

# Введение в анализ данных

Лекция 2

Задачи анализа данных

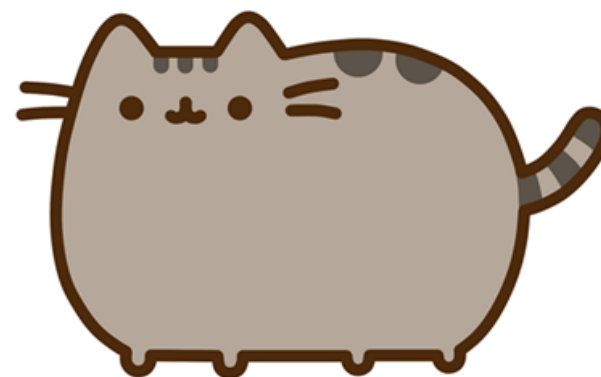
Евгений Соколов

[sokolov.evg@gmail.com](mailto:sokolov.evg@gmail.com)

НИУ ВШЭ, 2016

# Организационное

- [http://wiki.cs.hse.ru/Майнор\\_Интеллектуальный\\_анализ\\_данных/Введение\\_в\\_анализ\\_данных](http://wiki.cs.hse.ru/Майнор_Интеллектуальный_анализ_данных/Введение_в_анализ_данных)
- Почта: [hse.minor.dm@gmail.com](mailto:hse.minor.dm@gmail.com)
- Отзывы: <http://goo.gl/forms/RwdMxnChST>
- Переносы занятий
- Проекты



# Напоминание

- $\mathbb{X}$  — пространство объектов,  $\mathbb{Y}$  — пространство ответов
- $x = (x^1, \dots, x^d)$  — признаковое описание
- $X = (x_i, y_i)_{i=1}^{\ell}$  — обучающая выборка
- $a(x)$  — алгоритм, модель
- $Q(a, X)$  — функционал качества алгоритма  $a$  на выборке  $X$
- Обучение:  $a(x) = \arg \max_{a \in \mathcal{A}} Q(a, X)$

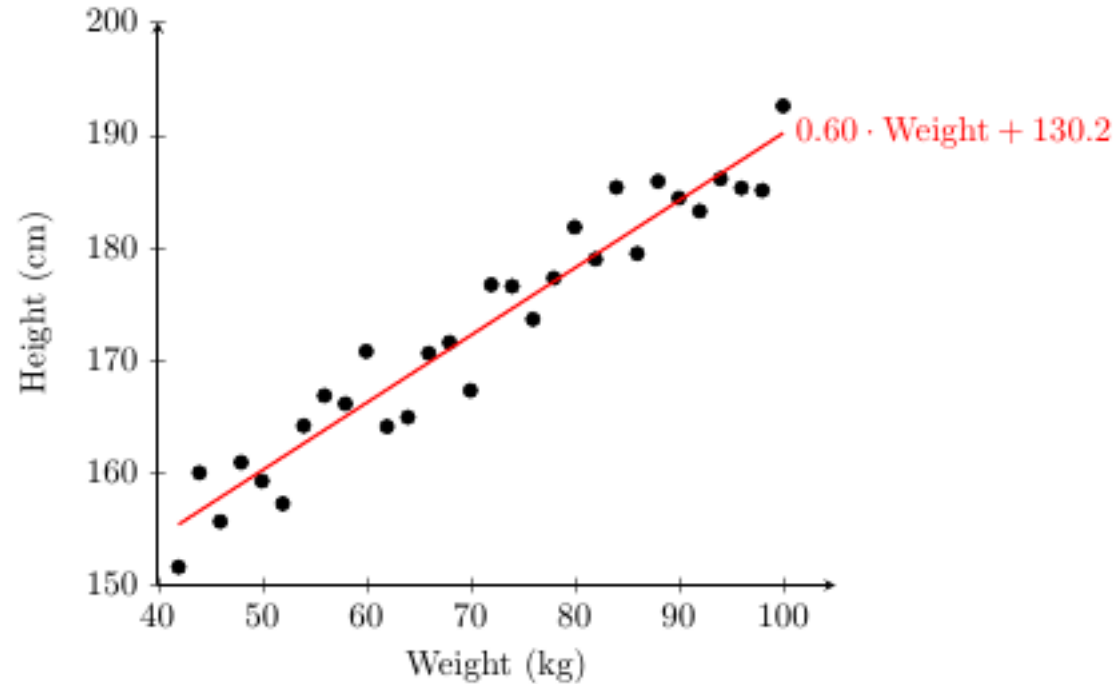
# Вопросы на сегодня

- Какие бывают ответы?
- Какие бывают признаки?
- Что такое переобучение?
- Какие задачи можно решать машинным обучением?

