

# Машинное обучение в гидрометеорологии

## Лекция №2. Технические средства анализа данных.

Михаил Иванович Варенцов ([mikhail.varentsov@srcc.msu.ru](mailto:mikhail.varentsov@srcc.msu.ru))

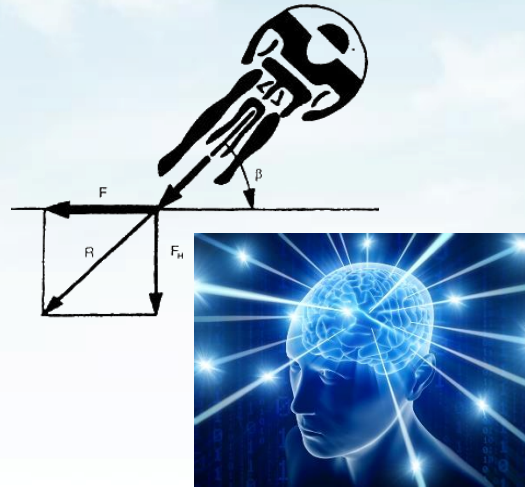
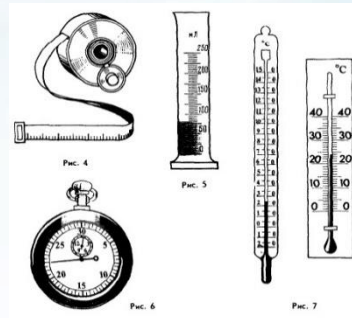
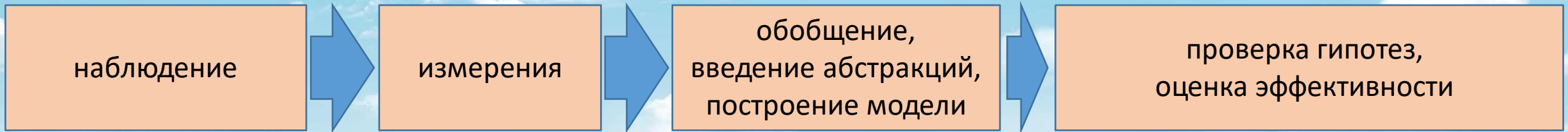
Михаил Алексеевич Криницкий ([krinitsky@sail.msk.ru](mailto:krinitsky@sail.msk.ru))

[ml4hydromet@ml4es.ru](mailto:ml4hydromet@ml4es.ru)

**ИНТЕЛЛЕКТ**  
ФОНД РАЗВИТИЯ НАУКИ И ОБРАЗОВАНИЯ

# Ранее в ML4hydromet...

## КАК проводятся физические исследования?

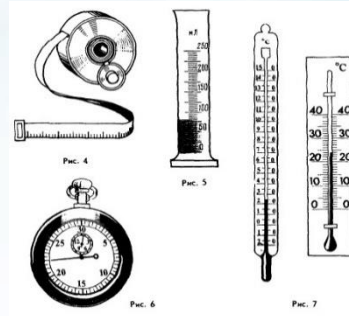
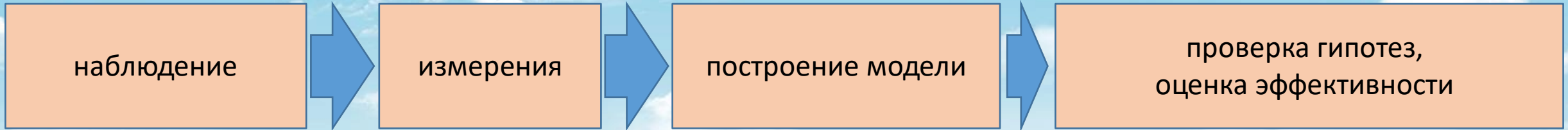


Настоящая наука начинается с тех пор, как начинают измерять.  
Точная наука немыслима без меры.  
Д.И. Менделеев

# Ранее в ML4hydromet...

Когда (человеку) непонятно, что происходит


все равно строим модель










~~обобщение ?  
введение абстракций ?~~






# Репозиторий курса

 **ML4hydromet-2024** Public


 Pin  Unwatch 1  Fork 0  Star 0


 main  1 Branch  0 Tags


 Add file  Code


 About


No description, website, or topics provided.

 Readme

 Activity

 0 stars

 1 watching

 0 forks

**Releases**

No releases published


[Create a new release](#)


**Packages**



No packages published













[Publish your first package](#)




**Contributors** 2

 mvarentsov

 MKrinitkiy Mikhail Krinitkiy

 MKrinitkiy Lect01 video added 1e6cbaa · 5 days ago  44 Commits

 datasets	data for SPB added	5 days ago
 homework	Update HW1_description.md	last week
 images	DASIO subcollection added	3 weeks ago
 presentations	MK presentation Lect01 corrected	5 days ago
 .gitignore	DASIO subcollection added	3 weeks ago
 DASIO-dataset-description.md	DISO3 dataset description added	3 weeks ago
 DCIPP_dataset-description.md	Update DCIPP_dataset-description.md	last week
 DDM-dataset-description.md	Update DDM-dataset-description.md	last week
 DISO3-dataset-description.md	DISO3 dataset description added	3 weeks ago
 DUHI-dataset-description.md	Update DUHI-dataset-description.md	5 days ago
 HW1_description.md	data for SPB added	5 days ago
 README.md	Lect01 video added	5 days ago

 README  

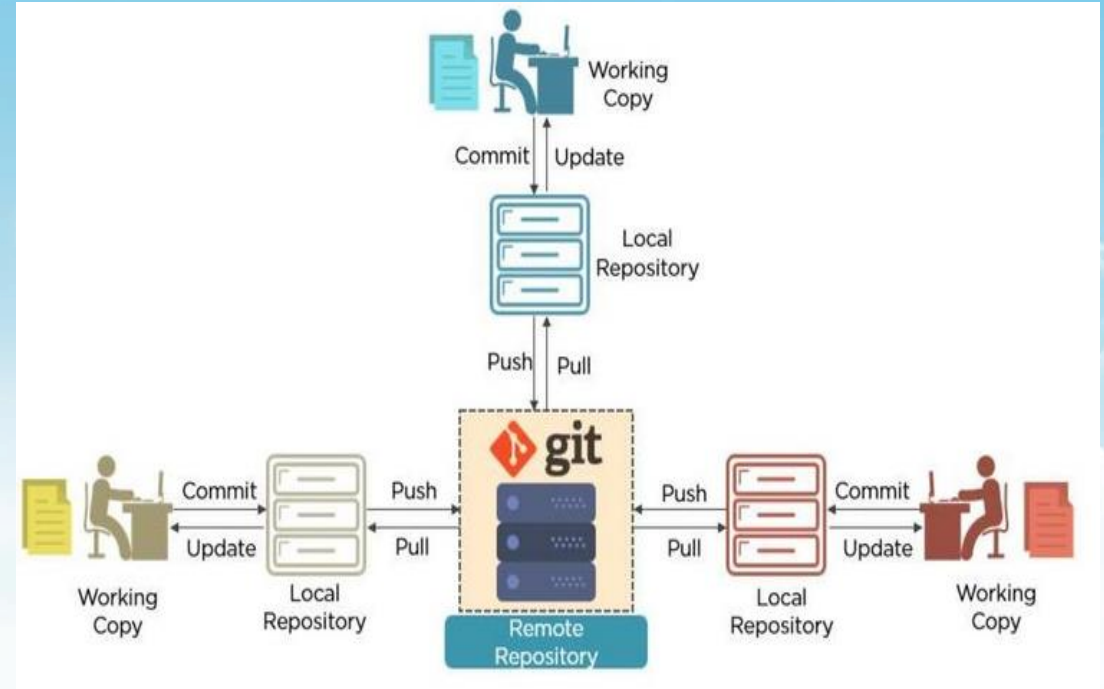
## ML4hydromet-2024

Курс "Машинное обучение в гидрометеорологии" для магистров I года обучения географического факультета МГУ имени М.В. Ломоносова.

