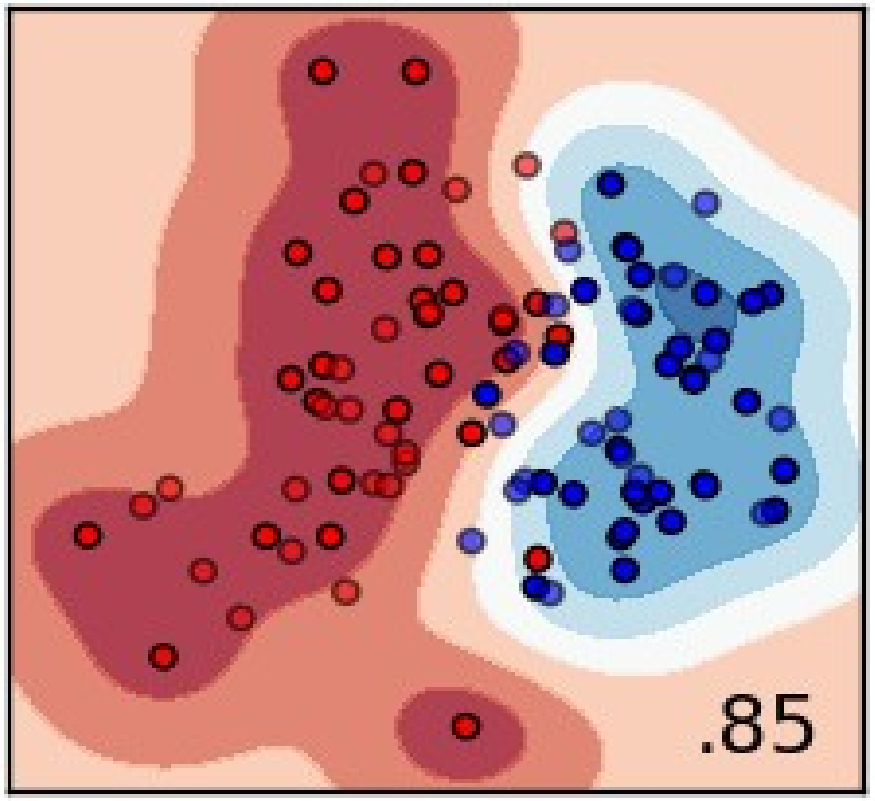
Лекция 2.

