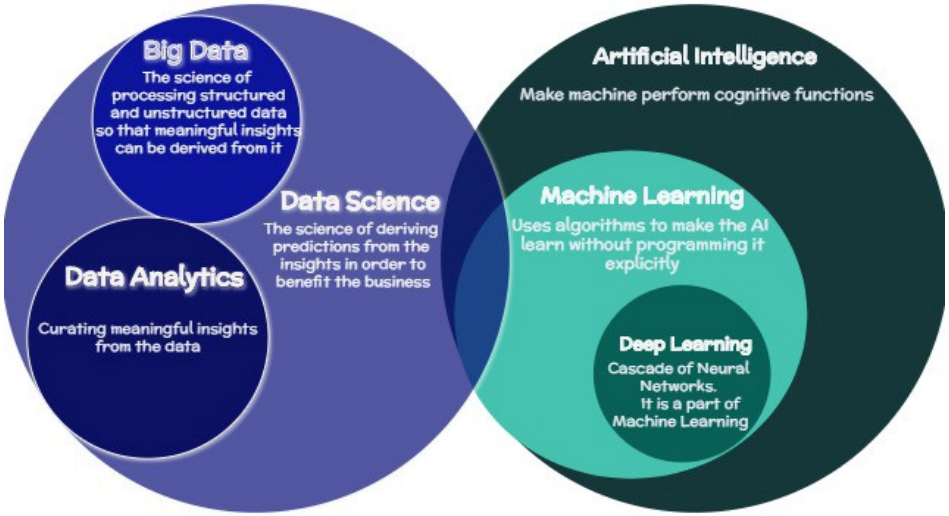
ЛЕКЦИЯ 1 - Data Science
и смежные с ней дисциплины

12 декабря 2022 г.

Основные дисциплины



Основные дисциплины

"The brain is the most complex object in the known universe."

C. Koch, chief scientific officer of the Allen Institute for Brain Science, 2013

- Data Science — выявление скрытых и существенных закономерностей в данных

Основные дисциплины

"The brain is the most complex object in the known universe."

C. Koch, chief scientific officer of the Allen Institute for Brain Science, 2013

- Data Science — выявление скрытых и существенных закономерностей в данных
- Искусственный Интеллект — создание вычислительных систем по образу и подобию биологического мозга

Основные дисциплины

"The brain is the most complex object in the known universe."

C. Koch, chief scientific officer of the Allen Institute for Brain Science, 2013

- Data Science — выявление скрытых и существенных закономерностей в данных
- Искусственный Интеллект — создание вычислительных систем по образу и подобию биологического мозга
- Машинное обучение

Основные дисциплины

"The brain is the most complex object in the known universe."

C. Koch, chief scientific officer of the Allen Institute for Brain Science, 2013

- Data Science — выявление скрытых и существенных закономерностей в данных
- Искусственный Интеллект — создание вычислительных систем по образу и подобию биологического мозга
- Машинное обучение
- *Глубинное обучение*

Data Science explained

История 1

Рис.: Горящий business jet

Data Science explained

История 2

Будущая юная мама

Data Science explained

История 3

Рис.: Мировой закусон!

Коэффициент Пирсона

$$\rho_{xy} = \frac{COV(x, y)}{\sigma_x \sigma_y},$$

где

Коэффициент Пирсона

$$\rho_{xy} = \frac{COV(x, y)}{\sigma_x \sigma_y},$$

где

- $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ – стандартное отклонение данных x

Коэффициент Пирсона

$$\rho_{xy} = \frac{COV(x, y)}{\sigma_x \sigma_y},$$

где

- $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ – стандартное отклонение данных x
- $\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$ – стандартное отклонение данных y

Коэффициент Пирсона

$$\rho_{xy} = \frac{COV(x,y)}{\sigma_x \sigma_y},$$

где

- $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ – стандартное отклонение данных x
- $\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$ – стандартное отклонение данных y
- $cov(x,y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j>i}^{n}(x_i - x_j)(y_i - y_j)$ – ковариация между данными x,y

Data Science explained

Inferential versus Descriptive Statistics...