Типы ответов и признаков

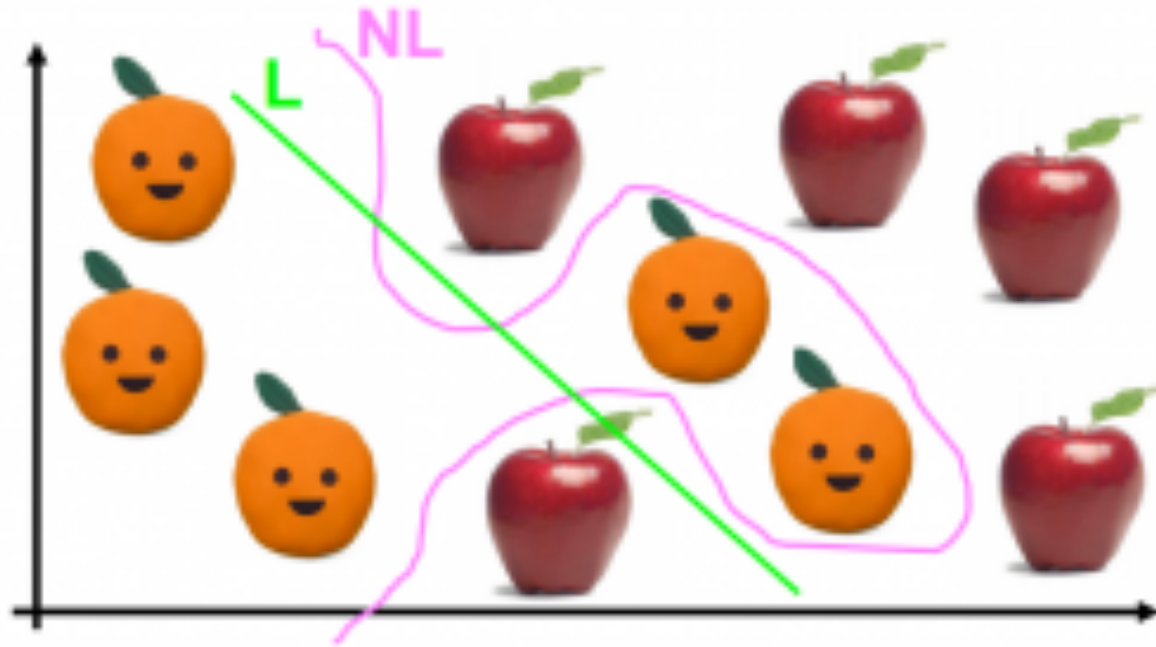
# Регрессия

- Вещественные ответы:  $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



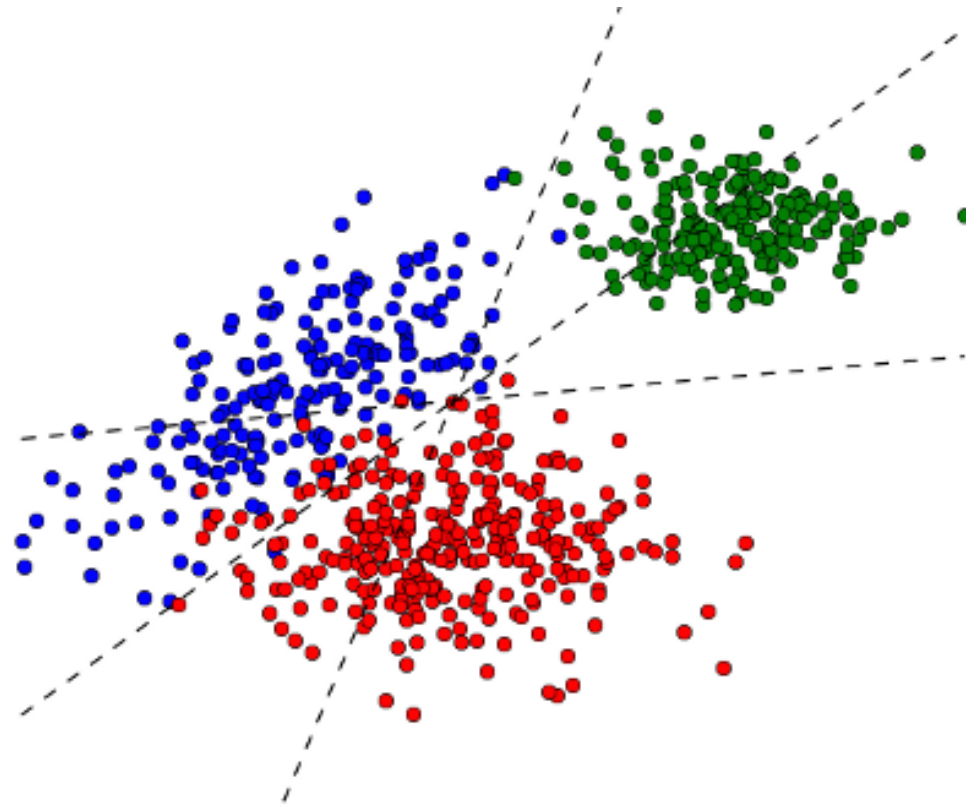
# Классификация

- Конечное число ответов:  $|\mathbb{Y}| < \infty$
- Бинарная классификация:  $\mathbb{Y} = \{-1, +1\}$



# Классификация

- Многоклассовая классификация:  $\mathbb{Y} = \{1, 2, \dots, K\}$





# Классификация

- Классификация с пересекающимися классами:  $\mathbb{Y} = \{0, 1\}^K$ 
  - (multi-label classification)
- Ответ — набор из  $K$  нулей и единиц
- $i$ -й элемент ответа — принадлежит ли объект  $i$ -му классу
  
- Какие темы присутствуют в статье?
- (математика, биология, экономика)

# Ранжирование

- Набор документов  $d_1, \dots, d_n$
- Запрос  $q$
- Задача: отсортировать документы по *релевантности* запросу
- $a(q, d)$  — оценка релевантности

# Ранжирование

Яндекс

картинки с котиками — 5 млн ответов



Найти

Поиск

Картинки

Видео

Карты

Маркет

Ещё



**Картинки с кошками | Fun Cats — Забавные коты**

[funcats.by](#) > [pictures/](#) ▼

**Картинки с кошками.** Прикольные коты. 777 **изображений**. ... 32 **изображения**. Кошки Стамбула. 41 **изображение**. Веселые котята.



**Уморные котики (57 фото) » Бяки.нет | Картинки**

[byaki.net](#) > **Картинки** > [14026-umornye-kotiki-57...](#) ▼

Бяки нет! . NET. Уморные **котики (57 фото)**. 223. Комментарии:9Автор:4ertonok  
Просмотров:161 395 **Картинки**28-10-2008, 00:03.



**Смешные картинки кошек с надписями | Лолкот.Ру**

[lolkot.ru](#) ▼

Смешные **картинки** для новых приколов! Сделать свой прикол очень просто. ... **Котик** верит в чудеса. Он в носке подарок ищет...



**Красивые картинки и фото кошек, котят и котов**

[foto-zverey.ru](#) > **Кошки** ▼

**Фото и картинки** кошек и котят потрясающей красоты и нежности. Здесь мы собрали такие **изображения**, которые всегда вызывают море положительных эмоций...



**Обои для рабочего стола Котят | картинки на стол Котят**

[7fon.ru](#) > Чёрные обои и **картинки** > Обои котят ▼

**Картинки Котят** с 1 по 15. **Обои** для рабочего стола Котят. ... Скачать **Картинки Котят** на рабочий стол бесплатно.

# Прогнозирование временных рядов

- Позже — на примере

# Построение рекомендательных систем

- Позже — на примере

# Кластеризация

- $Y$  — отсутствует
  - Нужно найти группы похожих объектов
  - Сколько таких групп?
  - Как измерить качество?
- 
- Пример: сегментация пользователей мобильного оператора

# Типы признаков

- $f_j$  —  $j$ -й признак
- $D_j$  — множество значений признака

# Бинарные признаки

- $D_j = \{0, 1\}$
- Доход клиента выше среднего по городу?
- Цвет фрукта — зеленый?



# Вещественные признаки

- $D_j = \mathbb{R}$
- Возраст
- Площадь квартиры
- Количество звонков в колл-центр

# Категориальные признаки

- $D_j$  — неупорядоченное множество
- Цвет глаз
- Город
- Образование (может быть упорядоченным)
- Очень трудны в обращении

# Порядковые признаки

- $D_j$  — упорядоченное множество
- Военское звание
- Роль в фильме (первого плана, второго плана, массовка)
- Тип населенного пункта

# Множественнозначные признаки

- (set-valued)
- $D_j$  — множество всех подмножеств некоторого множества
- Какие фильмы посмотрел пользователь?
- Какие слова входят в текст?

