# 1. Основы языка Python

## Часть 1. Python и интеллектуальный анализ данных (Data Mining) – введение.

**В этом разделе:**

- Зачем нужен Python при анализе данных?
- Как установить пакет Anaconda и начать с ним работать?

**Александр Владимирович Толмачев**

axtolm@gmail.com

## Определения. Интеллектуальный анализ данных (Data Mining)

**1) Data Mining** — это **процесс обнаружения в «сырых» данных знаний**, необходимых для **принятия решений** в различных сферах человеческой деятельности (Григорий Пятецкий-Шапиро, 1992 г.).

При этом знания должны быть ранее неизвестными, нетривиальными, практически полезными и доступными интерпретации.

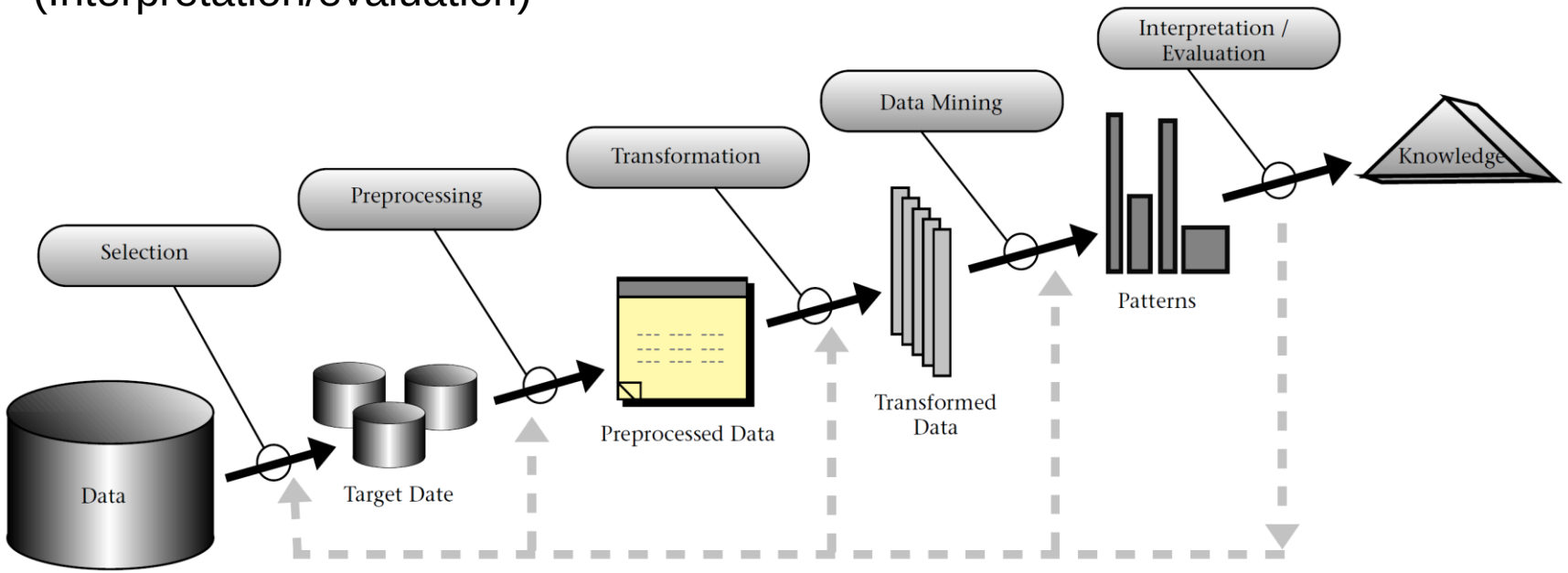
**2) Data Mining** — это **современная концепция анализа данных**, предполагающая, что:

- данные могут быть неточными, неполными, противоречивыми, разнородными, и при этом иметь гигантские объёмы;
- алгоритмы анализа данных могут обладать «элементами интеллекта», в частности, способностью обучаться по прецедентам, а их разработка также требует значительных интеллектуальных усилий;
- процессы переработки сырых данных в информацию, а информации в знания не могут быть выполнены вручную и требуют нетривиальной автоматизации.

**Data Mining** - часть более общего процесса **извлечения знаний из баз данных** («Knowledge Discovery in Databases" или KDD)[1].

**Этапы KDD:**

1. Отбор данных (Selection )
2. Предварительная обработка данных (Pre-processing)
3. Преобразование данных (Transformation)
4. Интеллектуальный анализ данных (Data Mining)
5. Интерпретация и оценка результатов (Interpretation/evaluation)




**Python – полноценный инструмент KDD, используемый на всех этапах.**

**Почему Python?**

- прост в освоении,
- широко распространен,
- много библиотек для решения прикладных задач,
- open source проект,
- кроссплатформенный, работает с CPU и GPU,
- активно развивается,
- есть своя философия.