Data Science explained

Inferential versus Descriptive Statistics...
STATISTICS versus STATISTIC...

Data Science explained

Наиболее популярные статистики

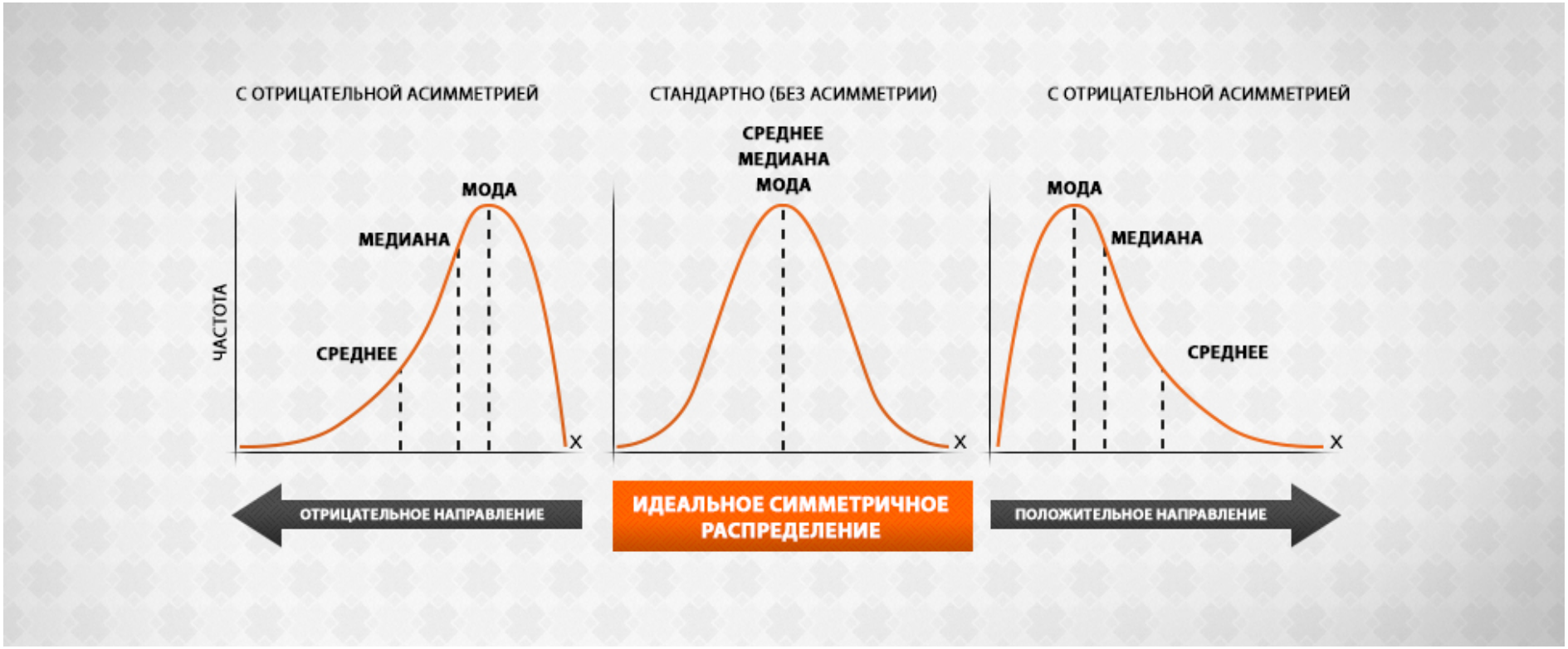
- Среднее арифметическое
- Медиана
- Мода
- Среднее геометрическое
- Взвешенное среднее
- Размах (интервал изменения)
- Дисперсия

Data Science explained

Наиболее популярные статистики

- Среднее арифметическое
- Медиана
- Мода
- Среднее геометрическое
- Взвешенное среднее
- Размах (интервал изменения)
- Дисперсия