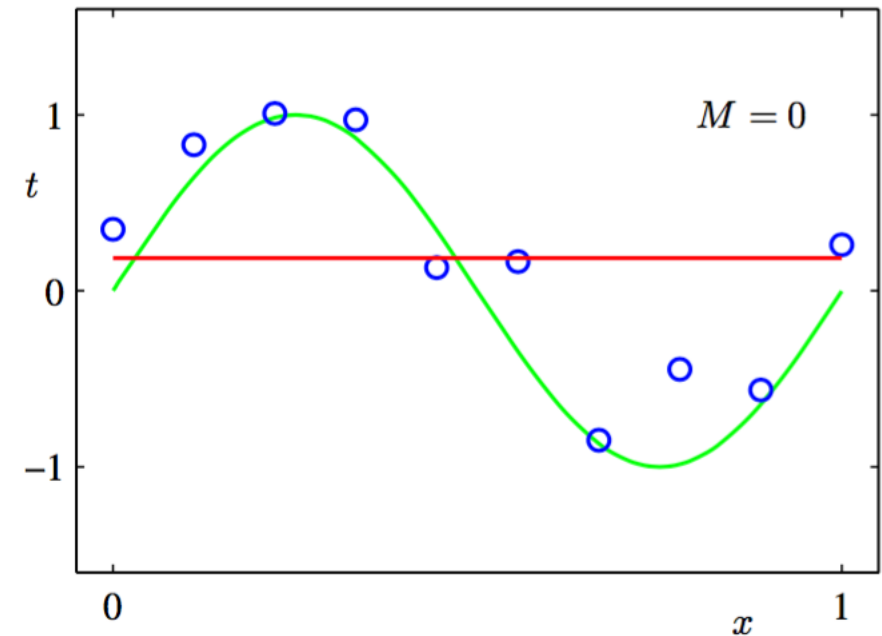
Обобщающая способность

# Обобщающая способность

- Выбираем алгоритм с лучшим качеством на обучающей выборке
- Как он будет вести себя на новых данных?
- Смог ли он выразить  $y$  через  $x$ ?

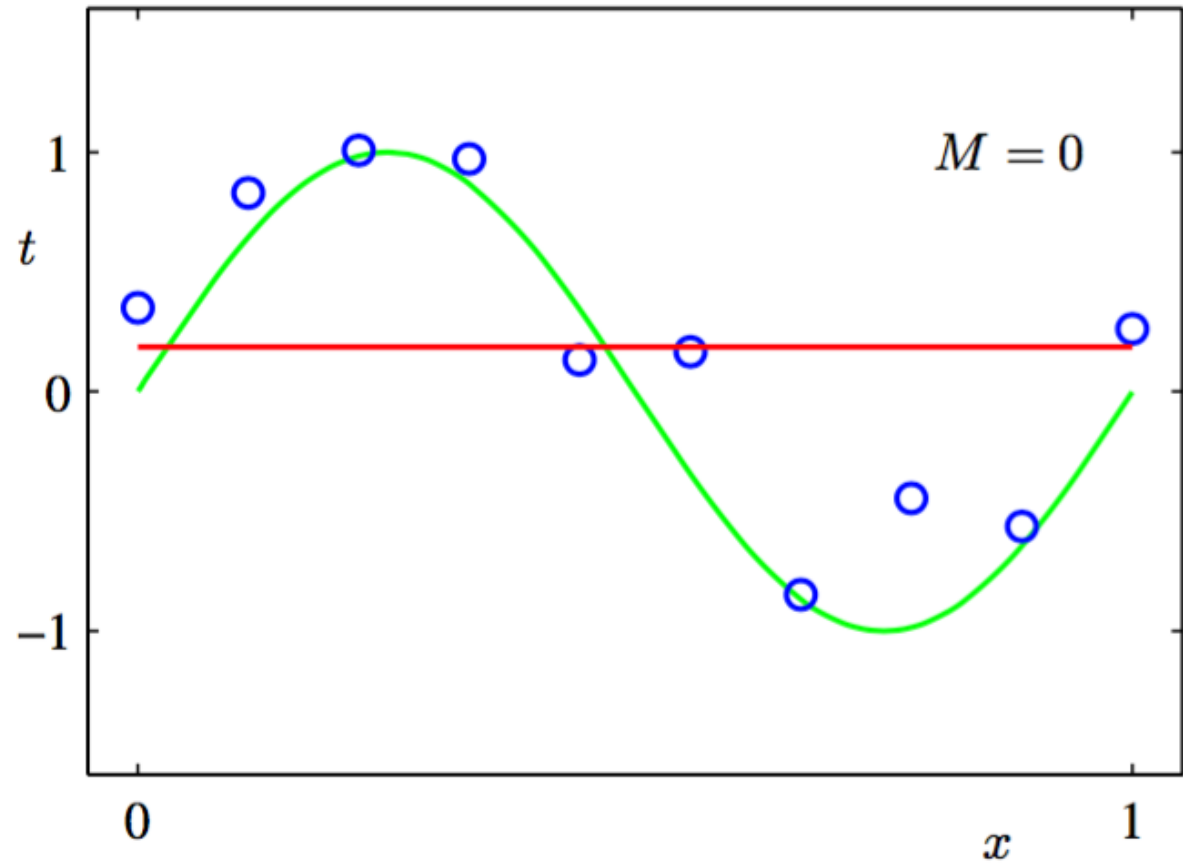
# Обобщающая способность

- Зеленый — истинная зависимость
- Красный — прогноз алгоритма
- Синий — выборка
- Линейный алгоритм