[1] Fayyad U., Piatetsky-Shapiro G., & Smyth P. From data mining to knowledge discovery in databases. Ai Magazine, vol. 17, no. 3, pp. 37-54, 1996

## Подробнее о Data Mining

**Шесть классов задач, которые решают с помощью Data Mining:**

1. Классификация (Classification)
2. Регрессия (Regression)
3. Кластеризация (Clustering)
4. Обобщение (Summarization)
5. Моделирование зависимостей (Dependency modelling)
6. Обнаружение аномалий (Change and deviation detection)

**Вручную их не выполнить, а для автоматизации нужны инструменты.**
**Мы будем использовать Python.**

## Работать с Python будем с помощью пакета Anaconda Individual Edition для Windows

- Популярность у 25M+ пользователей в мире.
- Установка на Linux, Windows, Mac OS.
- Основные библиотеки для работы с данными (250+) идут в комплекте поставки.
- 7.5K+ библиотек доступны в облаке.
- Open-source для Data Mining и Machine Learning.


**Установка Anaconda Individual Edition**:

Скачать установщик https://www.anaconda.com/products/individual и запустить его.

Подробности процесса установки: https://docs.anaconda.com/anaconda/install/

# Anaconda Individual Edition - когда и что из компонентов будем использовать в работе


**IDE Spyder**

Интегрированная среда разработки

→ код, требующий **отладки**

→ относительно объемный код

→ код с длительным временем исполнения

Для **отладки** есть все необходимое:

✓ менеджер переменных

✓ точки останова и пошаговое выполнение

✓ профайлер


**Jupyter Notebook**

Web приложение для работы с документами (текст + код на Python + результаты его выполнения)

→ алгоритмы с пошаговым выводом результатов

→ код и результаты, которыми нужно поделиться

→ отчеты для программистов и непрограммистов


**Conda**

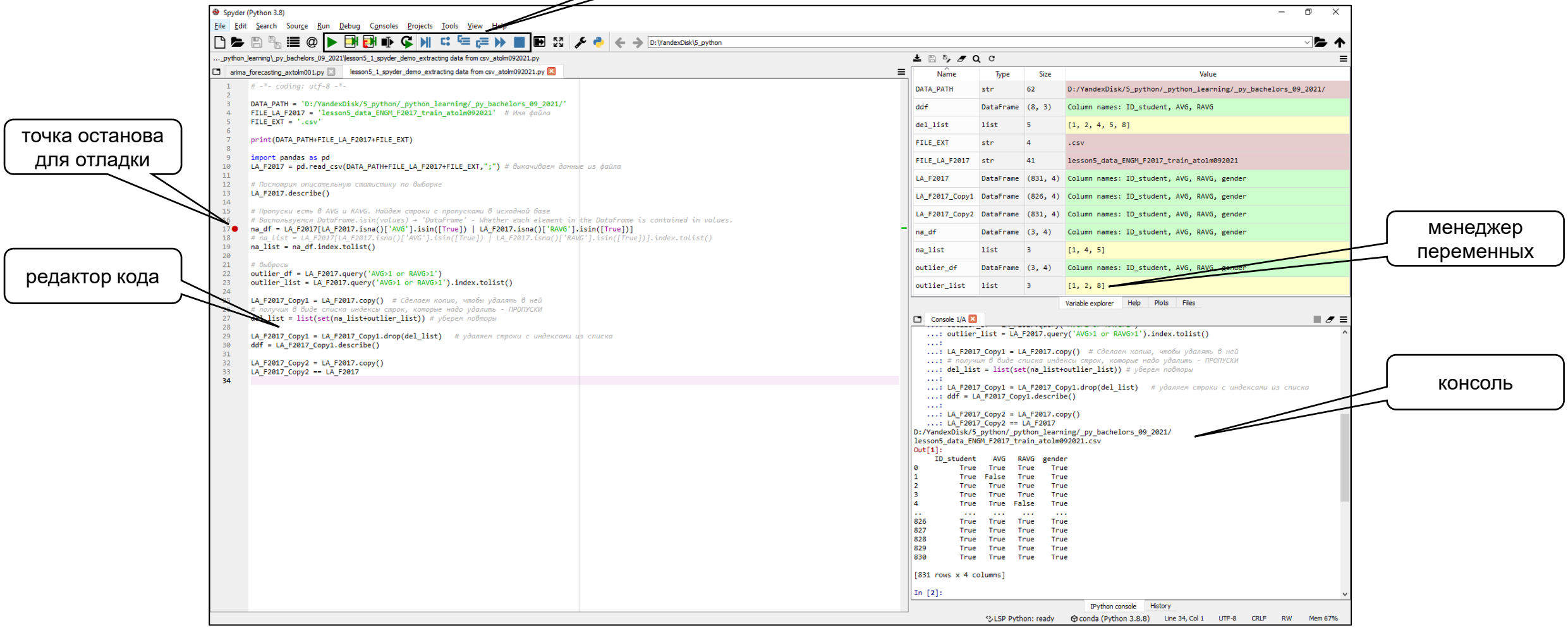
Менеджер для управления библиотеками и окружением