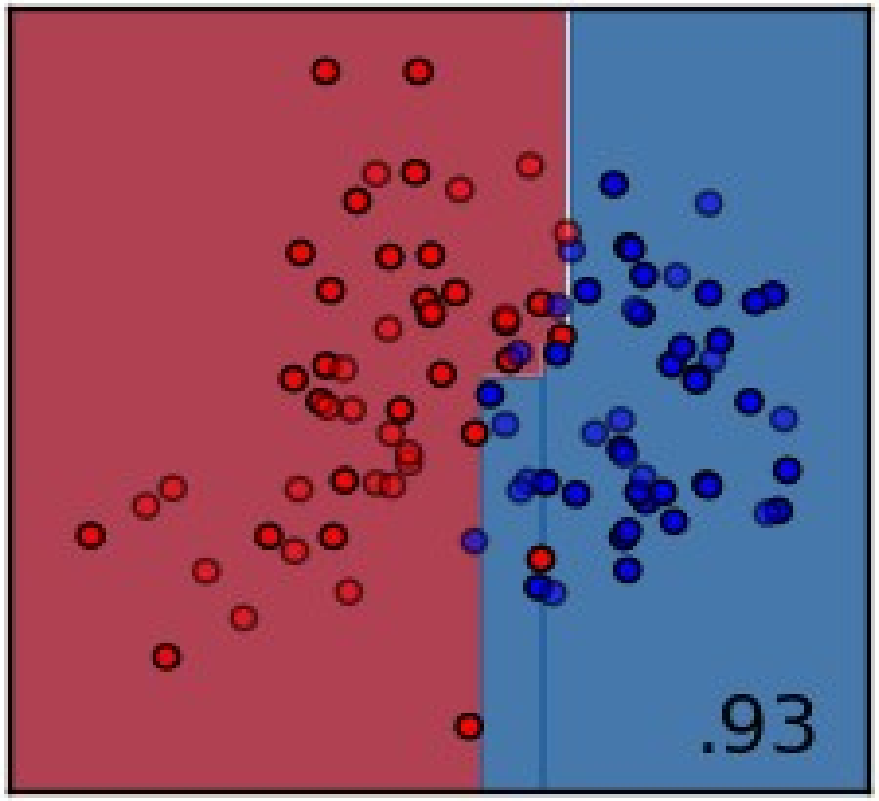
Задачи ML.
Основные понятия ML

Павел Владимирович
Слипенчук
PavelMSTU@stego.su
ИУ-8

Москва
09.09.2016

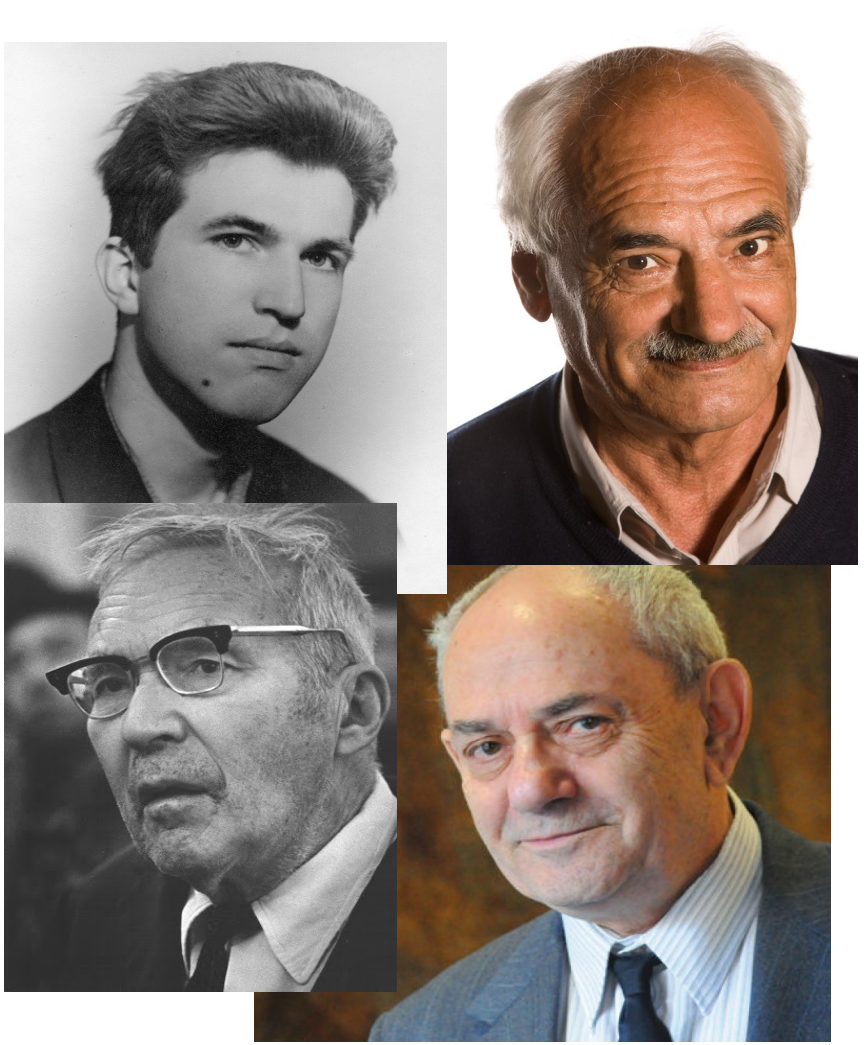
Классификация
(Classification)





Классификация
(Classification)