<https://github.com/mvarentsov/ML4hydromet-2024>

# Репозитории Git – что это и зачем?

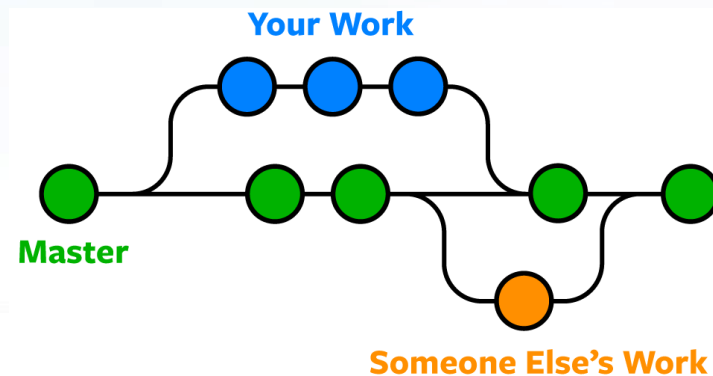
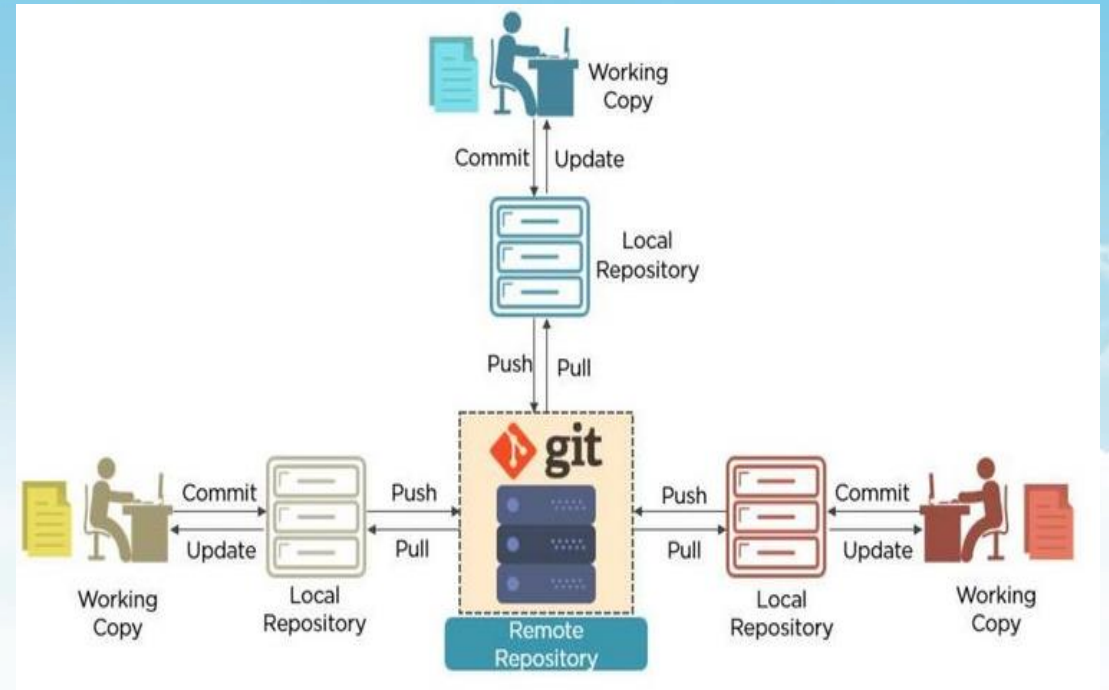
- ❑ Git - система управления версиями
- ❑ Придумал Линус Торвалдс при разработке ядра Линукс
- ❑ Современный стандарт для совместной разработки, в том числе в научной сфере и Data Science
- ❑ Задачи Git:
  - Синхронизация
  - Резервное копирование
  - Отслеживание и отмена изменений
  - Командная работа





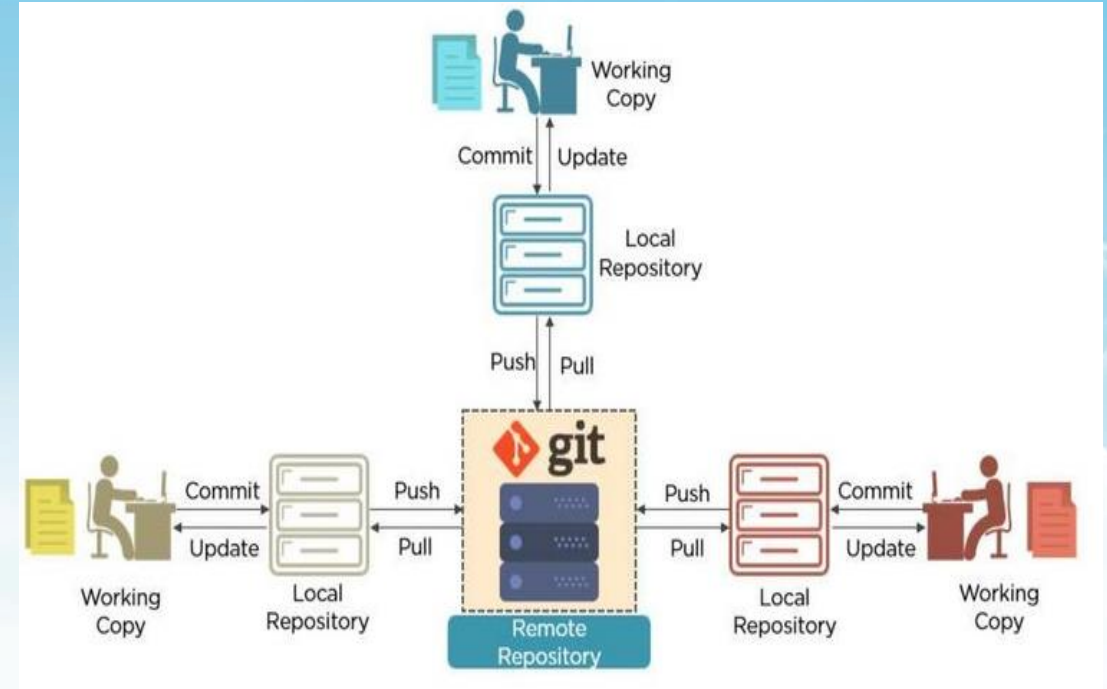
# Основные понятия Git

- ❑ **Репозиторий** — место, где хранится код (или данные)
  - Локальный
  - Удаленный (remote)
- ❑ **Коммит (commit)** — зафиксированное и неизменяемое состояние репозитория. Чаще всего их создают, когда:
  - Создан новый функционал
  - Добавлен новый блок на верстке
  - Исправлены ошибки по коду
  - Завершен рабочий день
- ❑ **Ветка (branch)** - независимая последовательность коммитов в хронологическом порядке



# Основные команды Git

- ❑ **git clone** – получения локальной копии существующего Git-репозитория  
`git clone https://github.com/mvarentsov/Urban-climate-modelling4HSE.git`
- ❑ **git fetch** – загрузка содержимого из удаленного репозитория (без изменения локального репозитория)
- ❑ **git pull = Git fetch + Git merge** – загрузка содержимого из удаленного репозитория и обновление локального репозитория
- ❑ **git add** – добавить файл в список тех, которые отслеживаются для текущего коммита
- ❑ **git commit** – зафиксировать текущие изменения
- ❑ **git push** – выгрузка содержимого локального репозитория в удаленный репозиторий.



