*





## Пример: задача классификации цветков ириса [Фишер, 1936]

$n = 4$  признака,  $|Y| = 3$  класса, длина выборки  $\ell = 150$ .



## Модель алгоритмов (предсказательная модель)

*Модель* (predictive model) — параметрическое семейство функций

$$A = \{g(x, \theta) \mid \theta \in \Theta\},$$

где  $g: X \times \Theta \rightarrow Y$  — фиксированная функция,  
 $\Theta$  — множество допустимых значений параметра  $\theta$ .

**Пример.**

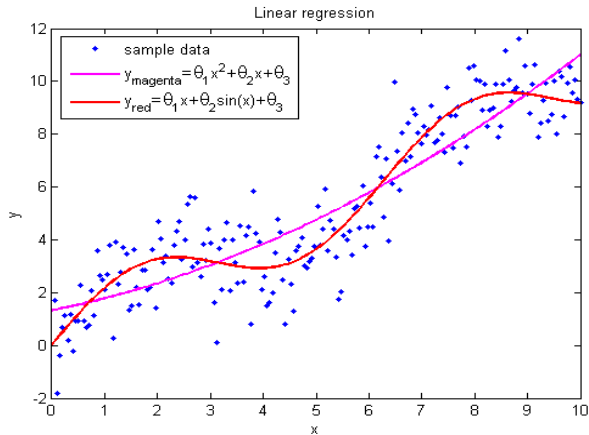
*Линейная модель* с вектором параметров  $\theta = (\theta_1, \dots, \theta_n)$ ,  $\Theta = \mathbb{R}^n$ :

$g(x, \theta) = \sum_{j=1}^n \theta_j f_j(x)$  — для регрессии и ранжирования,  $Y = \mathbb{R}$ ;

$g(x, \theta) = \text{sign} \sum_{j=1}^n \theta_j f_j(x)$  — для классификации,  $Y = \{-1, +1\}$ .

## Пример: задача регрессии, синтетические данные

$X = Y = \mathbb{R}$ ,  $\ell = 200$ ,  $n = 3$  признака:  $\{x, x^2, 1\}$  или  $\{x, \sin x, 1\}$



**Вывод:** признаковое описание можно задавать по-разному

## Метод обучения

*Метод обучения* (learning algorithm) — это отображение вида

$$\mu: (X \times Y)^\ell \rightarrow A,$$

которое произвольной выборке  $X^\ell = (x_i, y_i)_{i=1}^\ell$  ставит в соответствие некоторый алгоритм  $a \in A$ .

В задачах обучения по прецедентам всегда есть два этапа:

- *Этап обучения* (training):  
метод  $\mu$  по выборке  $X^\ell$  строит алгоритм  $a = \mu(X^\ell)$ .
- *Этап применения* (testing):  
алгоритм  $a$  для новых объектов  $x$  выдаёт ответы  $a(x)$ .

## Этап обучения и этап применения

**Этап обучения (train):**

метод  $\mu$  по выборке  $X^\ell = (x_i, y_i)_{i=1}^\ell$  строит алгоритм  $a = \mu(X^\ell)$ :

$$\boxed{\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix} \xrightarrow{\mu} a$$

**Этап применения (test):**

алгоритм  $a$  для новых объектов  $x'_i$  выдаёт ответы  $a(x'_i)$ .

$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1) \\ \dots \\ a(x'_k) \end{pmatrix}$$

## Функционалы качества

$\mathcal{L}(a, x)$  — функция потерь (loss function) — величина ошибки алгоритма  $a \in A$  на объекте  $x \in X$ .

**Функции потерь для задач классификации:**

- $\mathcal{L}(a, x) = [a(x) \neq y(x)]$  — индикатор ошибки;

**Функции потерь для задач регрессии:**

- $\mathcal{L}(a, x) = |a(x) - y(x)|$  — абсолютное значение ошибки;
- $\mathcal{L}(a, x) = (a(x) - y(x))^2$  — квадратичная ошибка.

*Эмпирический риск* — функционал качества алгоритма  $a$  на  $X^\ell$ :

$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(a, x_i).$$

## Сведение задачи обучения к задаче оптимизации

Метод минимизации эмпирического риска:

$$\mu(X^\ell) = \arg \min_{a \in A} Q(a, X^\ell).$$

Пример: метод наименьших квадратов ( $Y = \mathbb{R}$ ,  $\mathcal{L}$  квадратична):

$$\mu(X^\ell) = \arg \min_{\theta} \sum_{i=1}^{\ell} (g(x_i, \theta) - y_i)^2.$$

Проблема обобщающей способности:

- найдём ли мы «закон природы» или *переобучимся*, то есть подгоним функцию  $g(x_i, \theta)$  под заданные точки?
- будет ли  $a = \mu(X^\ell)$  приближать функцию  $y$  на всём  $X$ ?
- будет ли  $Q(a, X^k)$  мало на новых данных — контрольной выборке  $X^k = (x'_i, y'_i)_{i=1}^k$ ,  $y'_i = y(x_i)$ ?