- Например: мужчины – женщины



М | Ж

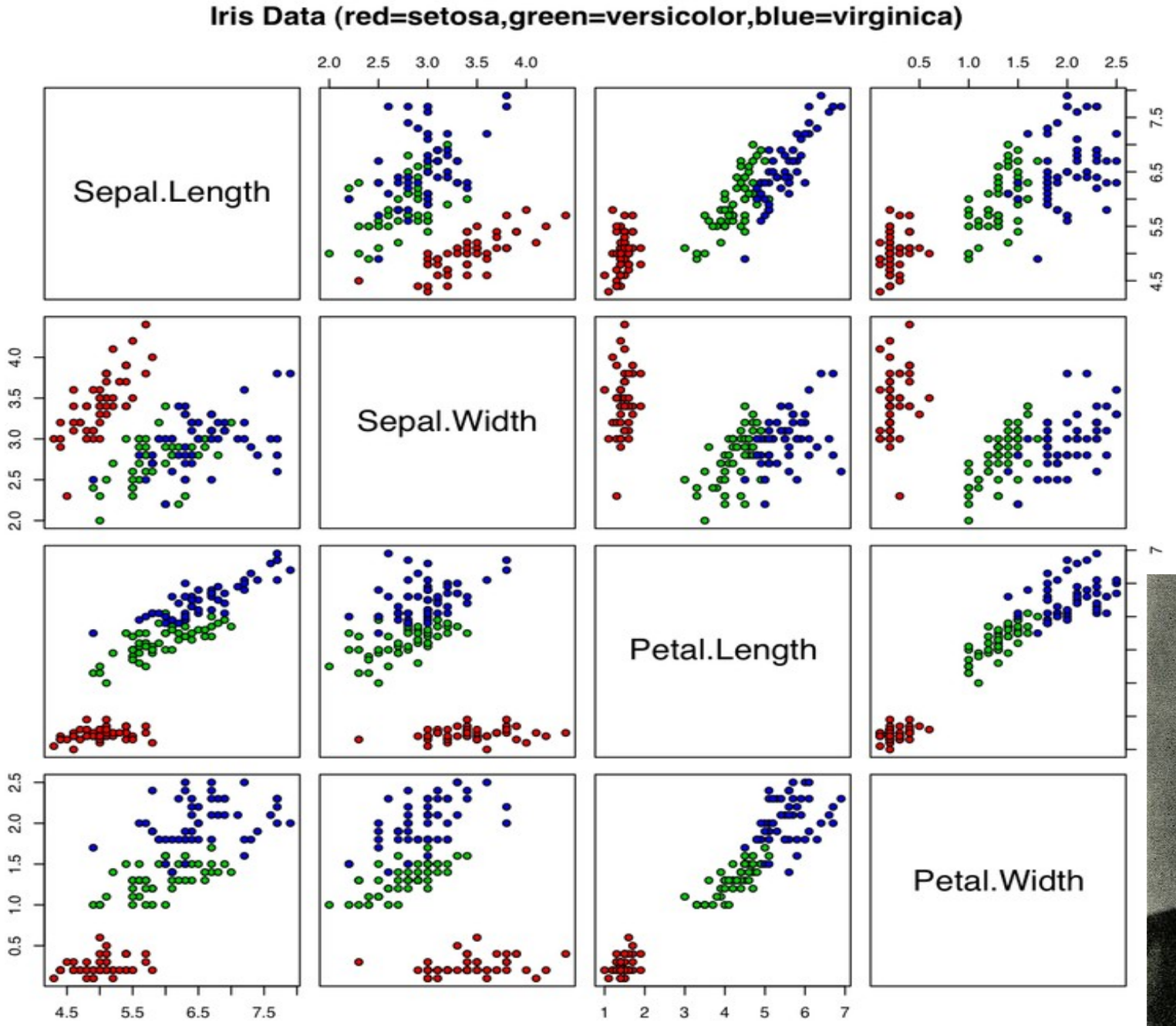
=3=

Классификация
(Classification)

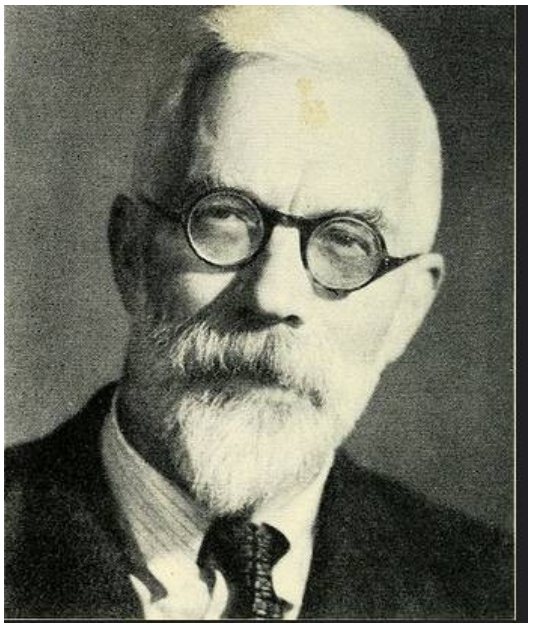
- **Классы**
- **Выборка, обучающая выборка, контрольная выборка.**
- **Характеристики, признаки, контрибьютеры**
- **Алгоритм классификации**
- **Подгонка** (fitting)
- **Проверка** (scoring)

Классификация
(Classification)

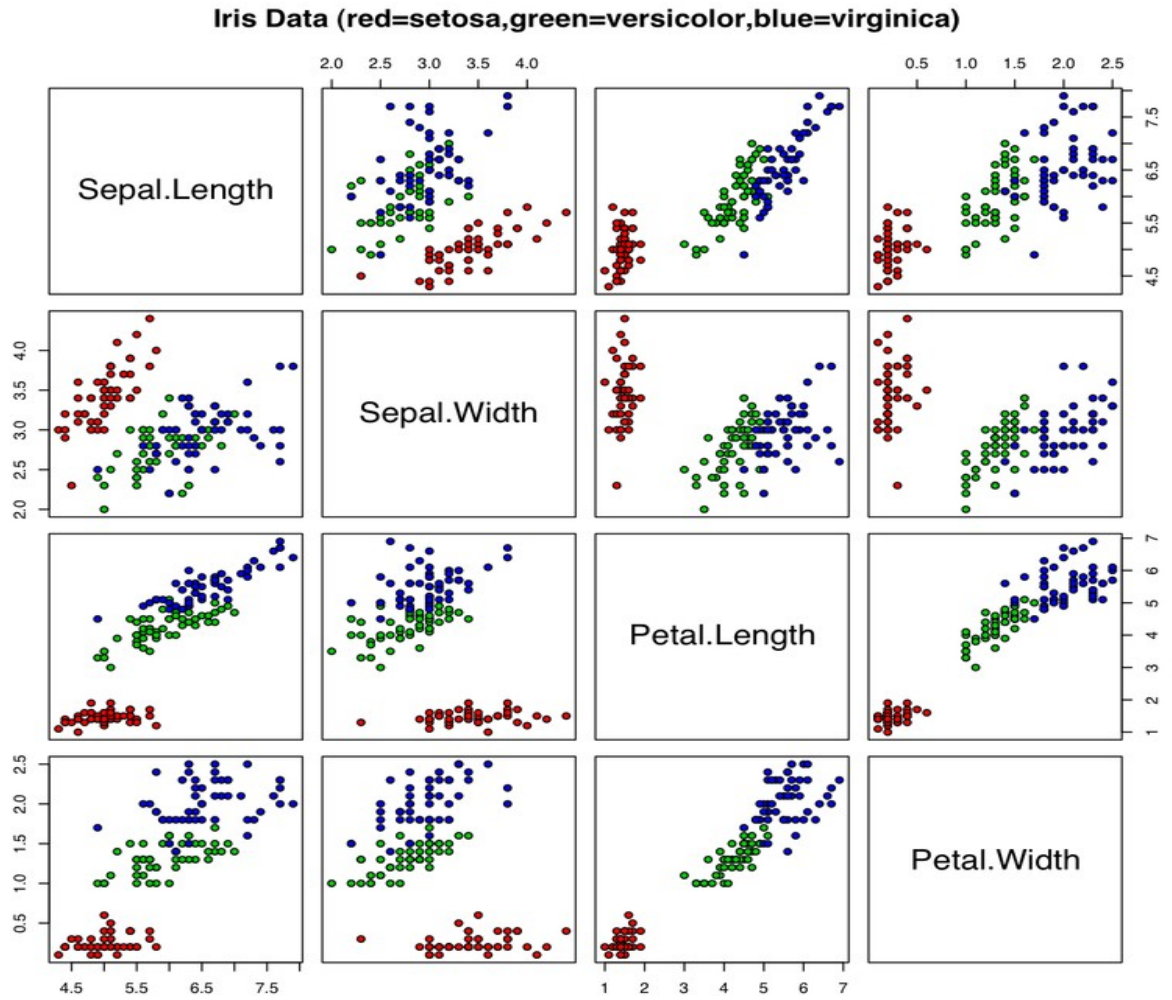
- **Выброс** – нестандартные данные, которые могут помешать классификации