# Графические интерфейсы для Git

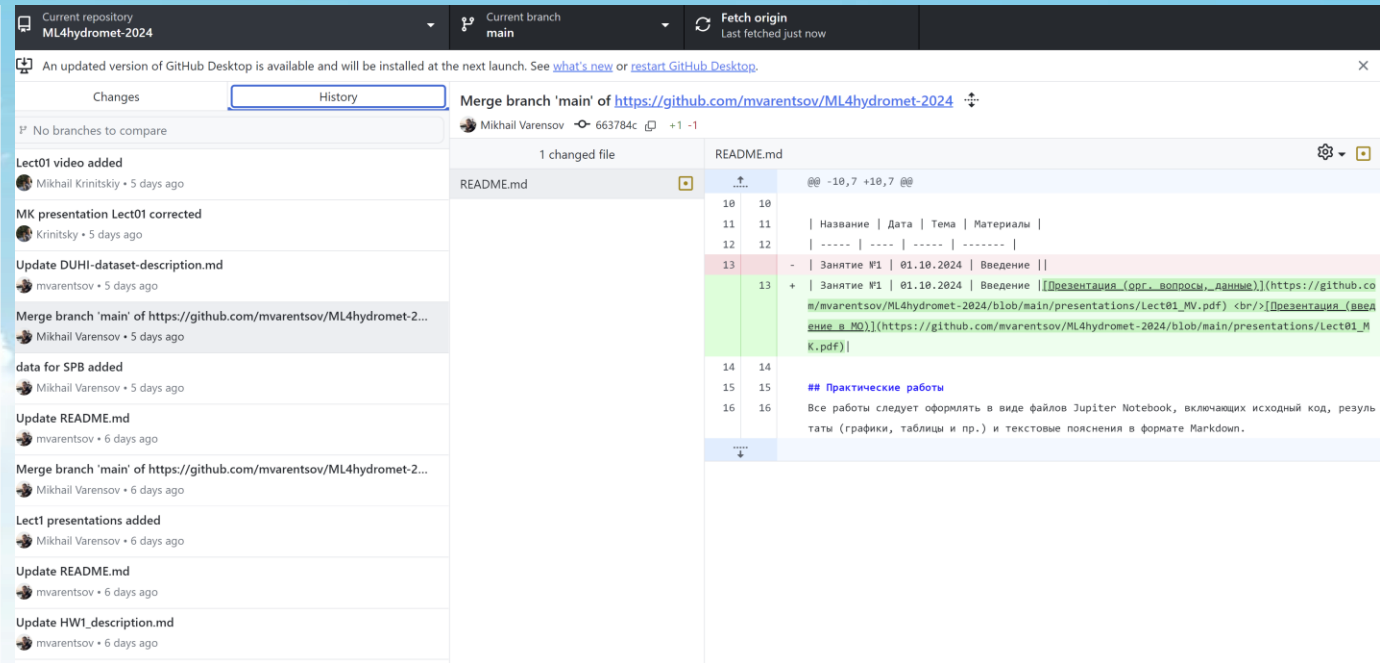
## ❑ GitHub – облачный Git-сервер

- Альтернатива – создать собственный сервер, например на базе GitLab

## ❑ GitHub Desktop – графический клиент для GitHub.

## ❑ Есть и альтернативы

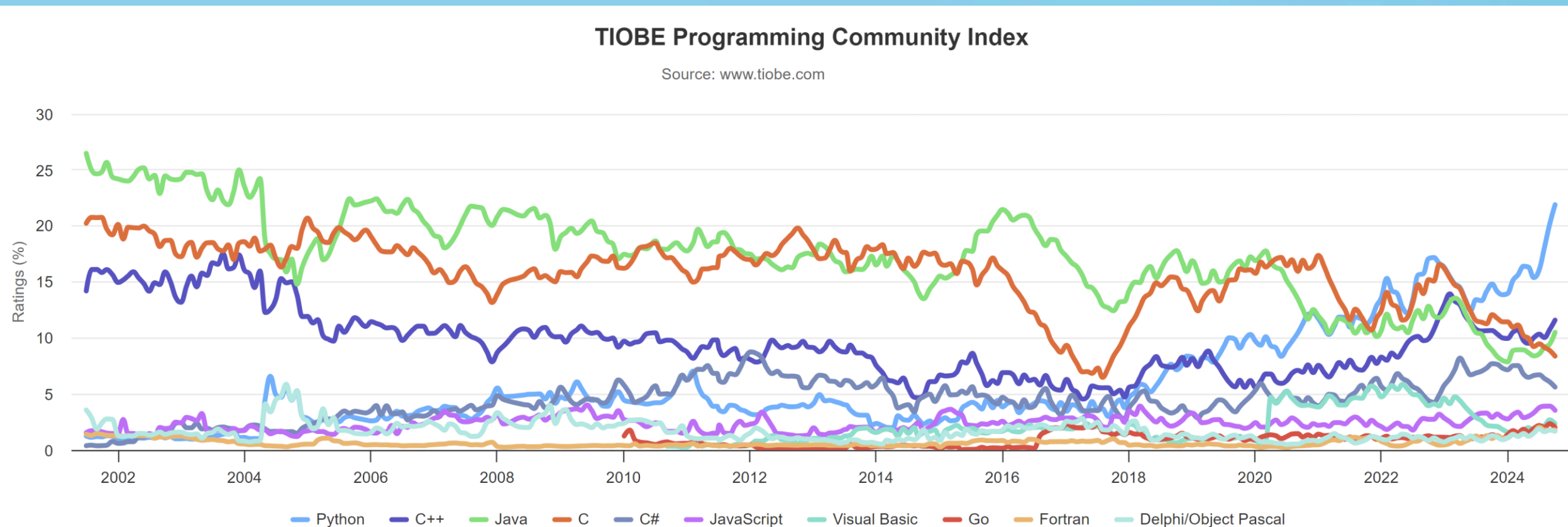
- Sourcetree (Windows, macOS и Linux)
- GitKraken (Windows, macOS)
- ...