# Обобщающая способность

- Без признаков
- Константный алгоритм

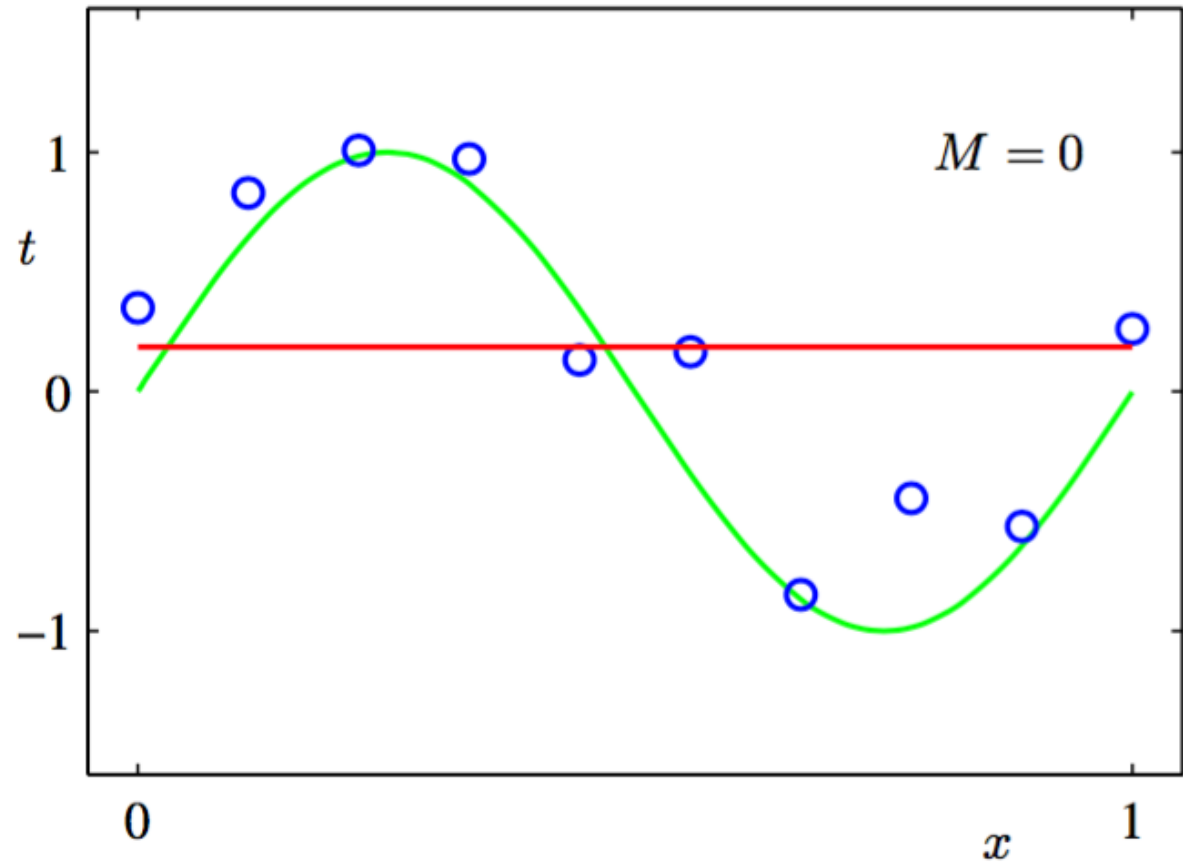




# Обобщающая способность

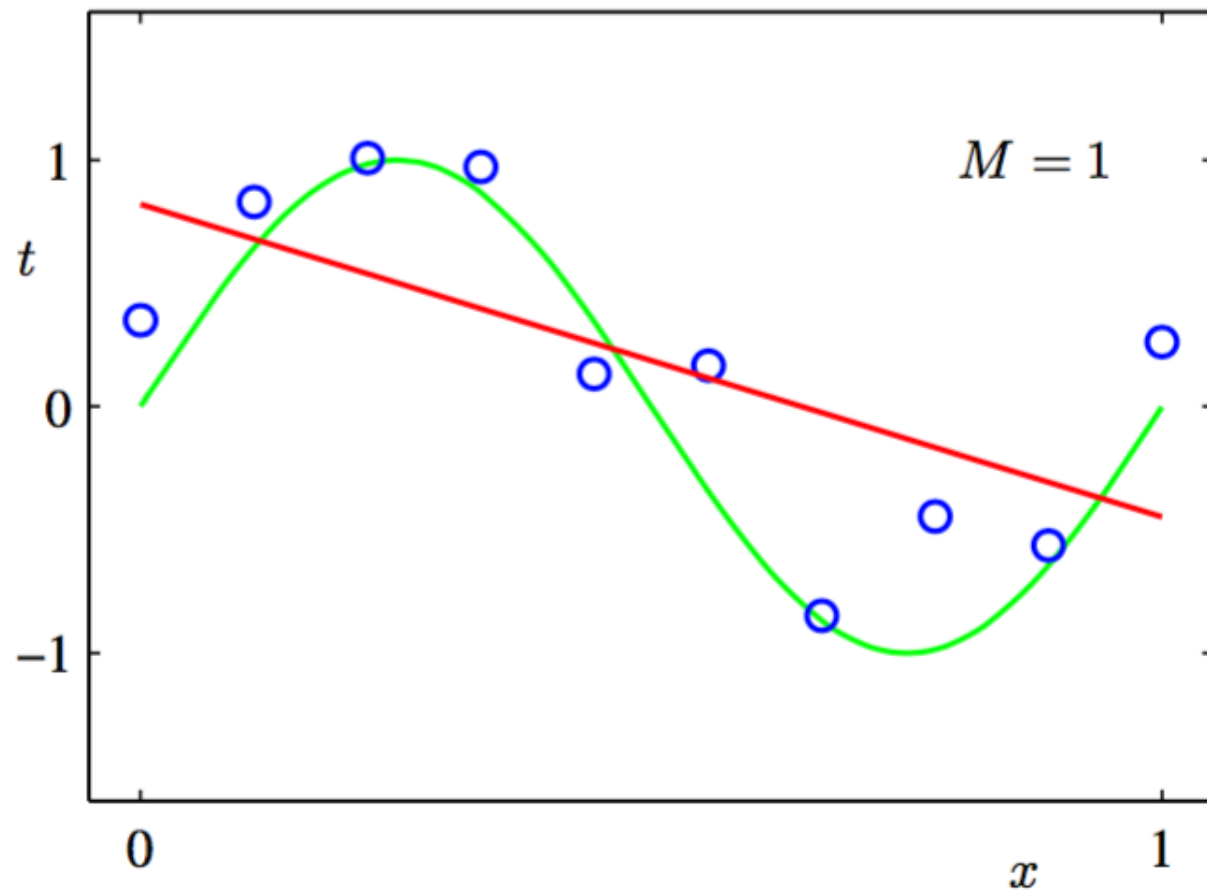
- Без признаков
- Константный алгоритм

**Недообучение**



# Обобщающая способность

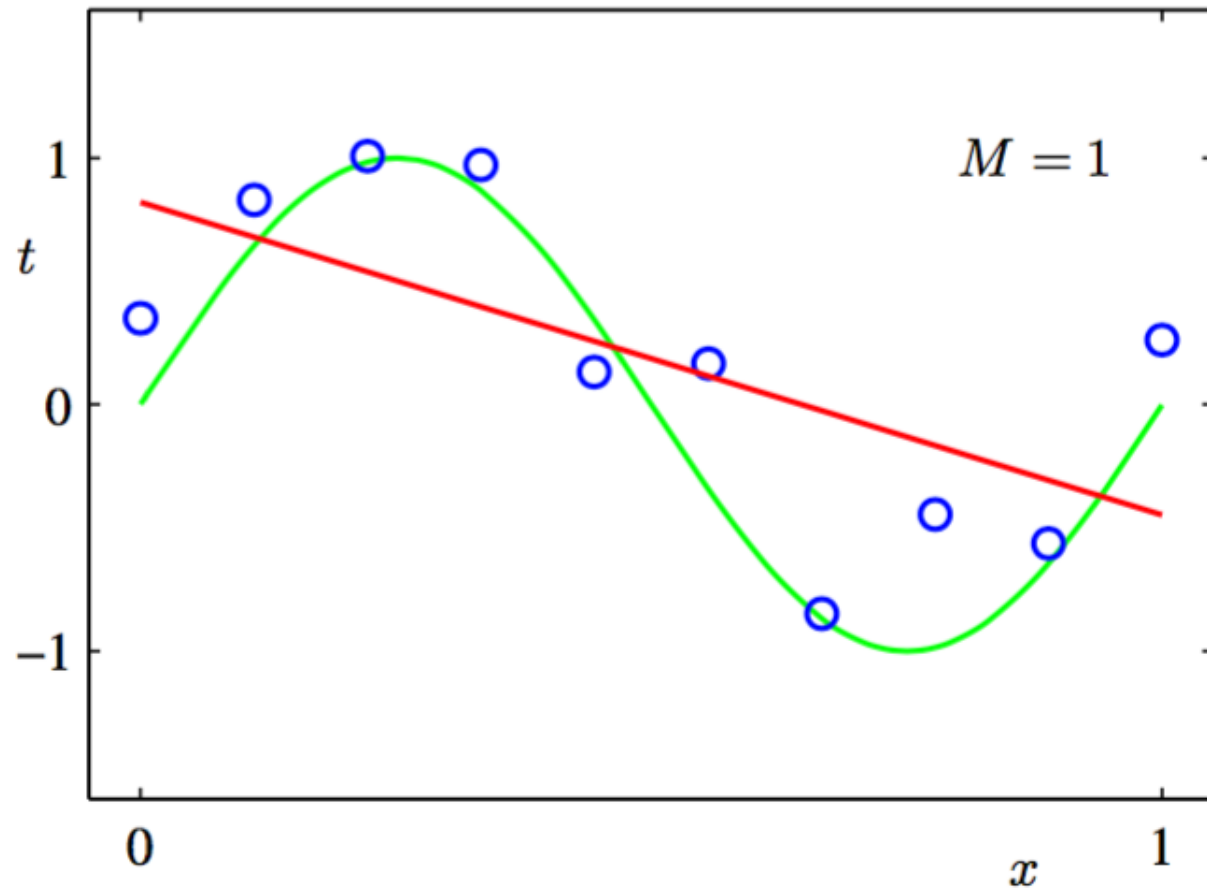
- 1 признак
- $x$



# Обобщающая способность

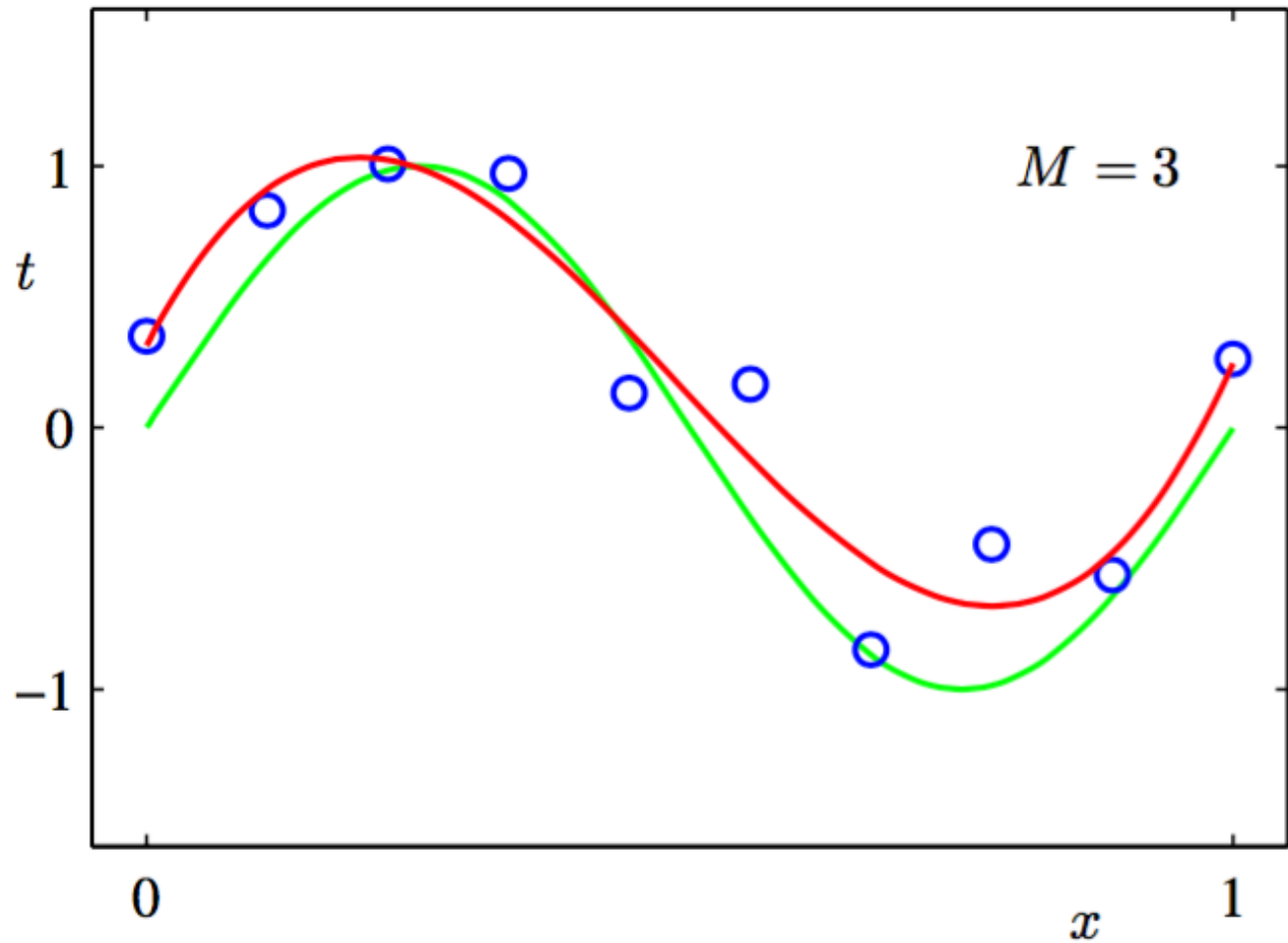
- 1 признак
- $x$

Недообучение



# Обобщающая способность

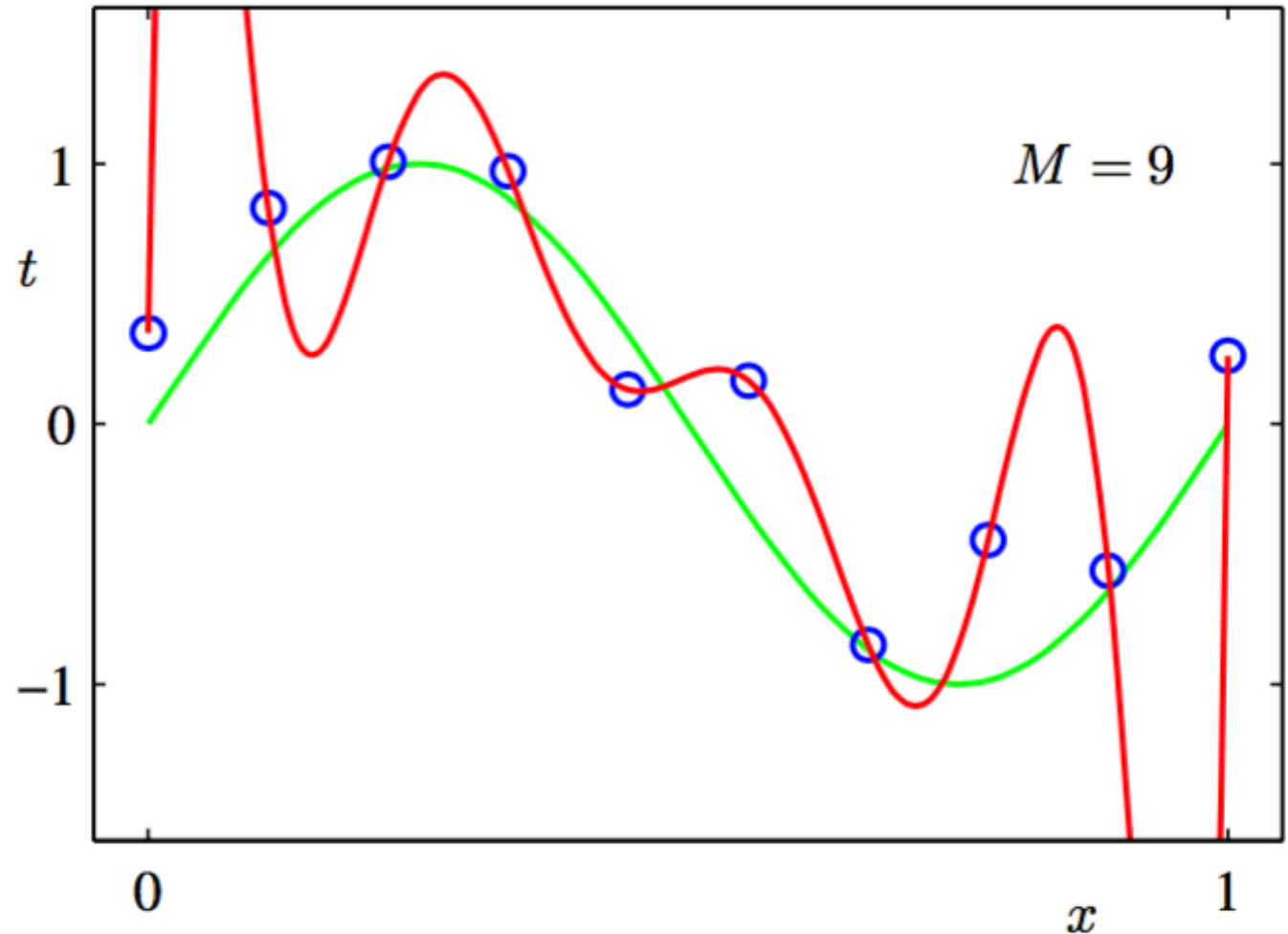
- 3 признака
- $x, x^2, x^3$



# Обобщающая способность

- 9 признаков
- $x, x^2, x^3, x^4, \dots, x^9$

**Переобучение  
(overfitting)**



# Обобщающая способность

- Недообучение — **плохое** качество на обучении и на новых данных
- Переобучение — **хорошее** качество на обучении, **плохое** на новых данных
- Переобучение — алгоритм запоминает ответы, а не находит закономерности

# Как выявить переобучение?

- Хороший алгоритм — хорошее качество на обучении
- Переобученный алгоритм — хорошее качество на обучении
- По обучающей выборке очень сложно выявить переобучение



# Как выявить переобучение?

- Отложенная выборка — данные, на которых не обучались