“Ирисы Фишера” (1936)



Рональд Эйлмер Фишер (1890 – 1962)



- **Классы**
- **Выборка**
- **Признаки**
- **Алгоритм классификации (?)**
- **Подгонка** (fitting) **(?)**
- **Проверка** (scoring) **(?)**

Характеристика, признак, контрибьютор

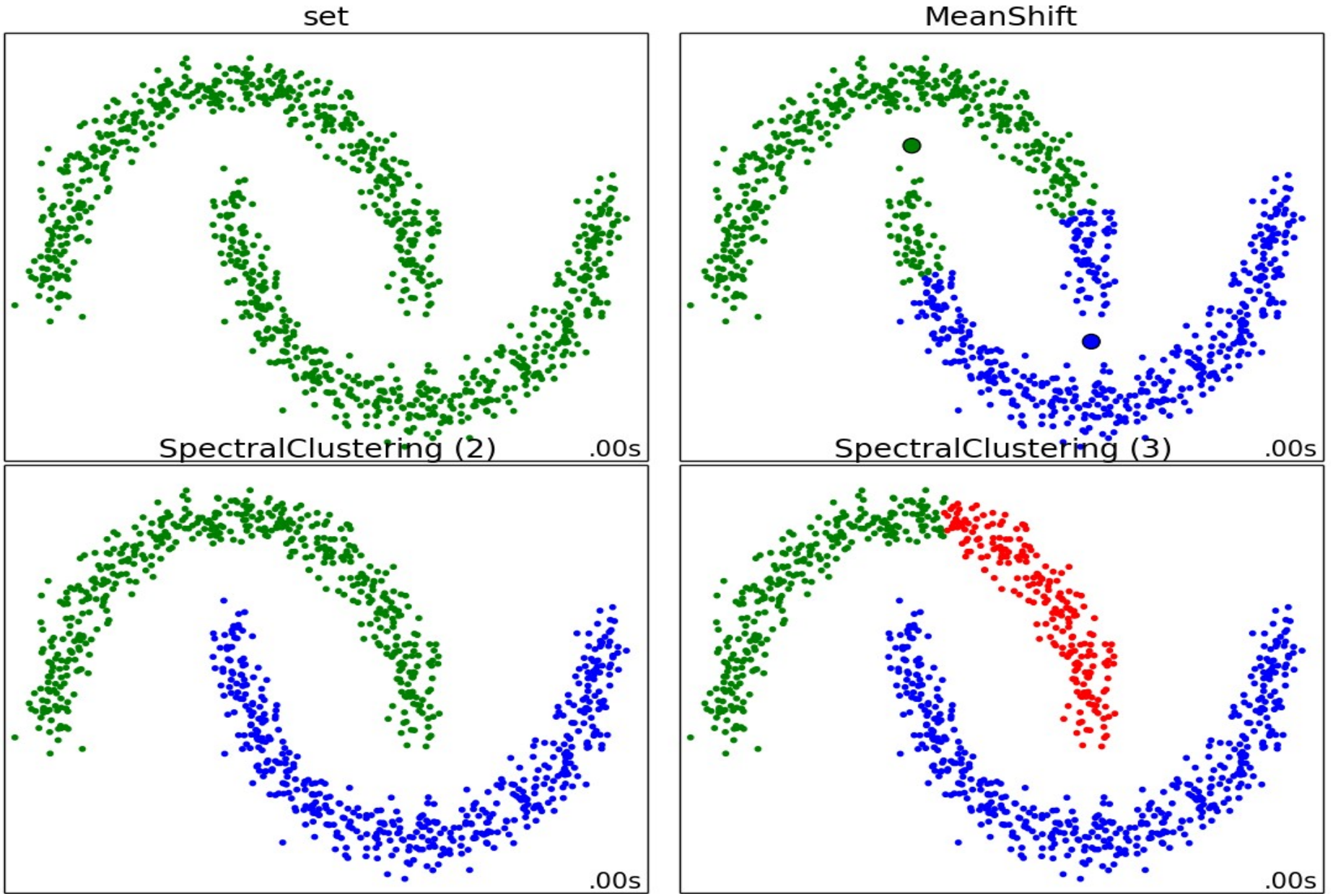
- **Характеристика** – какое либо свойство ***любой*** строгости (математической) в описании.
- **Признак** – строгий мат.объект.
- **Контрибьютер** – совокупность признаков (возможно один) вносящий определенный вклад в априорную вероятность определения класса.