# Python для анализа данных

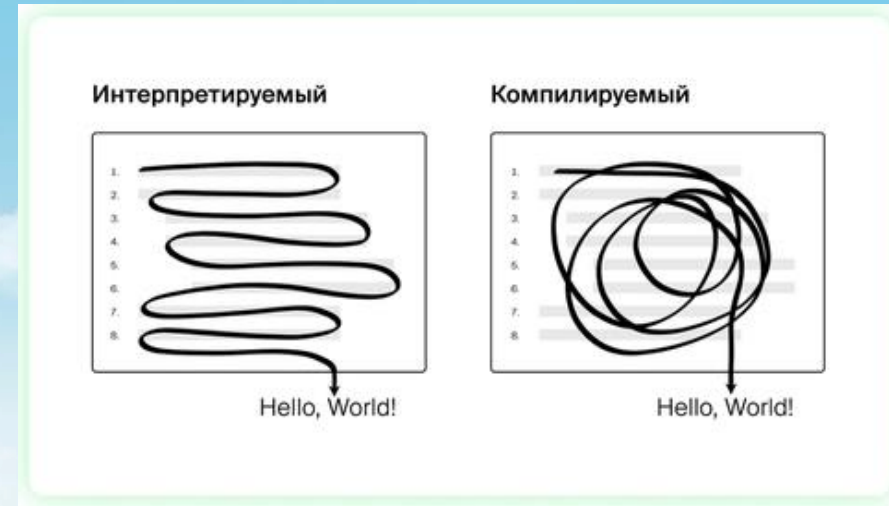
# Почему Python?



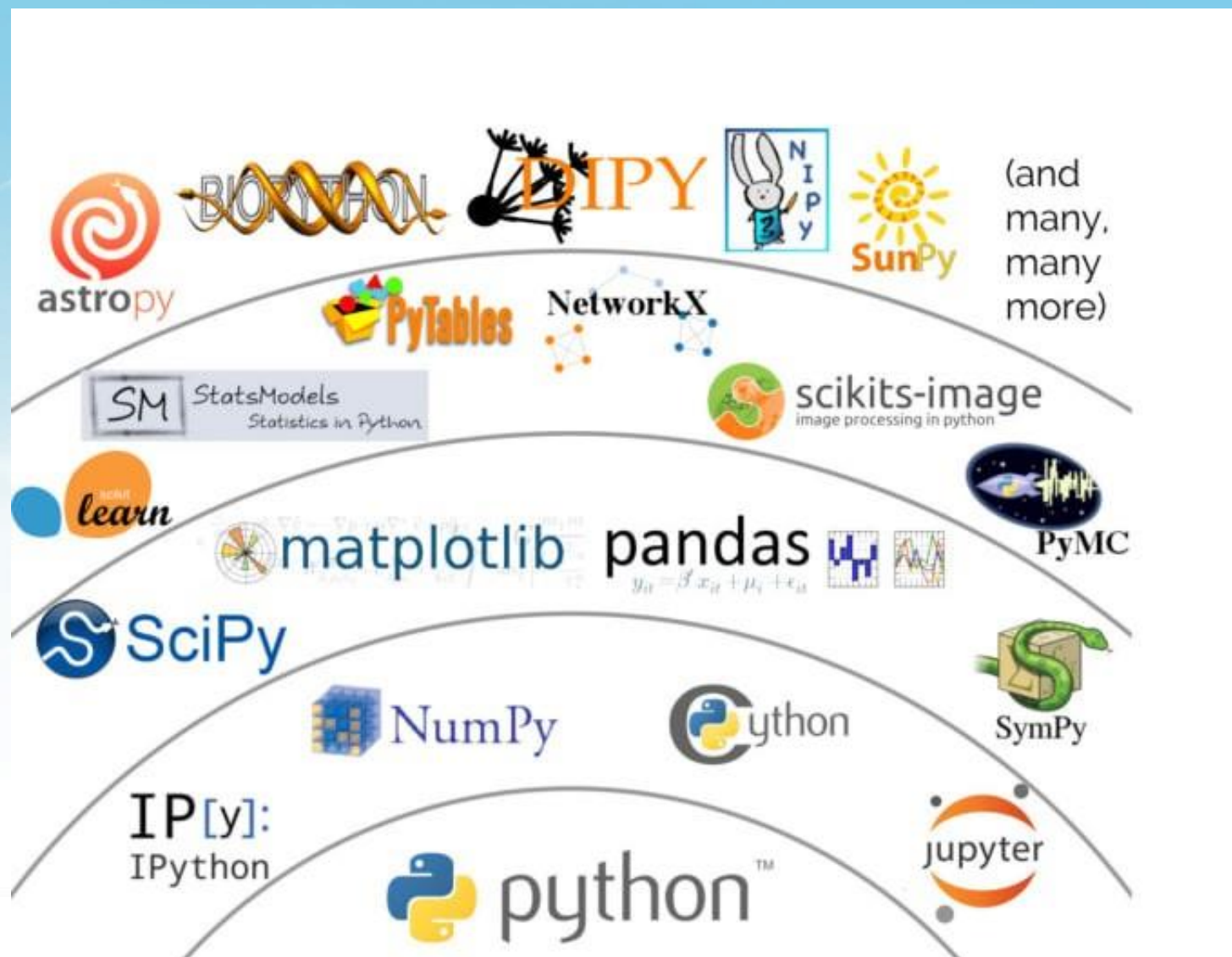
Рейтинг TIOBE (TIOBE Index) — это ежемесячный индекс, который отражает популярность языков программирования. Он основан на количестве поисковых запросов в различных поисковых системах, таких как Google, Bing, Yahoo!, Wikipedia и других.

# Экосистема Python

- ❑ Python – один из интерпретируемых языков программирования (наряду с R, Matlab, Julia и др.).
  - Исполняется ровно точно то, что написано
  - Исполняется построчно
  - Ошибки идентифицируются только в момент исполнения
- ❑ Ключевые элементы экосистемы:
  - **Дистрибутив Python** (например Anaconda)
  - **Среда выполнения:** python (встроенная), ipython, Jupiter notebook, Google Colab и пр.
  - **Среды разработки (IDE):** VS Code, PyCharm, Spyder, Jupiter Lab и пр.
  - **Менеджер пакетов** (pip, conda)
  - **Окружение (environment)**



# Python для анализа данных



# Python для анализа данных

## Наиболее важные для нашего курса:

- ❑ [NumPy](#) – работа с многомерными массивами, матрицами, математические операции
- ❑ [Pandas](#) – работа табличными данными
- ❑ [Matplotlib](#) – графики на все случаи жизни с ручной настройкой
- ❑ [Scikit-learn](#) – базовый уровень машинного обучения
- ❑ [Scipy](#) – статистический анализ

## Также могут пригодиться:

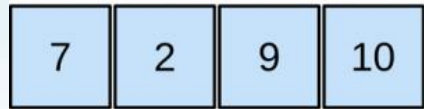
- ❑ [Xarray](#) – работа с многомерными массивами данных, имеющих пространственно-временную привязку (netcdf, grib, hdf)
- ❑ [Seaborn](#) – продвинутое графическое представление для статистического анализа
- ❑ [Rasterio](#) – работа с пространственными растровыми данными (geotiff)
- ❑ [Shapely](#), [geopandas](#) – работа с векторными пространственными данными