## Пример переобучения

Зависимость  $y(x) = \frac{1}{1 + 25x^2}$  на отрезке  $x \in [-2, 2]$ .

Признаковое описание  $x \mapsto (1, x^1, x^2, \dots, x^n)$ .

Модель полиномиальной регрессии

$$a(x, \theta) = \theta_0 + \theta_1 x + \dots + \theta_n x^n \quad \text{— полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(\theta, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \dots + \theta_n x_i^n - y_i)^2 \rightarrow \min_{\theta_0, \dots, \theta_n}.$$

Обучающая выборка:  $X^\ell = \{x_i = 4 \frac{i-1}{\ell-1} - 2 \mid i = 1, \dots, \ell\}$ .

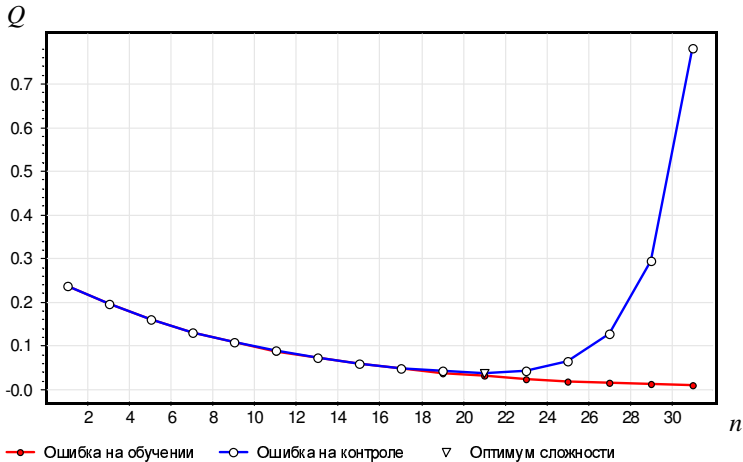
Контрольная выборка:  $X^k = \{x_i = 4 \frac{i-0.5}{\ell-1} - 2 \mid i = 1, \dots, \ell - 1\}$ .

Что происходит с  $Q(\theta, X^\ell)$  и  $Q(\theta, X^k)$  при увеличении  $n$ ?



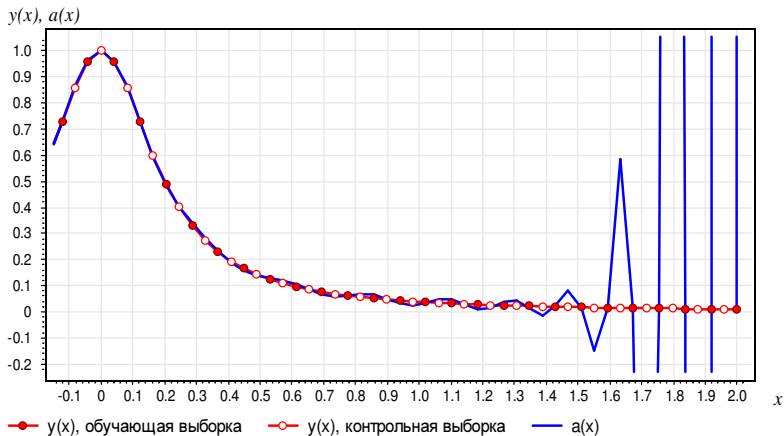
## Пример переобучения: эксперимент при $\ell = 50$ , $n = 1..31$

Переобучение — это когда  $Q(\mu(X^\ell), X^k) \gg Q(\mu(X^\ell), X^\ell)$ :



## Пример переобучения: эксперимент при $\ell = 50$

$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) \text{ — полином степени } n = 38$$



## Переобучение — одна из проблем машинного обучения

- 1 Из-за чего возникает переобучение?**
  - избыточная сложность пространства параметров  $\Theta$ , лишние степени свободы в модели  $g(x, \theta)$  «тратятся» на чрезмерно точную подгонку под обучающую выборку.
  - переобучение есть всегда, когда есть оптимизация параметров по конечной (заведомо неполной) выборке.
- 2 Как обнаружить переобучение?**
  - эмпирически, путём разбиения выборки на train и test.
- 3 Избавиться от него нельзя. Как его минимизировать?**
  - минимизировать одну из теоретических оценок;
  - накладывать ограничения на  $\theta$  (регуляризация);
  - минимизировать HoldOut, LOO или CV, но осторожно!