Признаки

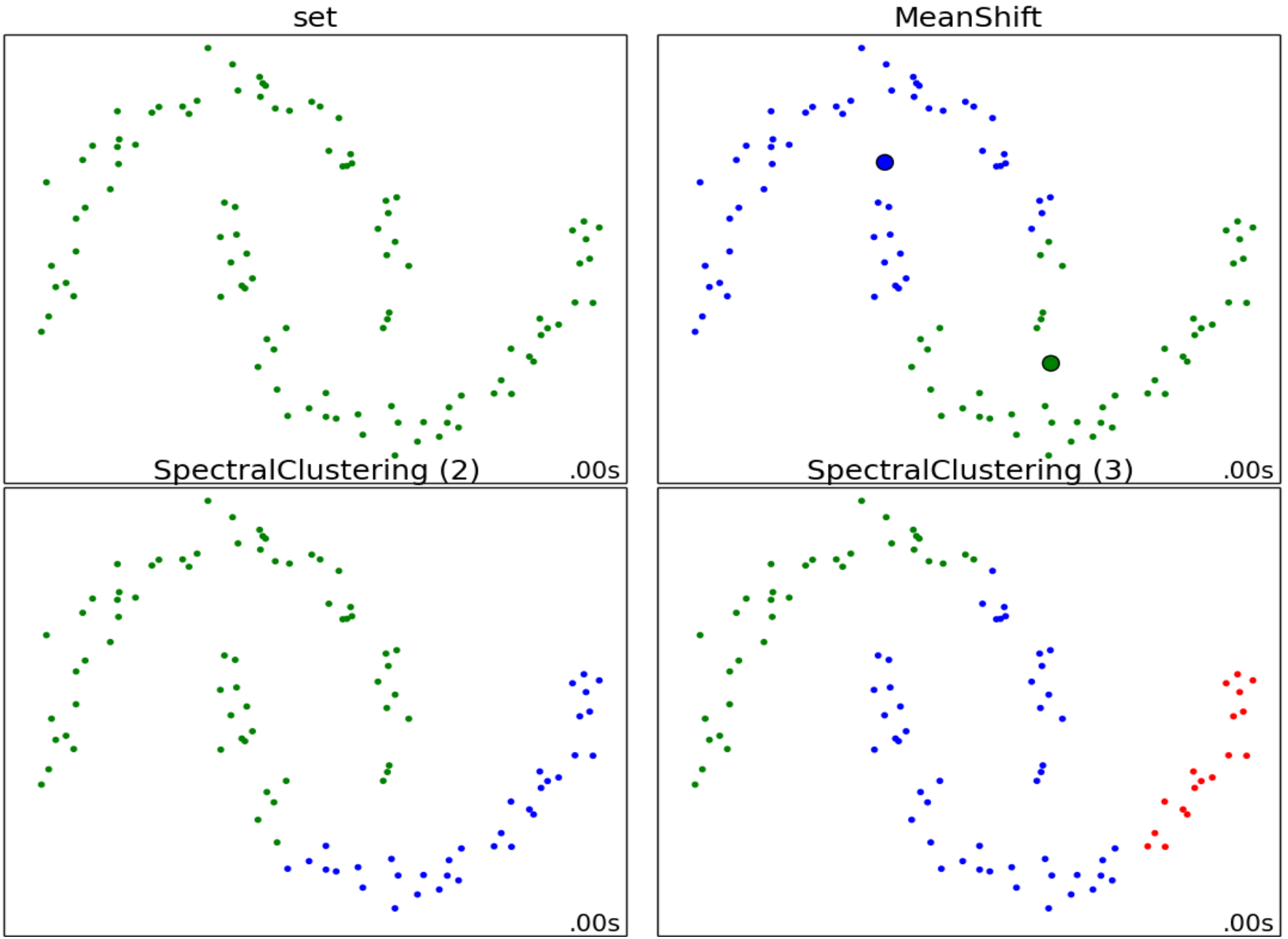
В 99% случаях используют:

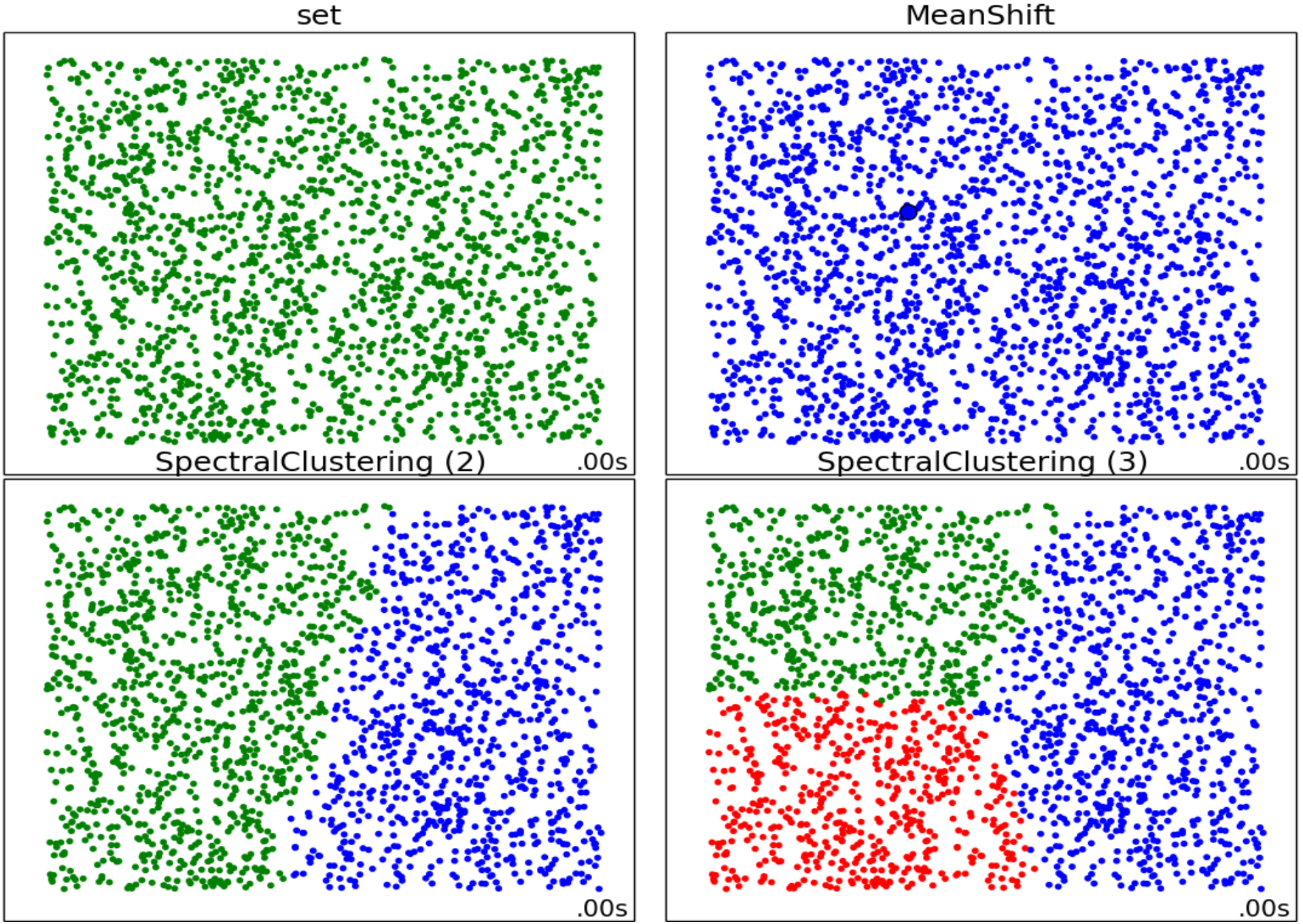
- **Сравнимые** – можно задать ***бинарное отношение***.
- **Несравнимые** = категориальные.
- Сравнимые бывают ***непрерывные*** и ***дискретные***.
- **Сложные (составной)** – это f(x1, …, xn), где xi – другой признак
- **Бинарный** признак – это дискретный признак, состоящий из двух значений: 0, 1 или -1, 1.