# Многомерные массивы: NumPy

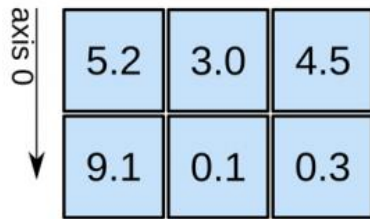
1D array



axis 0 →

shape: (4,)

2D array

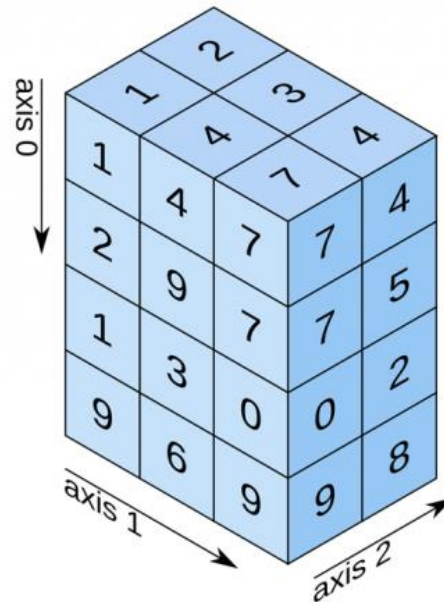


axis 0 ↓

axis 1 →

shape: (2, 3)

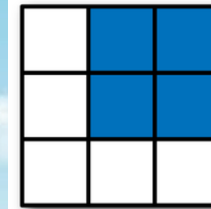
3D array



axis 0 ↓

axis 1 → axis 2 ↗

shape: (4, 3, 2)

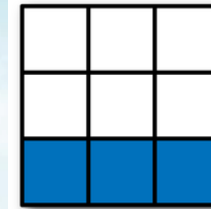


Expression

arr[:2, 1:]

Shape

(2, 2)



arr[2]

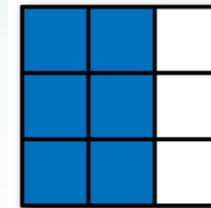
(3,)

arr[2, :]

(3,)

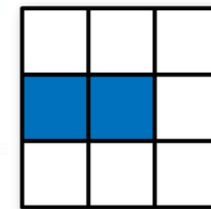
arr[2:, :]

(1, 3)



arr[:, :2]

(3, 2)



arr[1, :2]

(2,)

arr[1:2, :2]

(1, 2)

# Прямоугольные данные: Pandas

## Ключевые термины для прямоугольных данных

### Кадр данных (data frame)

Прямоугольные данные (подобно электронной таблице) — это базовая структура данных для статистических и автоматически обучающихся моделей.

### Признак (feature)

Столбец в таблице обычно называется признаком.

*Синонимы:* атрибут, вход, предсказатель, предиктор, переменная.

### Исход (outcome)

Многие проекты науки о данных предусматривают с предсказание исхода — нередко в формате да/нет (например, в табл. 1.1 это ответ на вопрос "Были ли торги состязательными или нет?"). Признаки иногда используются для предсказания исхода в эксперименте или статистическом исследовании.

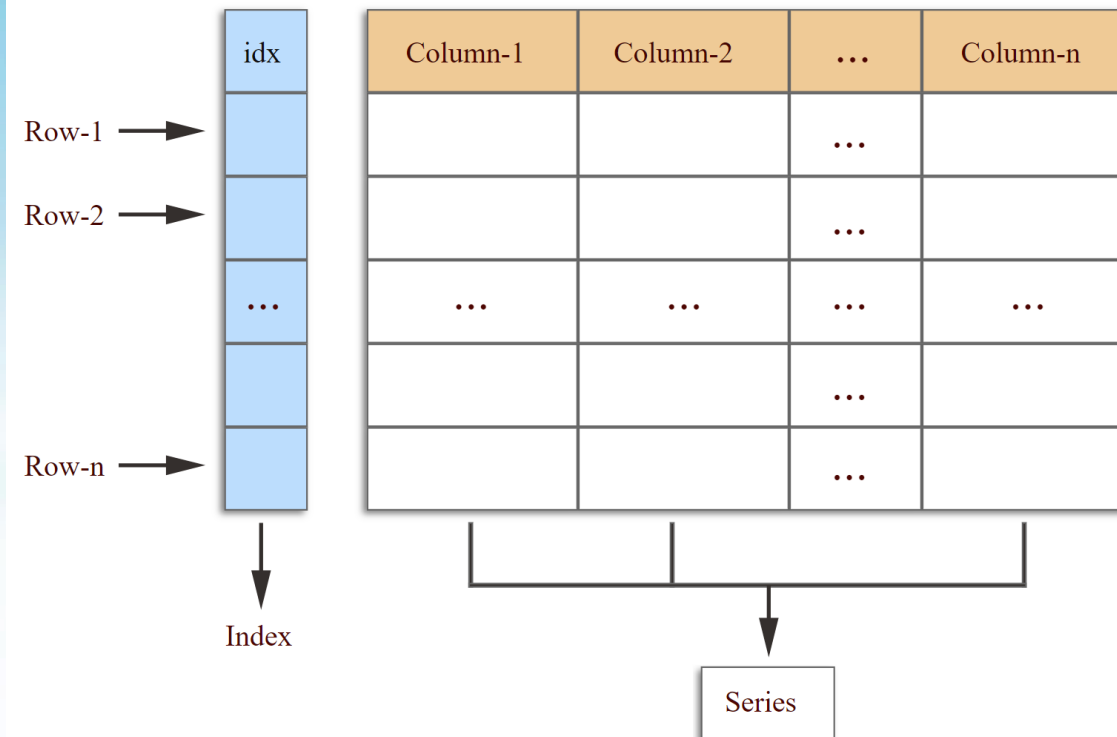
*Синонимы:* результат, зависимая переменная, отклик, цель, выход.

### Записи (records)

Строка в таблице обычно называется записью.

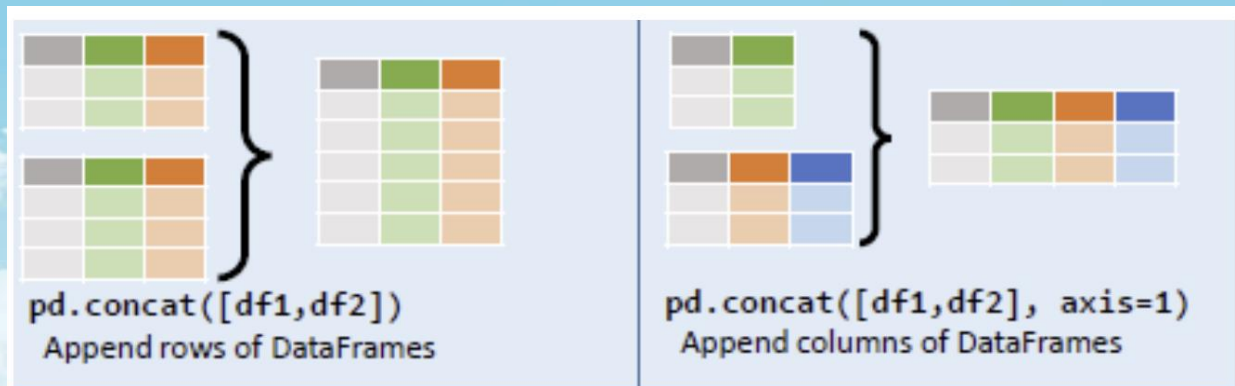
*Синонимы:* случай, пример, прецедент, экземпляр, наблюдение, шаблон, паттерн, образец.