## Эмпирические оценки обобщающей способности

- Эмпирический риск на тестовых данных (hold-out):

$$\text{HO}(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \rightarrow \min$$

- Скользящий контроль (leave-one-out),  $L = \ell + 1$ :

$$\text{LOO}(\mu, X^L) = \frac{1}{L} \sum_{i=1}^L \mathcal{L}(\mu(X^L \setminus \{x_i\}), x_i) \rightarrow \min$$

- Кросс-проверка (cross-validation),  $L = \ell + k$ ,  $X^L = X_n^\ell \sqcup X_n^k$ :

$$\text{CV}(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} Q(\mu(X_n^\ell), X_n^k) \rightarrow \min$$

- Эмпирическая оценка вероятности переобучения:

$$Q_\varepsilon(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} \left[ Q(\mu(X_n^\ell), X_n^k) - Q(\mu(X_n^\ell), X_n^\ell) \geq \varepsilon \right] \rightarrow \min$$

## 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.

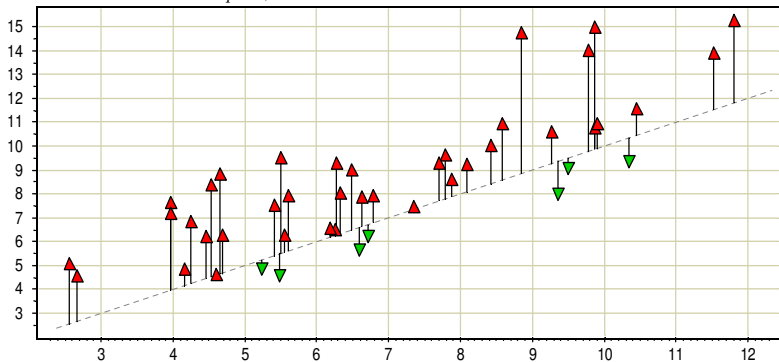
**Особенности задачи:**

- обычно много «пропусков» в данных;
- нужен интерпретируемый алгоритм классификации;
- нужно выделять *синдромы* — сочетания *симптомов*;
- нужна оценка вероятности отрицательного исхода.

## Задача медицинской диагностики. Пример переобучения

Задача предсказания отдалённого результата хирургического лечения атеросклероза. Точки — различные алгоритмы.

*Частота ошибок на контроле, %*



*Частота ошибок на обучении, %*

## Задача кредитного скоринга

**Объект** — заявка на выдачу банком кредита.

**Классы** — bad или good.

**Примеры признаков:**

- **бинарные:** пол, наличие телефона, и т. д.
- **номинальные:** место проживания, профессия, работодатель, и т. д.
- **порядковые:** образование, должность, и т. д.
- **количественные:** возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т. д.

**Особенности задачи:**

- нужно оценивать вероятность дефолта  $P(\text{bad})$ .

## 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:**

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



## Задача категоризации текстовых документов

**Объект** — текстовый документ.

**Классы** — рубрики иерархического тематического каталога.

**Примеры признаков:**

- **номинальные:** автор, издание, год, и т. д.
- **количественные:** для каждого термина — частота в тексте, в заголовках, в аннотации, и т. д.

**Особенности задачи:**

- лишь небольшая часть документов имеют метки  $y_i$ ;
- документ может относиться к нескольким рубрикам;
- в каждом ребре дерева свой классификатор на 2 класса.

## Задачи биометрической идентификации личности

Идентификация личности по отпечаткам пальцев



Идентификация личности по радужной оболочке глаза



Особенности задач:

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

## 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

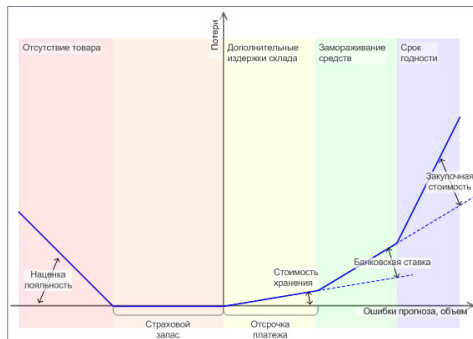
Объект — тройка (товар, магазин, день).

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:**

- демографические данные: возраст, достаток и т.д.,
- цены на недвижимость поблизости,
- маркетинговые данные: наличие школ, офисов и т.д.