- Кросс-валидация
- Меры сложности модели



Задачи анализа данных

# Медицинская диагностика

- Объект — пациент в определенный момент времени
- Ответ — диагноз
- Классификация с пересекающимися классами

# Медицинская диагностика — признаки

- Бинарные: пол, головная боль, слабость, и т.д.
- Порядковые: тяжесть состояния, желтушность, и т.д.
- Вещественные: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т.д.

# Медицинская диагностика — особенности

- Много пропусков в данных (missing data)
- Недостаточный объем данных
- Алгоритм должен быть интерпретируемым
- Нужна оценка вероятности для каждого заболевания

# Кредитный скоринг

- Объект — заявка на выдачу кредита банком
- Ответ — вернет ли клиент кредит
- Бинарная классификация

# Кредитный скоринг — признаки

- Бинарные: пол, наличие телефона, и т.д.
- Категориальные: место жительства, профессия, семейный статус, работодатель, и т.д.
- Порядковые: образование, должность, и т.д.
- Вещественные: возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т.д.

# Кредитный скоринг — особенности

- Нужно оценивать вероятность дефолта

# Предсказание оттока клиентов

- Объект — абонент в определенный момент времени
- Ответ — уйдет или не уйдет в следующем месяце
- Бинарная классификация



# Предсказание оттока клиентов — признаки

- Бинарные: корпоративный клиент, подключенные услуги, и т.д.
- Категориальные: регион проживания, тарифный план, и т.д.
- Вещественные: длительность разговоров, количество СМС, частота оплаты, объем трафика, и т.д.

# Предсказание оттока клиентов — особенности

- Нужно оценивать вероятность ухода
- Сверхбольшие выборки
- Исходные данные — сырые логи

# Стоимость недвижимости

- Объект — квартира в Москве
- Ответ — стоимость в рублях
- Регрессия

# Стоимость недвижимости — признаки

- Бинарные: наличие балкона, мусоропровода, лифта, охраны, парковки, и т.д.
- Категориальные: район города, тип дома (кирпичный/блочный/панельный/монолит), ближайшая станция метро и т.д.
- Вещественные: число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т.д.

# Стоимость недвижимости — особенности

- Выборка неоднородная, меняется со временем
- Разнотипные признаки
- Нужна интерпретируемая модель