Data Science explained



Data Science explained

Inferential Statisitics

Выносит суждения о популяции с помощью выборок

Data Science explained

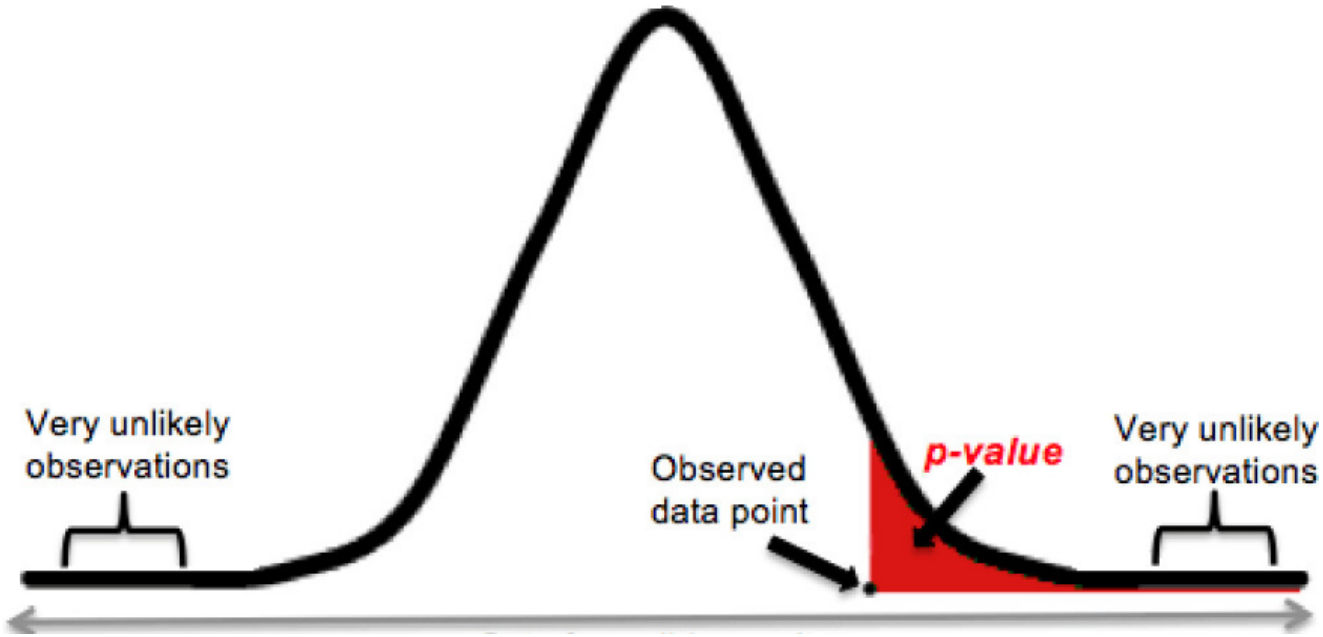
Inferential Statisitics

Выносит суждения о популяции с помощью выборок

Проблема:
Sampling variability / Вариабельность выборок.