## Pandas Data structure

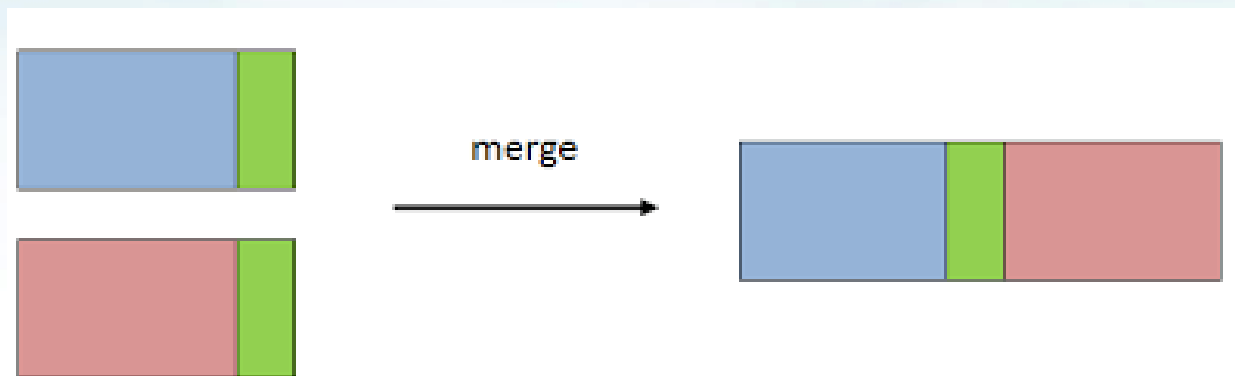


# Конкатенация данных в Pandas

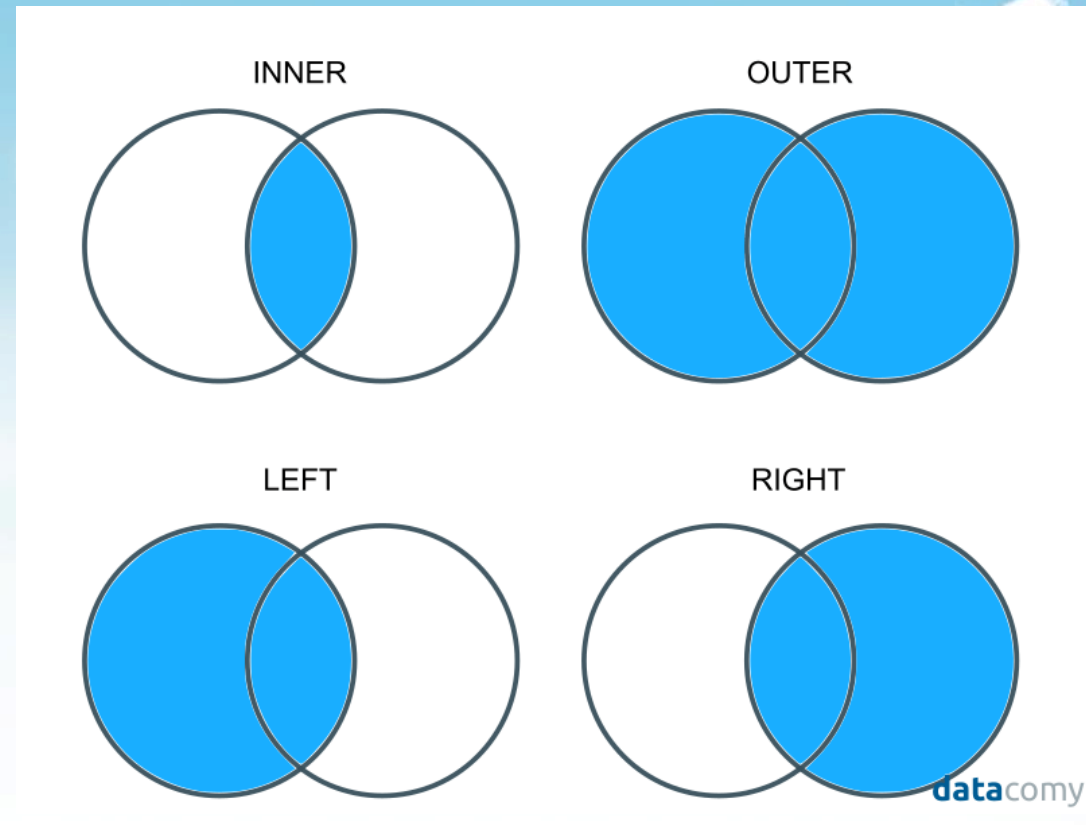
## pd.concat



## pd.merge

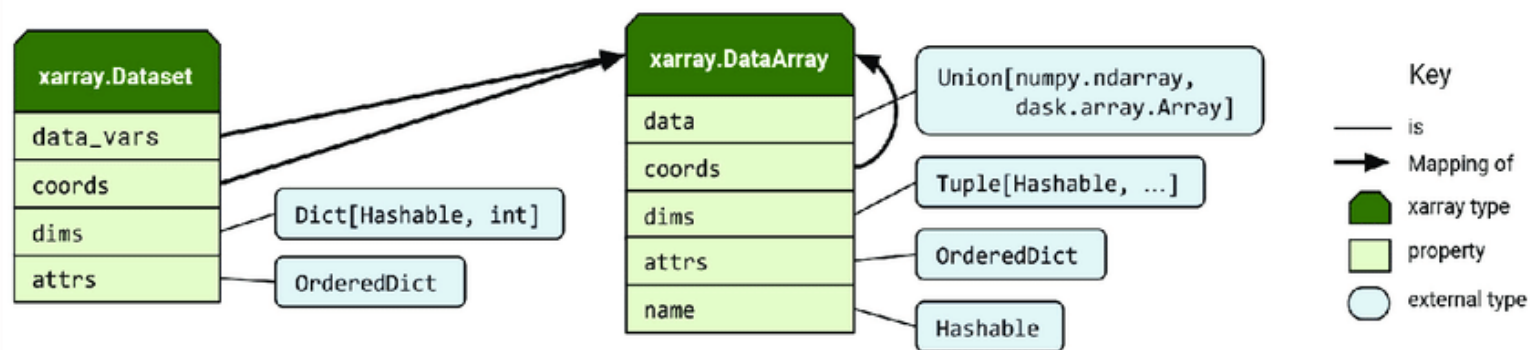
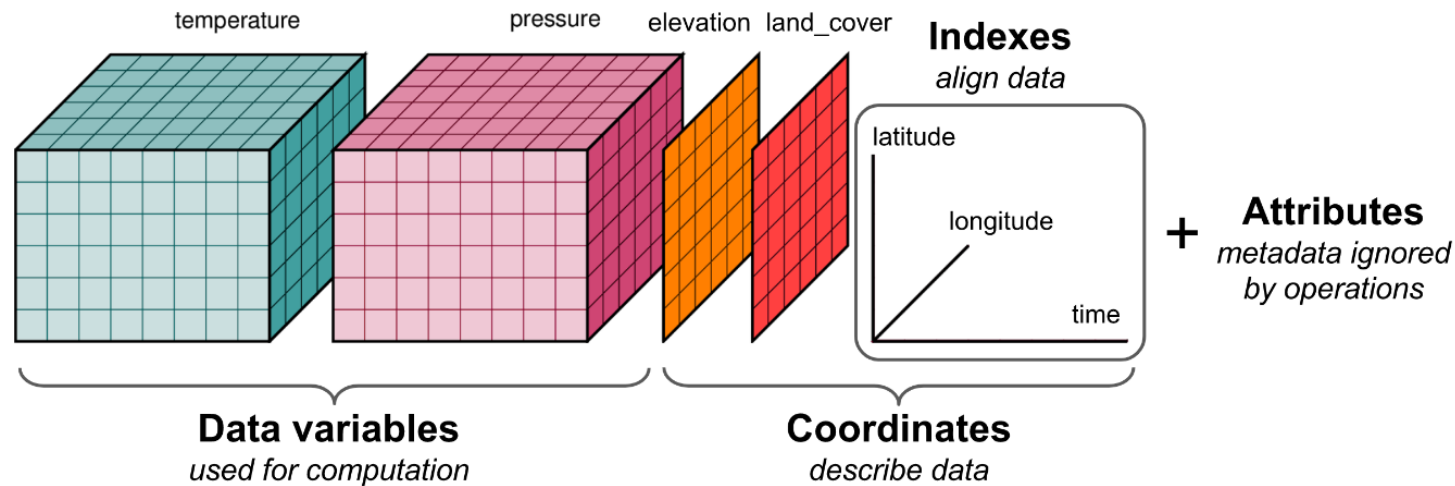