Кластеризация
(Clustering)

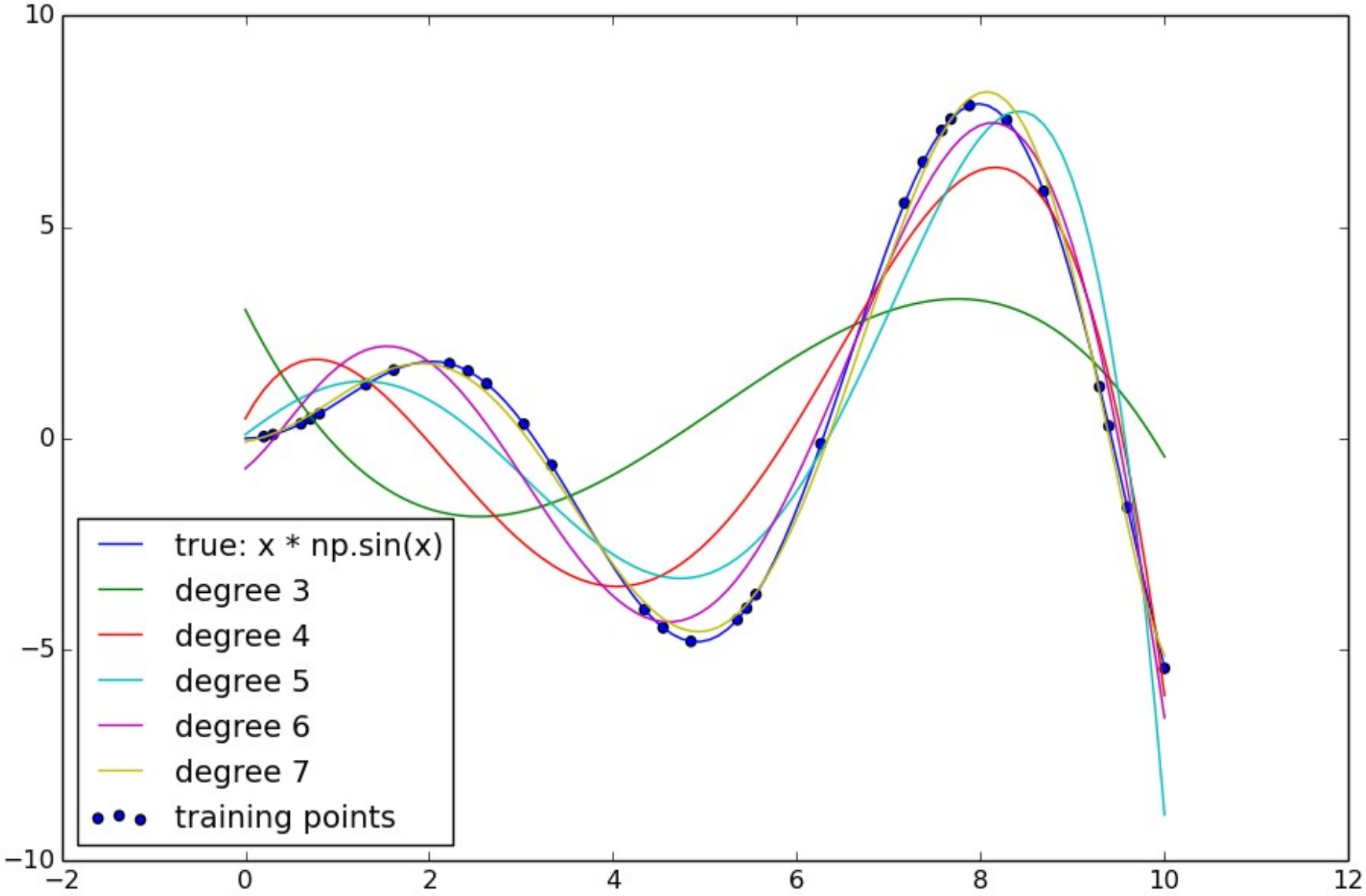


Объем выборки: 100 точек




set
MeanShift
SpectralClustering (2)
.00s
SpectralClustering (3)
.00s
.00s
.00s

Регрессия
(Regression)



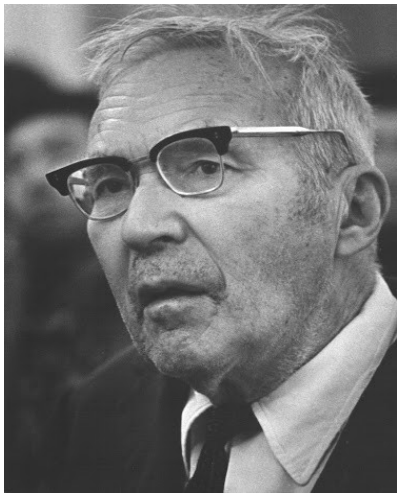
Who is who?

Регрессия
Regression
[rɪˈgrɛʃ(ə)n]

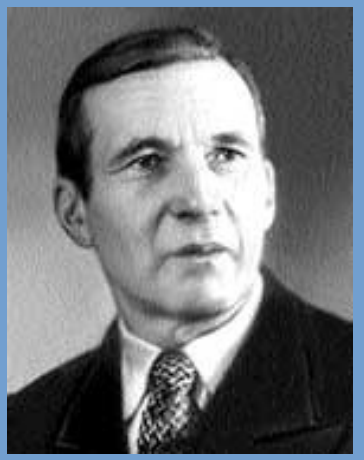
Интерполяция
Interpolation
[ɪnˌtɜːpəˈleɪʃ(ə)n]

Аппроксимация
Approximation
[əˌprɒksɪˈmeɪʃ(ə)n]