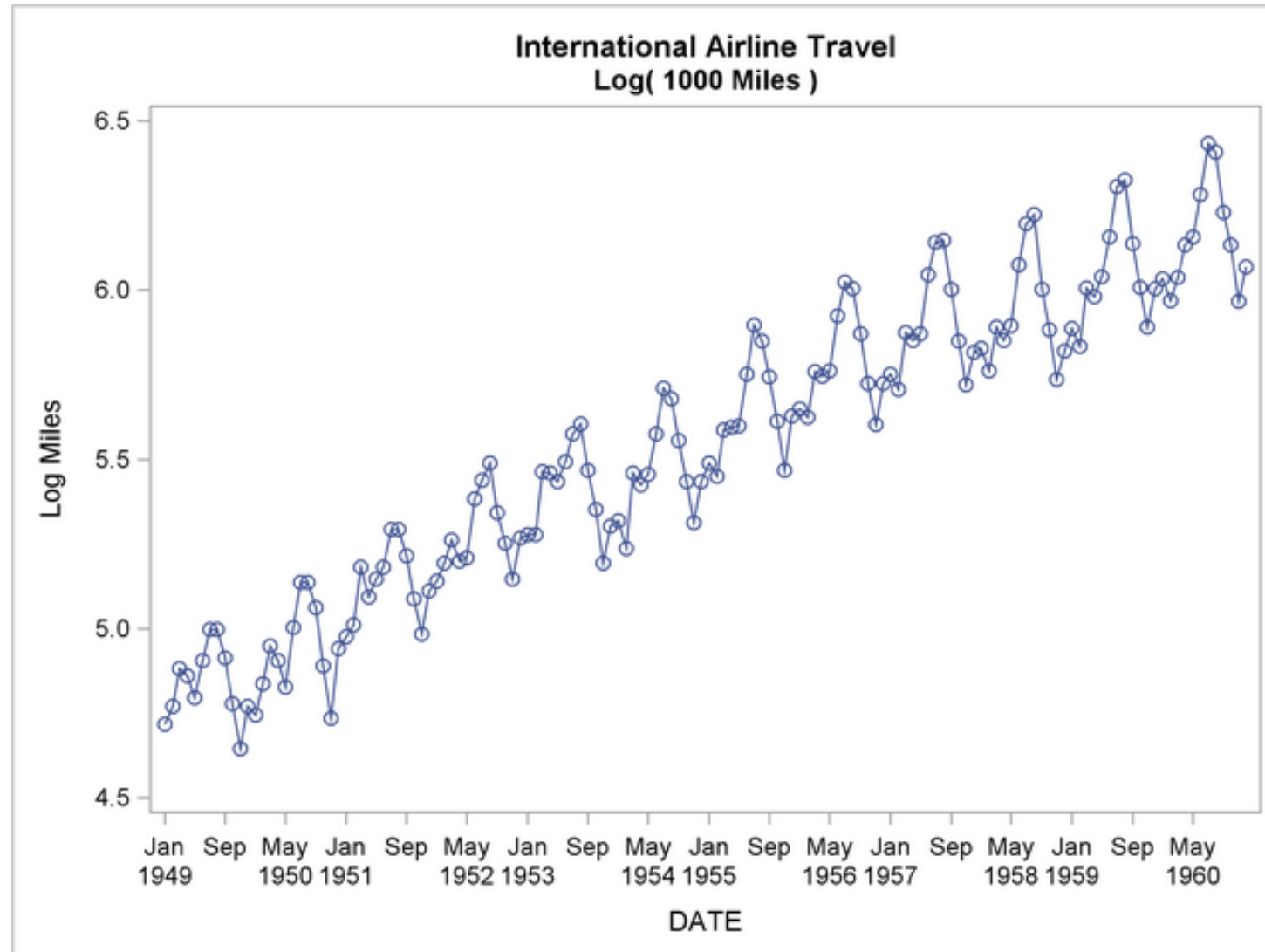
# Прогнозирование продаж

- Объект — тройка (товар, магазин, день)
- Ответ — объем продаж
- Регрессия
- Прогнозирование временных рядов

# Прогнозирование продаж — признаки

- Бинарные: выходной день, праздник, промоакция, и т.д.
- Вещественные: продажи в прошлые дни