## DataFrame.join



- Функция **join()** объединяет два кадра данных по индексу.
- Функция **merge()** объединяет два кадра данных по любому указанному вами столбцу.

# Многомерные пространственные данные: xarray



Библиотека Xarray  
<https://docs.xarray.dev/en/stable/>



# Растровые геоданные: rasterio

Библиотека rasterio

File connection

```
src = rasterio.open('data/srtm.tif')
src

<open DatasetReader name='data/srtm.tif' mode='r'>
```

dict  
Metadata

```
src.meta

{'driver': 'GTiff',
 'dtype': 'uint16',
 'nodata': 65535.0,
 'width': 465,
 'height': 457,
 'count': 1,
 'crs': CRS.from_epsg(4326),
 'transform': Affine(0.0008333333332777796, 0.0, -113.23958321278403,
 0.0, -0.0008333333332777843, 37.512916763165805)}
```

ndarray  
Values

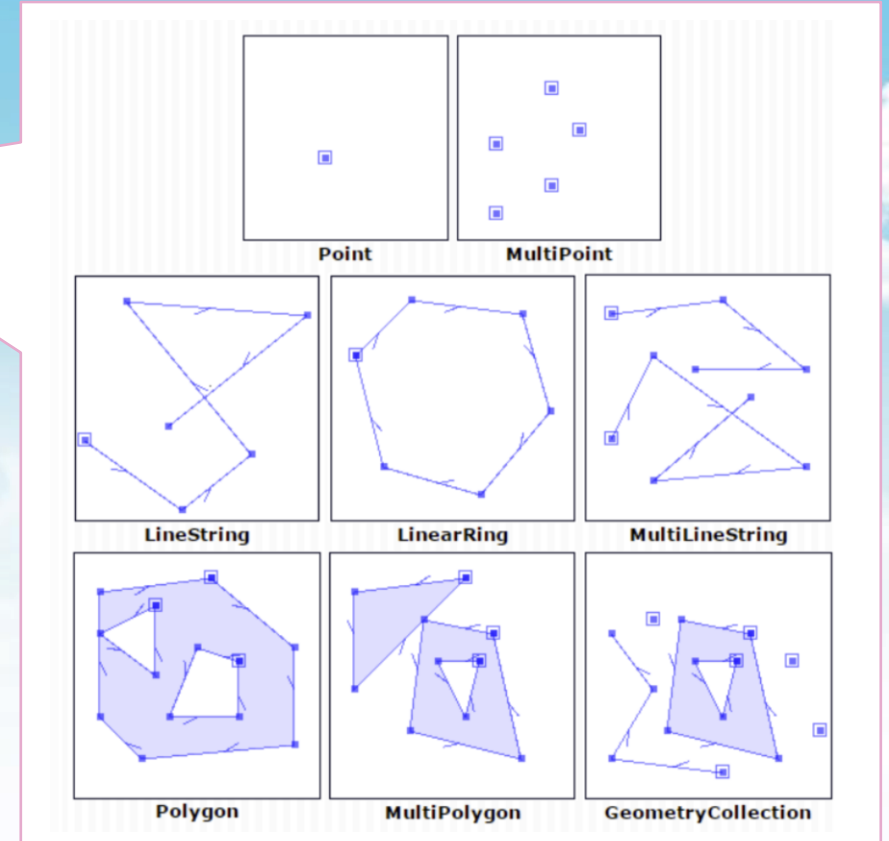
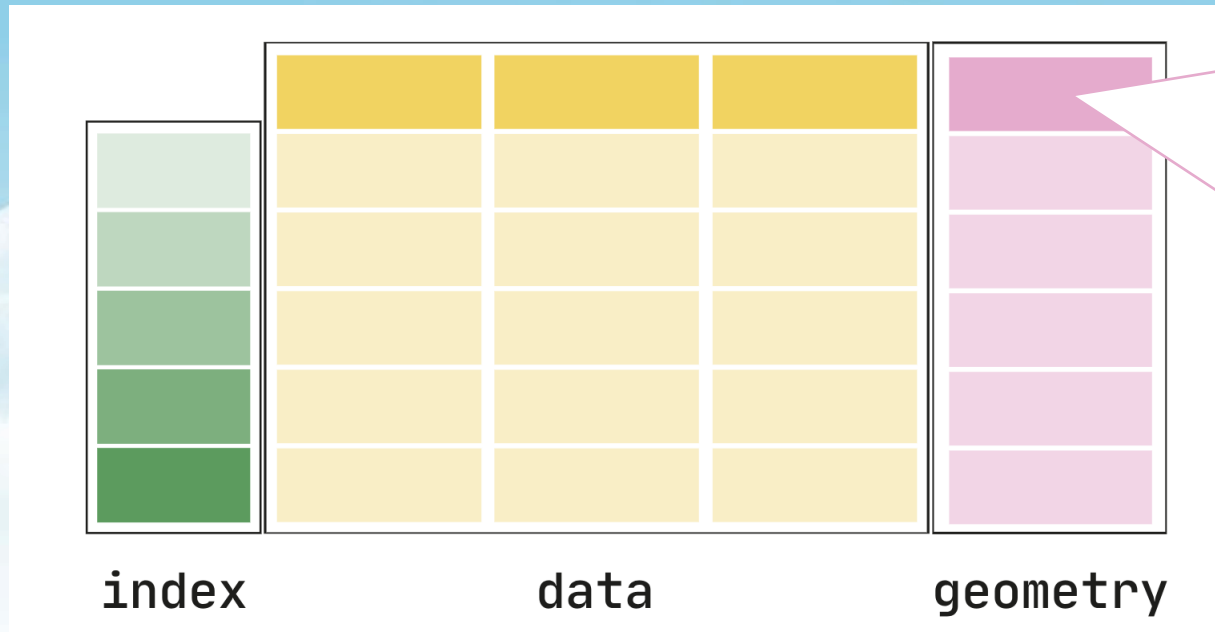
```
src.read(1)

array([[1728, 1718, 1715, ..., 2654, 2674, 2685],
       [1737, 1727, 1717, ..., 2649, 2677, 2693],
       [1739, 1734, 1727, ..., 2644, 2672, 2695],
       ...,
       [1326, 1328, 1329, ..., 1777, 1778, 1775],
       [1320, 1323, 1326, ..., 1771, 1770, 1772],
       [1319, 1319, 1322, ..., 1768, 1770, 1772]], dtype=uint16)
```