→ установка библиотек

→ создание виртуальных окружений

# IDE Spyder – интегрированная среда разработки в составе Anaconda Distribution

https://www.spyder-ide.org/



https://jupyter.org/

**The Jupyter Notebook** – это open-source **web приложение,** которое позволяет создавать **документы**, содержащие **тексты и рисунки**, **код на Python и результаты его выполнения**; конвертировать документы в HTML и другие форматы.

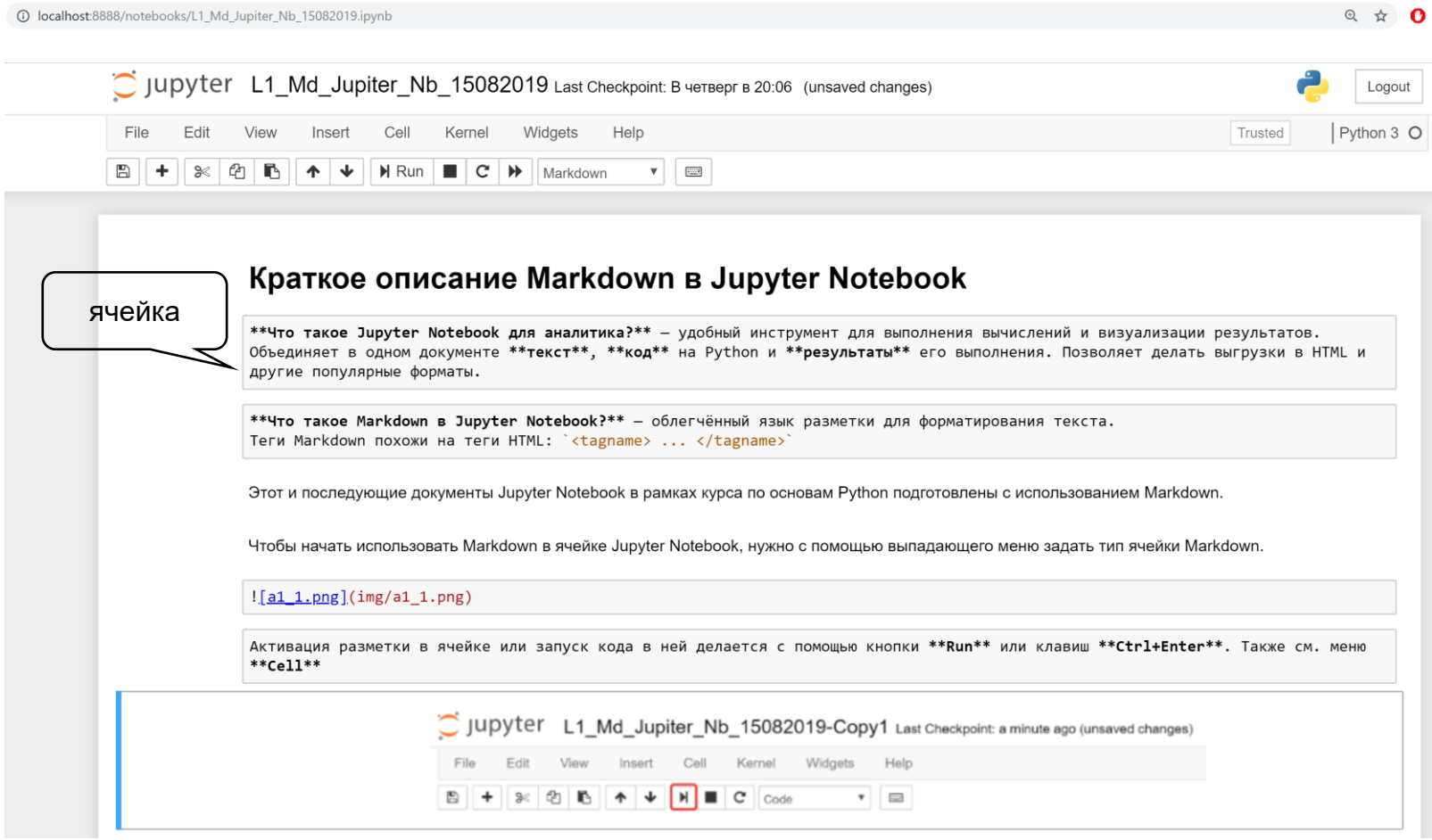
Поставляется в составе Anaconda Distribution