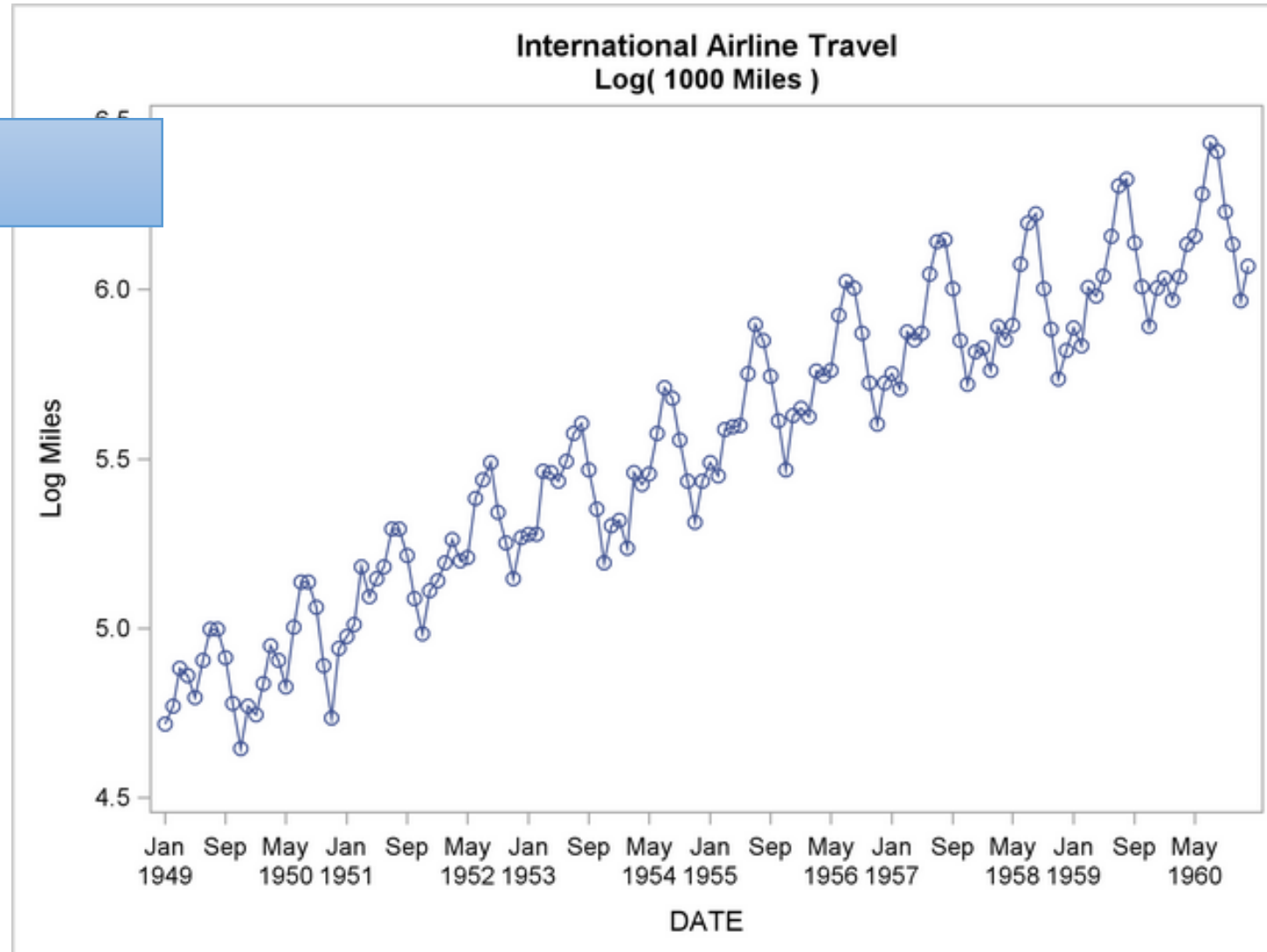
# Временные ряды





# Временные ряды

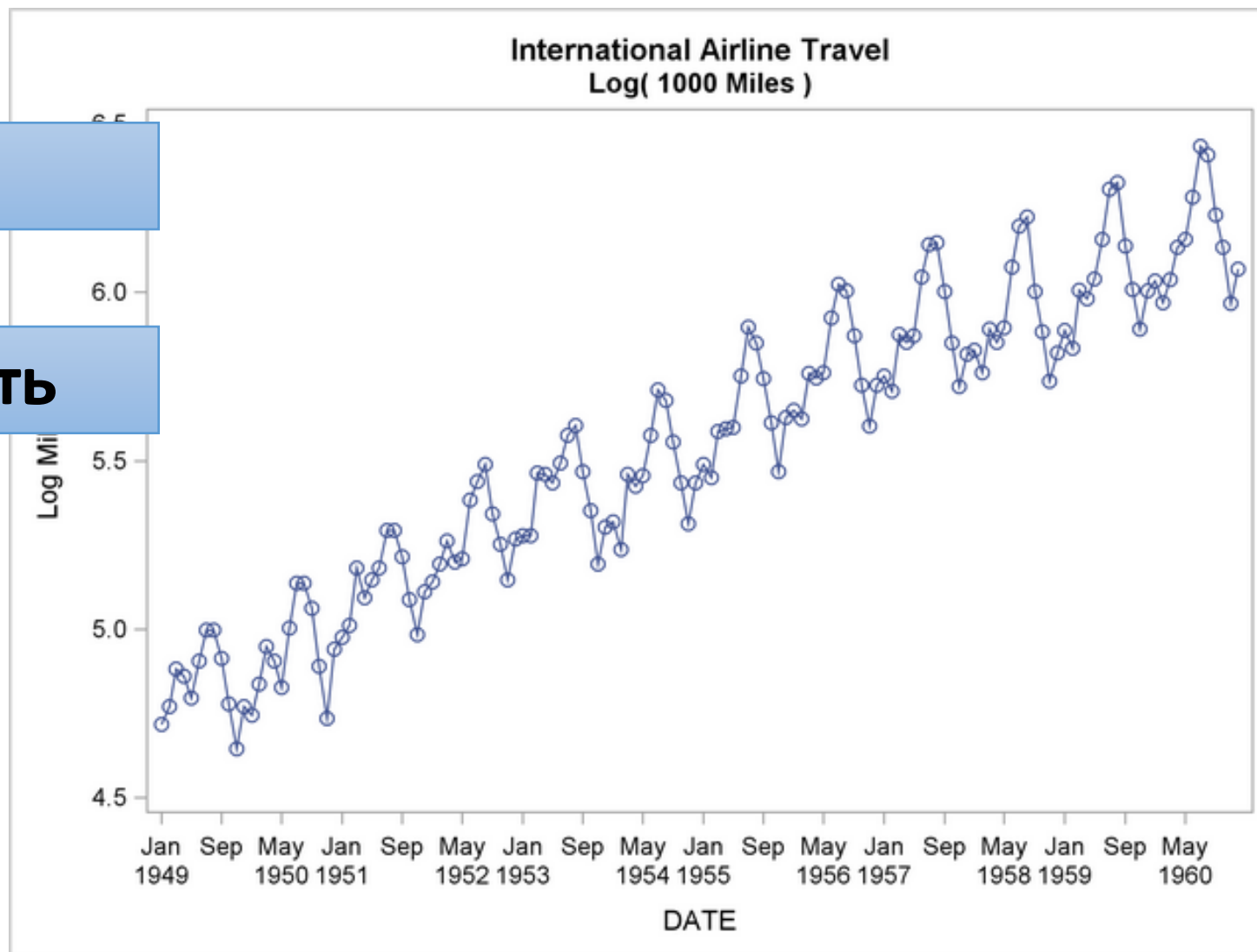
Тренд



# Временные ряды

Тренд

Сезонность



# Avito Context Ad Clicks Prediction

- [kaggle.com](https://www.kaggle.com)
- Объект — тройка (пользователь, запрос, баннер)
- Ответ — кликнет ли пользователь по баннеру
- Классификация

# Avito Context Ad Clicks Prediction — признаки

- Все действия пользователя на сайте
- Профиль пользователя (браузер, устройство, IP-адрес)
- История показов и кликов для других пользователей
- 10 таблиц с сырыми данными

# Avito Context Ad Clicks Prediction — особенности

- Надо изобретать признаки
- Сотни миллионов показов
- Размер подготовленной выборки — терабайты
- Нужны технологии и алгоритмы работы с большими данными

# Рекомендательная система фильмов

- Объект — пара (пользователь, фильм)
- Ответ — понравится ли пользователю фильм?
- Регрессия? Классификация?

# Рекомендательная система — признаки

- Оценки фильмов от пользователей
- Возможно, профиль пользователя
- Возможно, информация о фильме

# Рекомендательная система — Imhonet

## Оценки фильма Любопытное стечение обстоятельств

Een Bizarre Samenloop Van Omstandigheden, A Curious Conjunction of Coincidences

[Фильмы](#) / [Комедии](#) / [обстоятельств](#) /



**Да, Вам стоит смотреть фильм «Любопытное стечение обстоятельств»**

Людям, с оценками, похожими на [Ваши](#), этот фильм **нравится**

**А ещё они рекомендуют Вам [31 фильм](#)**

Ваша прогнозируемая оценка фильма после его просмотра

8.2

Смотрели? Оцените

[Не рекомендовать](#)

[Про фильм](#)

[Онлайн](#)

[Скачать](#)

[Отзывы](#)

[Персоны](#)

[Кадры](#)

**Оценки**

[Похожие](#)

### Распределение оценок



Всего оценок — 9

### Кому больше нравится





# Рекомендательная система — Amazon

## Frequently Bought Together



Price For All Three: **\$86.01**

 Add all three to Cart

 Add all three to Wish List

[Show availability and shipping details](#)

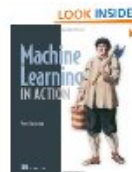
- ☒ **This item:** Machine Learning for Hackers by Drew Conway Paperback **\$33.87**
- ☒ Machine Learning in Action by Peter Harrington Paperback **\$25.75**
- ☒ Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback **\$26.39**

## Customers Who Bought This Item Also Bought

Page 1 of 17



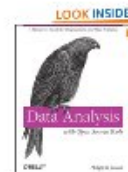
Programming Collective Intelligence: Building ...  
› Toby Segaran  
★★★★☆ (84)  
Paperback  
**\$26.39**



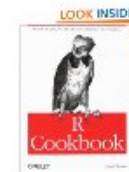
Machine Learning in Action  
› Peter Harrington  
★★★★☆ (10)  
Paperback  
**\$25.75**



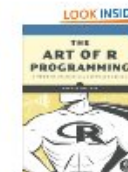
Mining the Social Web: Analyzing Data from ...  
› Matthew A. Russell  
★★★★☆ (19)  
Paperback  
**\$26.36**



Data Analysis with Open Source Tools  
› Philipp K. Janert  
★★★★☆ (29)  
Paperback  
**\$24.05**



R Cookbook (O'Reilly Cookbooks)  
› Paul Teetor  
★★★★☆ (18)  
Paperback  
**\$32.43**



The Art of R Programming: A Tour of Statistical ...  
Norman Matloff  
★★★★☆ (29)  
Paperback  
**\$25.06**

Are any of these items inappropriate for this page? [Let us know](#)

# Рекомендательная система — особенности

- Много метрик для оптимизации: число кликов по рекомендациям, число новых для пользователя товаров, разнообразие предлагаемых жанров, и т.д.
- Особый вид данных: (пользователь, фильм/товар, оценка)
- Получение оценки — явный и неявный отклик

# Резюме

- Много типов признаков — у всех свои особенности
- Много постановок задач — у всех свои особенности
- Много особенностей у конкретных прикладных задач

# На следующей лекции

- Линейная алгебра
- Векторы и матрицы
- Системы линейных уравнений