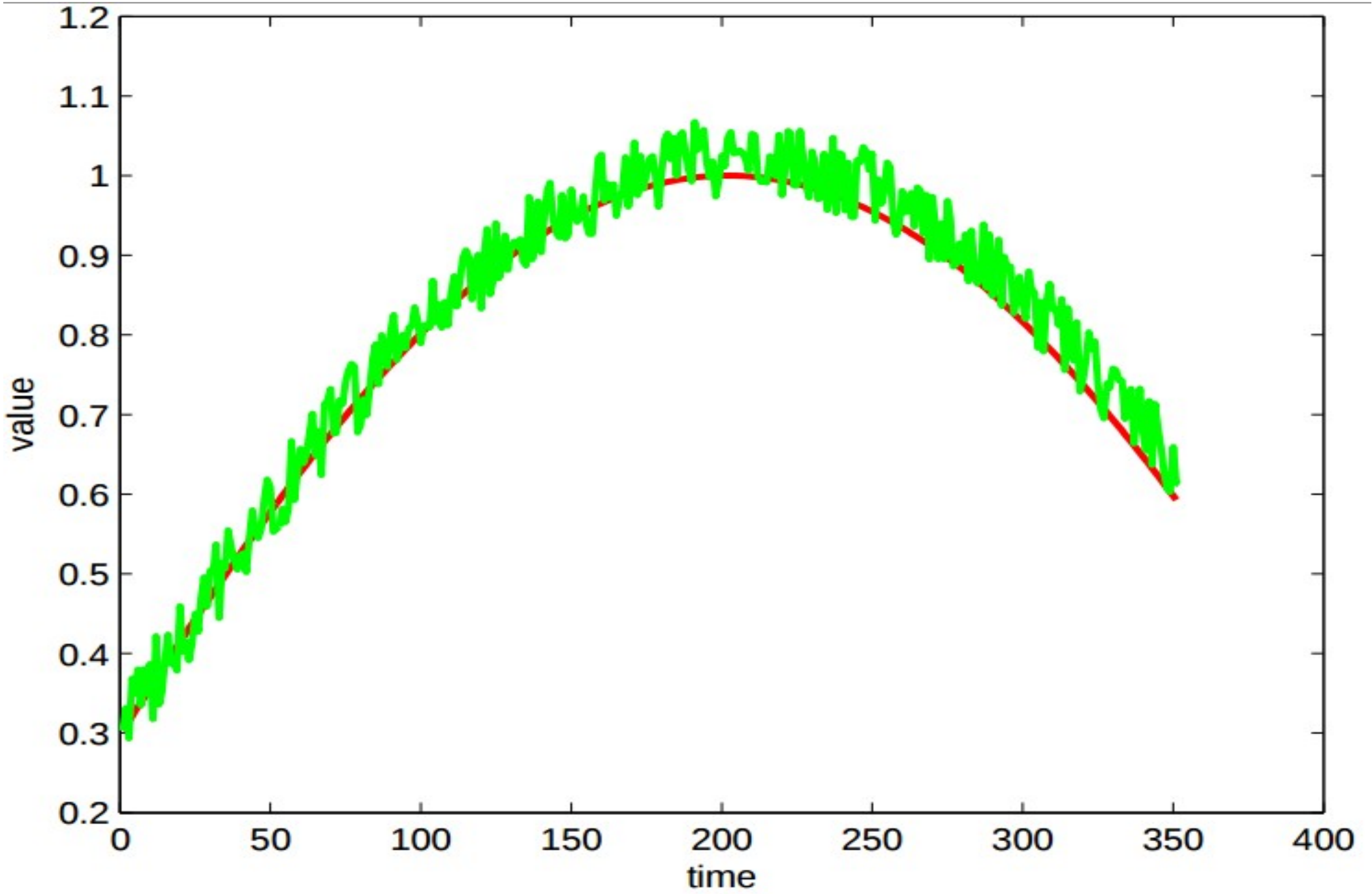
Идентификация



Объект



Прогнозирование



Прогнозирование

~~Предсказание~~

–5 очков

~~Предвидение~~

–20 очков

~~Пророчество~~

–100 очков

Data Mining

- ~~*Добыча Данных*~~
- ~~*Интеллектуальный Анализ Данных*~~
- ~~*Глубинный Анализ Данных*~~

Business Intelligence

- ~~*Бизнес Интеллигенция*~~
- ~~*Бизнес Разведка*~~
- ~~*Бизнес Интелект*~~

Резюме

- **Задачи**: классификация, кластеризация, регрессия, идентификация, интерполяция(~), аппроксимация(~)
- **Понятия**: классы, выборка, признаки, подгонка (обучение), проверка, выброс

Рекомендуемые источники информации.

- Курс ШАД'а Яндекса
- Червоненкис Алексей Яковлевич. Компьютерный анализ данных
- Вики Константина Вячеславовича Воронцова: http://www.machinelearning.ru/
- Ветров Дмитрий Петрович
- Свой проект (курсовая? Диплом?)
- **Здравый смысл**

Здравый смысл?

Пакеты Python

- http://matplotlib.org/
- http://scikit-learn.org/
- Pandas
- Numpy
- Skipy
- ... …. ...