<https://rasterio.readthedocs.io/en/stable/#>



# Векторные геоданные



 Shapely +  pandas =  GeoPandas

<https://geopandas.org/en/stable/>



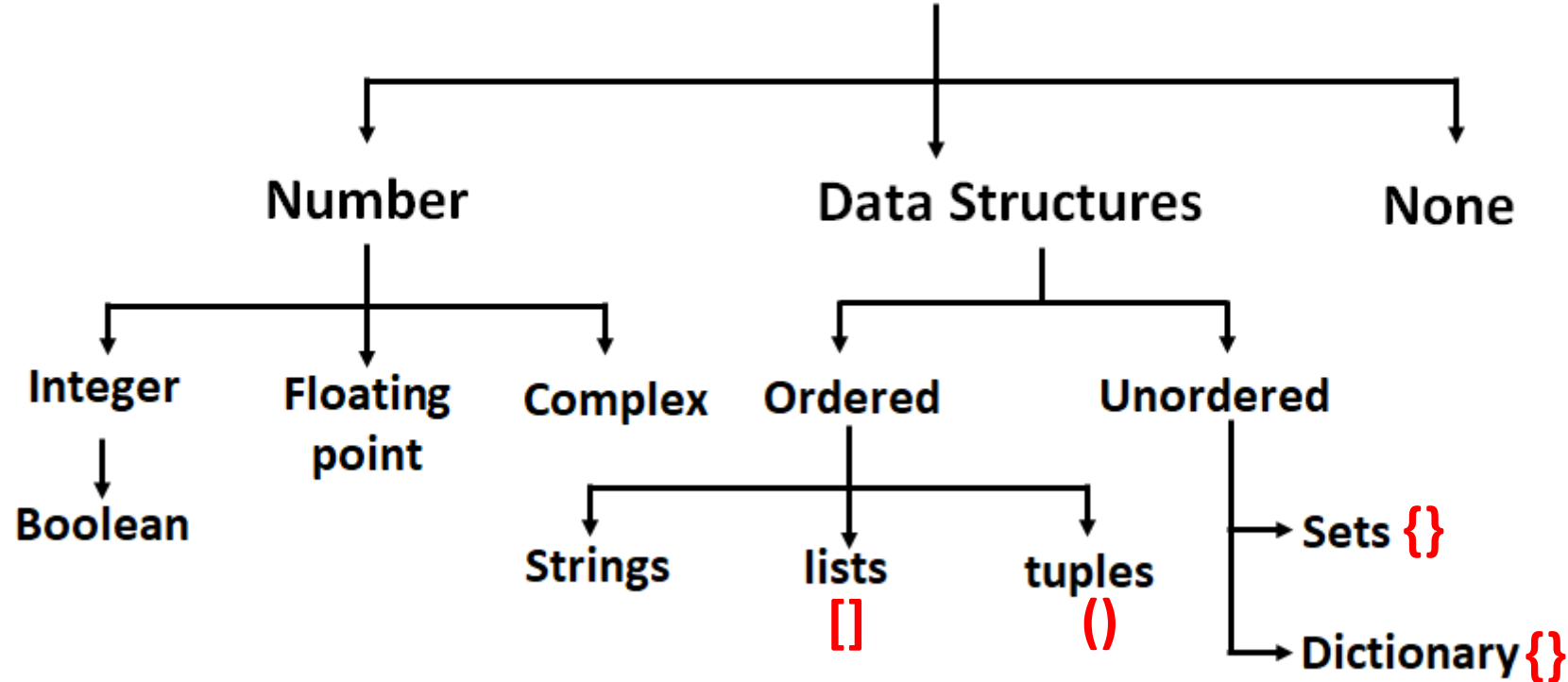
**Перейдем к практике**

# Дополнительные слайды

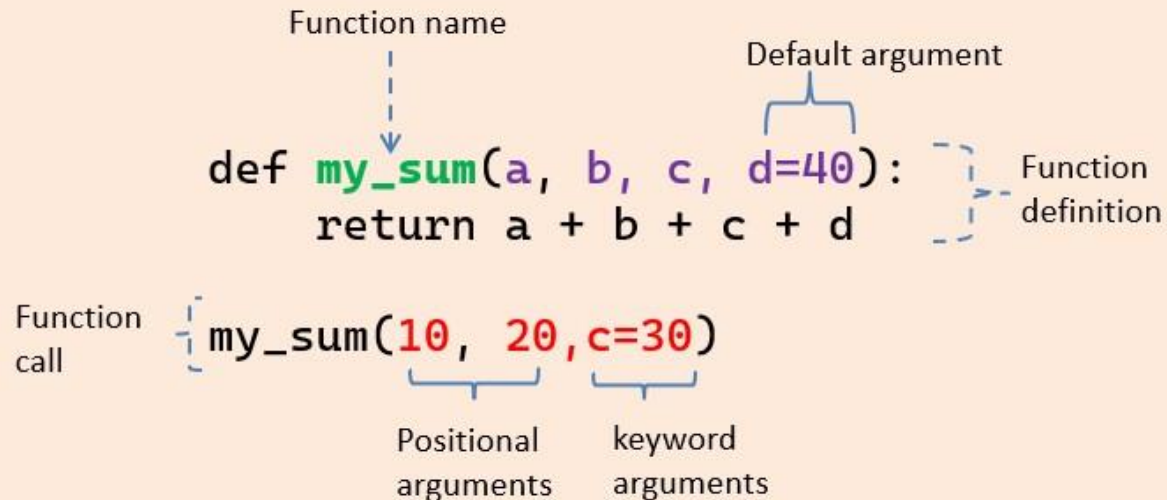
# Базовые типы данных в Python

*teachoo*

## Python Data Types



# Функции и классы в Python



- **Positional argument** values get assigned as per the sequence. Now `a=10` and `b=20`
- **Keyword arguments** are those arguments where values get assigned to the arguments by their keyword
- **Default arguments:** Assign default values to the argument using the '=' operator at the time of function definition

```
class Student:  
  
    def __init__(self, name, marks):  
        self.name = name  
        self.marks = marks  
  
    def check_pass_fail(self):  
        if self.marks >= 40:  
            return True  
        else:  
            return False  
  
student1 = Student('Harry', 85)  
did_pass = student1.check_pass_fail()  
print(did_pass)  
  
student2 = Student('Janet', 30)  
did_pass = student2.check_pass_fail()  
print(did_pass)
```