**Особенности задачи:**

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

## Задача ранжирования поисковой выдачи

**Объект** — пара ⟨короткий текстовый запрос, документ⟩.

**Классы** — релевантен или не релевантен,  
разметка делается людьми — ассессорами.

**Примеры количественных признаков:**

- частота слов запроса в документе,
- число ссылок на документ,
- число кликов на документ: всего, по данному запросу.

**Особенности задачи:**

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

## Задача тематического информационного поиска

**Объект** — пара ⟨длинный текстовый запрос, документ⟩.

**Предсказать** — оптимальный порядок чтения документов.

**Примеры признаков:**

- близость тематических представлений пары текстов,
- широта/узость тематики,
- широта/узость, уникальность, актуальность терминологии,
- когнитивная сложность текста.

**Особенности задачи:**

- темы латентные, их надо сначала выявить;
- разнообразие лингвистической предобработки текста;
- графическая поисковая выдача в виде «карты знаний».

## Конкурс kaggle.com: Avito Context Ad Clicks Prediction

**Объект** — тройка ⟨пользователь, объявление, баннер⟩.

**Предсказать** — кликнет ли пользователь по контекстной рекламе, которую показали в ответ на его запрос на avito.ru.

**Сырые данные:**

- все действия пользователя на сайте,
- профиль пользователя (браузер, устройство и т. д.),
- история показов и кликов других пользователей по баннеру,
- ... всего 10 таблиц данных.

**Особенности задачи:**

- признаки надо придумывать;
- данных много — сотни миллионов показов;
- основной критерий качества — доход рекламной площадки;



## 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

## Достижения в анализе сложно структурированных данных

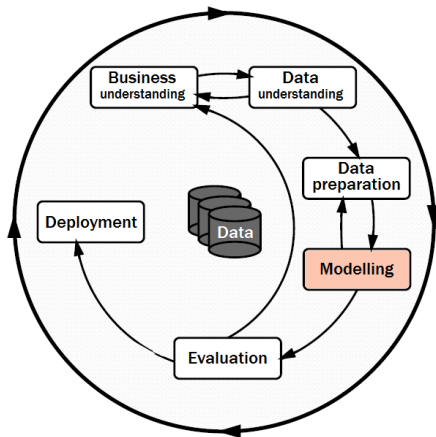
- **Статистический машинный перевод:**  
объект — предложение на естественном языке  
ответ — его перевод на другой язык
- **Перевод речи в текст:**  
объект — аудиозапись речи человека  
ответ — текстовая запись речи
- **Компьютерное зрение:**  
объект — изображение или последовательность видеок кадров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

Предпосылки этих прорывов:

- Большие и *чистые* данные (Big Data)
- Глубокие нейронные сети (Deep Learning)
- Рост вычислительной мощности (закон Мура, GPU)

## Межотраслевой стандарт интеллектуального анализа данных

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



Компании-инициаторы:

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

Шаги процесса:

- понимание бизнеса
- понимание данных
- предобработка данных
- инженерия признаков
- разработка моделей
- настройка параметров
- оценивание качества
- внедрение

## Эксперименты на реальных данных

### Эксперименты на конкретной прикладной задаче:

- цель — решить задачу как можно лучше
- важно понимание задачи и данных
- важно придумывать информативные признаки
- конкурсы по анализу данных: <http://www.kaggle.com>
- отечественная платформа: <http://DataRing.ru>

### Эксперименты на наборах прикладных задач:

- цель — протестировать метод в разнообразных условиях
- нет необходимости (и времени) разбираться в сути задач : (
- признаки, как правило, уже кем-то придуманы
- репозиторий UC Irvine Machine Learning Repository  
<http://archive.ics.uci.edu/ml> (308 задач, 09-02-2015)

## Эксперименты на синтетических данных

Используются для тестирования новых методов обучения.  
Преимущество — мы знаем истинную  $y(x)$  (ground truth)

### Эксперименты на синтетических данных:

- цель — отладить метод, выявить границы применимости
- объекты  $x_i$  из придуманного распределения (часто 2D)
- ответы  $y_i = y(x_i)$  для придуманной функции  $y(x)$
- двумерные данные + визуализация выборки

### Эксперименты на полу-синтетических данных:

- цель — протестировать помехоустойчивость модели
- объекты  $x_i$  из реальной задачи (+ шум)
- ответы  $y_i = a(x_i)$  для полученного решения  $a(x)$  (+ шум)

- **Основные понятия машинного обучения:**  
объект, ответ, признак, алгоритм, модель алгоритмов, метод обучения, эмпирический риск, переобучение.
- **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;
  - внедрение и эксплуатация.
- **Прикладные задачи машинного обучения:**  
очень много, очень разных,  
во всех областях бизнеса, науки, производства.