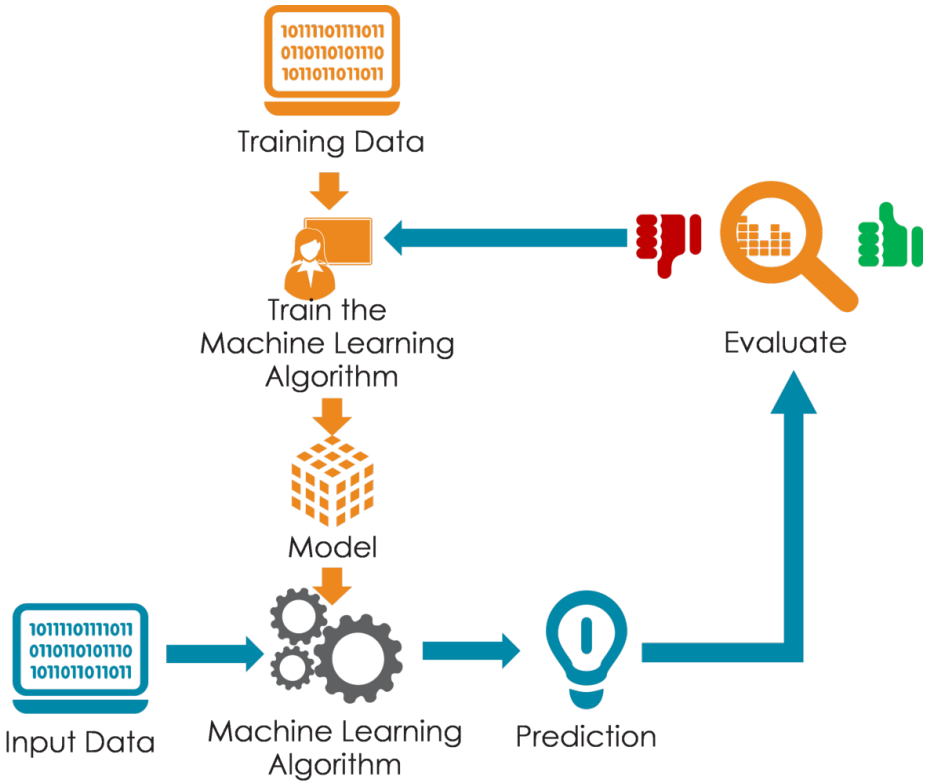
Data Science explained

Inferential Statisitics



Машинное обучение

Общая схема

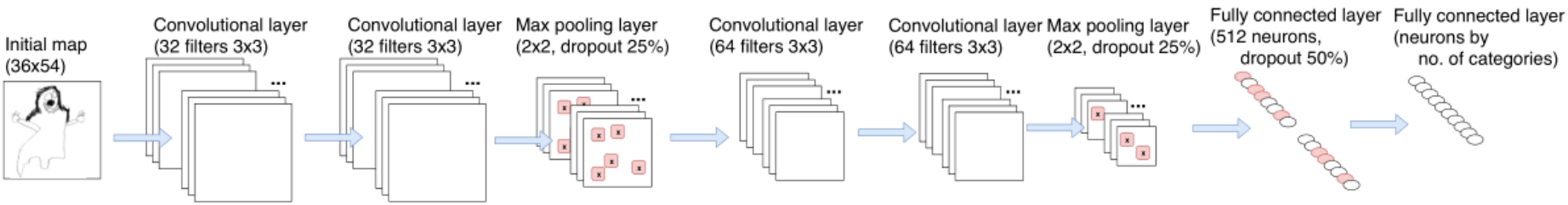


Машинное обучение explained

Красная шапочка - 2022

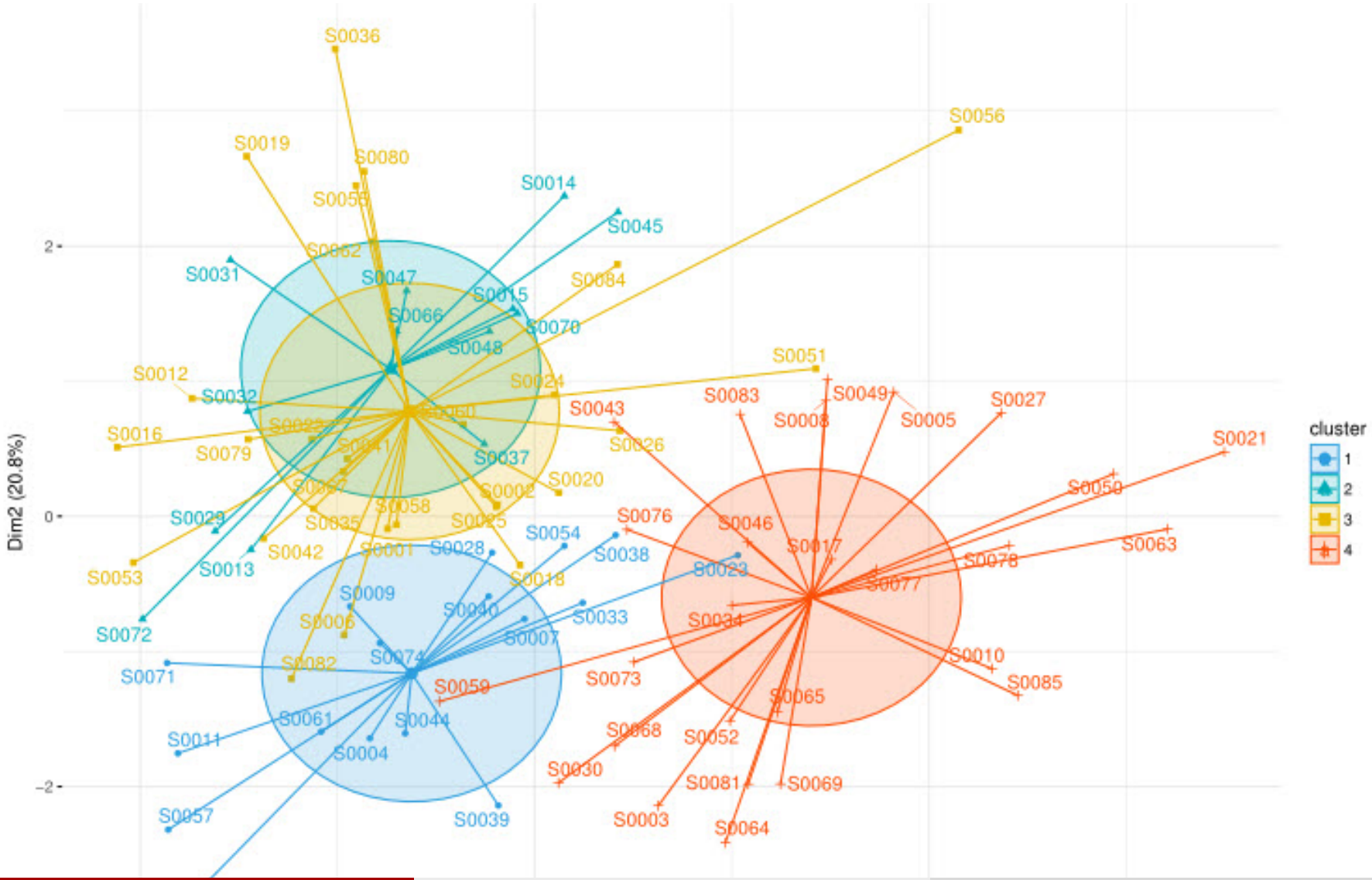
Глубинное обучение

Каскадные, "фрактальные"процессы



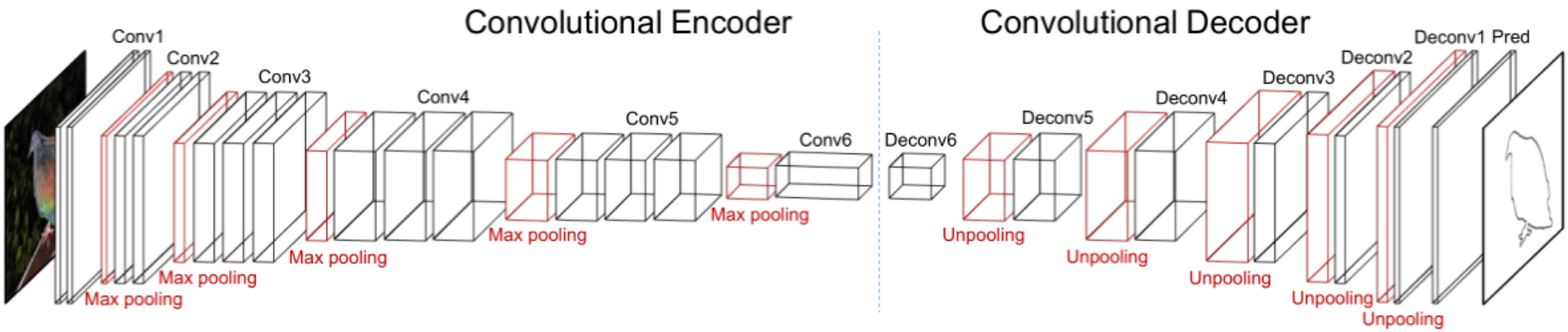
Обучение без учителя

Кластеризация

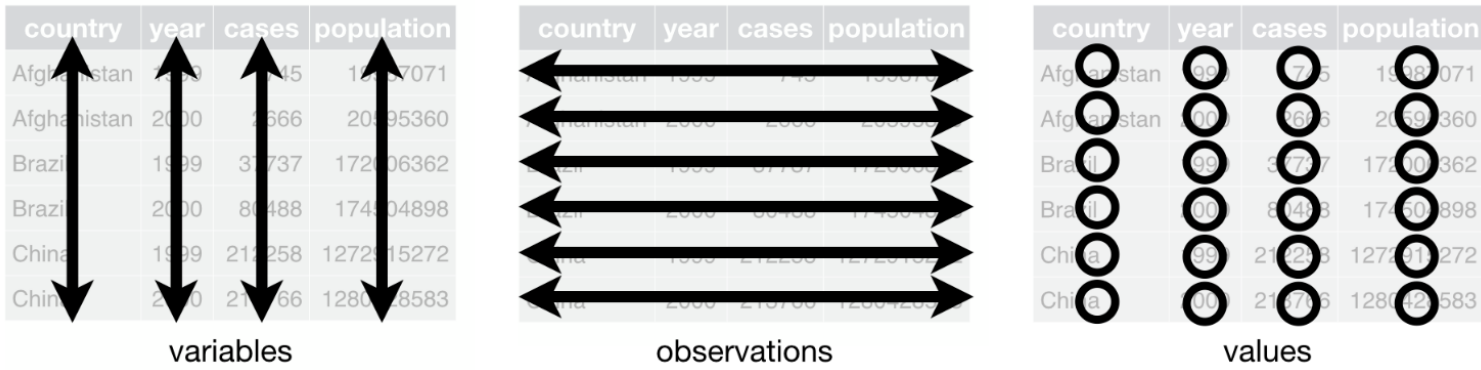


Обучение без учителя

Encoder – Decoder

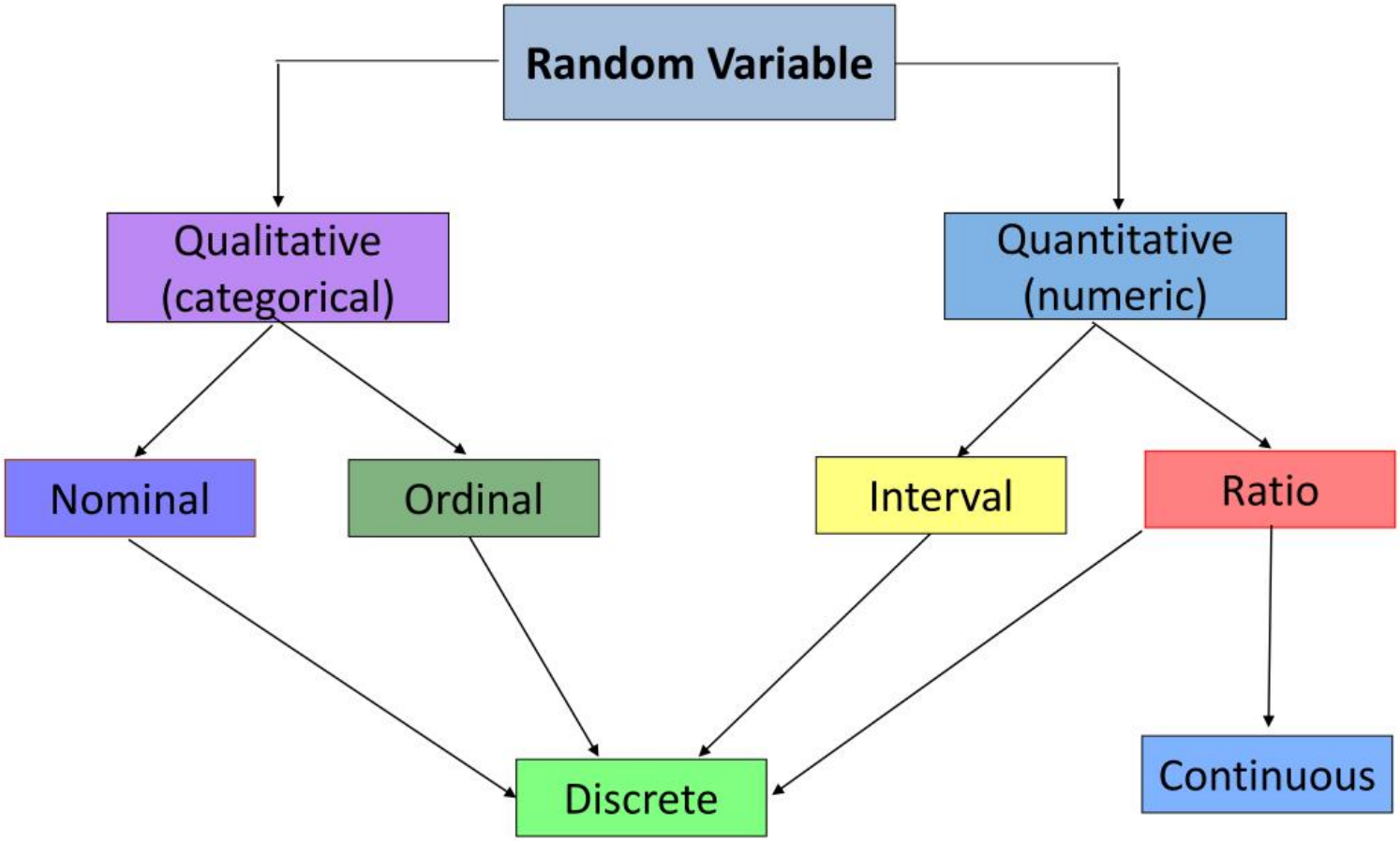


Данные


country
year
cases
population
variables
observations
values

Данные

Разновидности



Данные

Разновидности

- Quantitative/Количественные — значения основаны на объективных измерениях.

Данные

Разновидности

- Quantitative/Количественные — значения основаны на объективных измерениях.
- Qualiitative/Качественные, категориальные — относятся к объективным (мужчина.женщина; цвет) или субъективно положенным категориям (в восторге/нравится/нормально/так себе/отвратительно).

Данные

Разновидности

- Quantitative/Количественные — значения основаны на объективных измерениях.
- Qualiitative/Качественные, категориальные — относятся к объективным (мужчина.женщина; цвет) или субъективно положенным категориям (в восторге/нравится/нормально/так себе/отвратительно).
- Ordinal/Упорядоченные — ‘разница’ между классами субъективна.

Данные

Разновидности

- Quantitative/Количественные — значения основаны на объективных измерениях.
- Qualiitative/Качественные, категориальные — относятся к объективным (мужчина.женщина; цвет) или субъективно положенным категориям (в восторге/нравится/нормально/так себе/отвратительно).
- Ordinal/Упорядоченные — ‘разница’ между классами субъективна.
- Дискретные — определены с абсолютной точностью, как правило, целые числа (чмсло членов семьи).

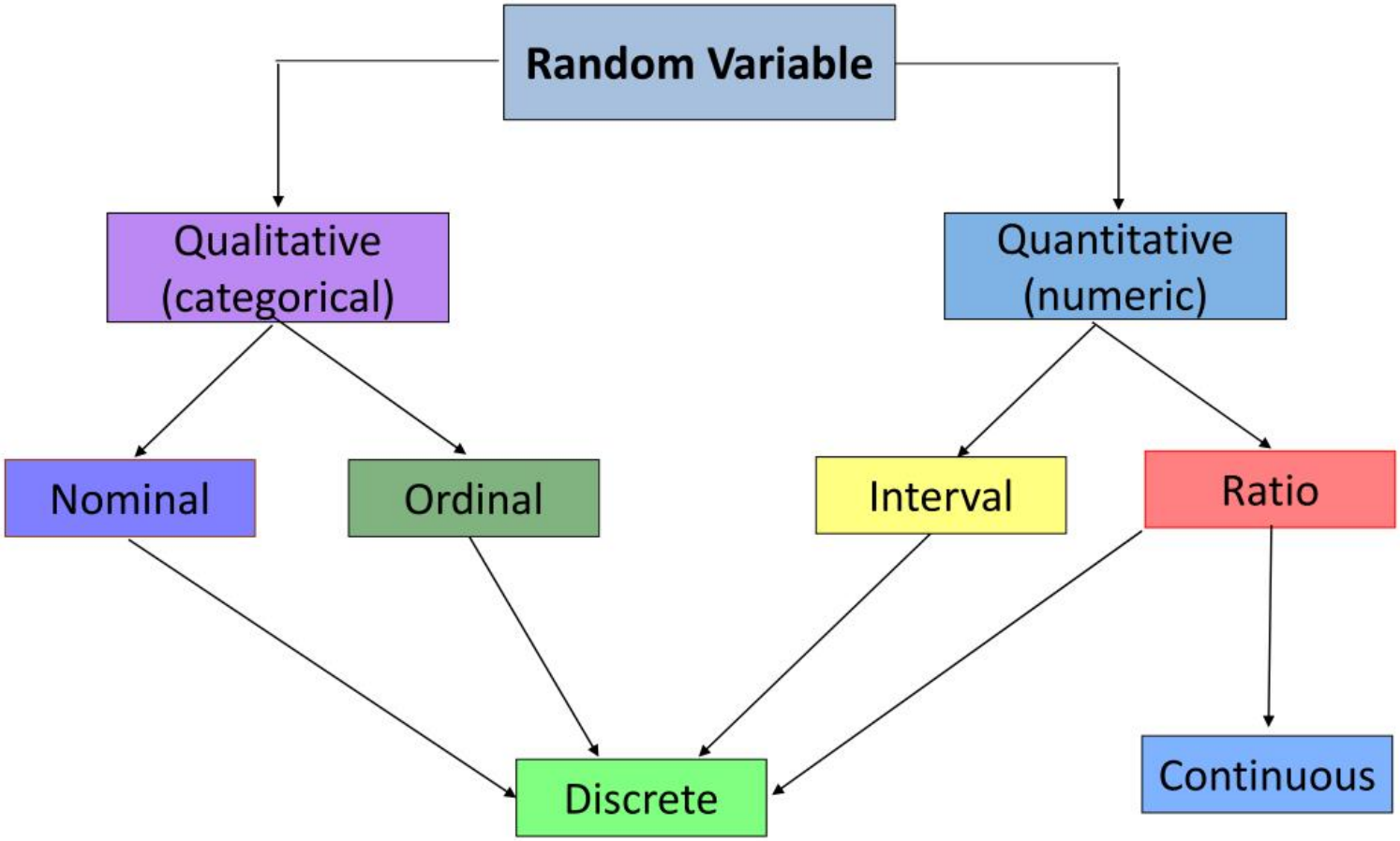
Данные

Разновидности

- Quantitative/Количественные — значения основаны на объективных измерениях.
- Qualiitative/Качественные, категориальные — относятся к объективным (мужчина.женщина; цвет) или субъективно положенным категориям (в восторге/нравится/нормально/так себе/отвратительно).
- Ordinal/Упорядоченные — ‘разница’ между классами субъективна.
- Дискретные — определены с абсолютной точностью, как правило, целые числа (чмсло членов семьи).
- Continuous/Непрерывные — как правило, действительные числа, конечная точность (рост человека может выражаться в метрах, сантиметрах, миллиметрах...)

Данные

Разновидности



О непрерывных данных

Градусы и градусы

- Непрерывные - интервальные — не имеют значимого нуля, *интервал* (разность) имеет смысл, *отношение* - нет.

О непрерывных данных

Градусы и градусы

- Непрерывные - интервальные — не имеют значимого нуля, *интервал* (разность) имеет смысл, *отношение* - нет.
- Непрерывные - ratio — имеют значимый нуль, имеет смысл и *интервал* (разность), и *отношение*.

Данные

Разновидности

Is TIME — interval or ratio data???

Примеры задач обучения

Задачи классификации

- Задача медицинской диагностики
- Задача оценивания заемщиков
- Задача предсказания ухода клиентов

Задачи восстановления регрессии

- Задача прогнозирования потребительского спроса
- Задача предсказания рейтингов
- Задача аппроксимации функции

Кодировка данных в таблицах:
не дайте ввести себя в заблуждение!