## Ячейка – ключевой элемент

В ячейке может быть:

- **текст**, который можно форматировать с помощью Markdown
- **код на Python** и результаты его выполнения
- **рисунок**

Есть еще **JupyterLab** – Web IDE для Jupyter notebooks

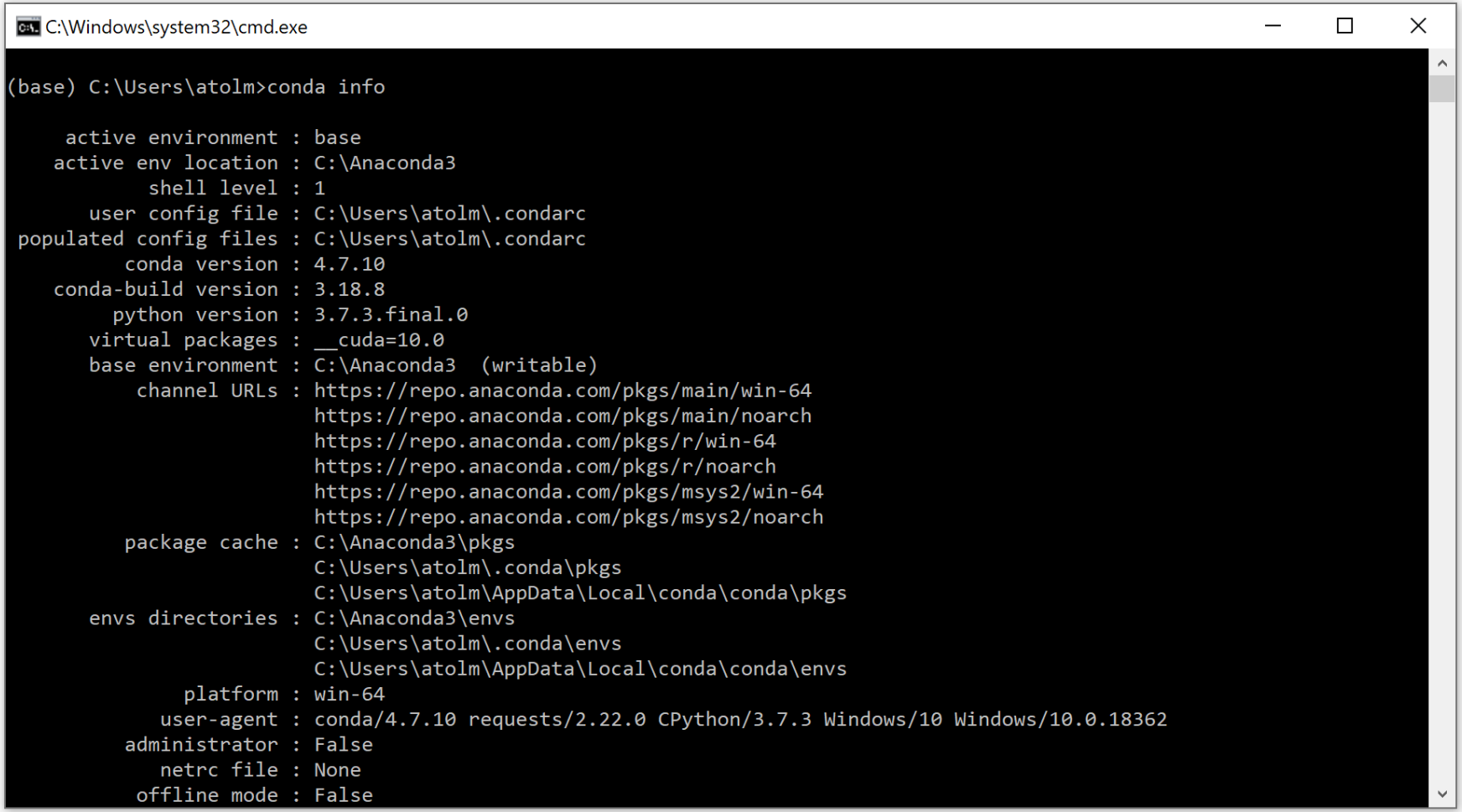


**Conda** – это open source менеджер для управления библиотеками и окружением, который работает на Windows, macOS, Linux.

Поставляется в составе Anaconda Distribution

Используем для установки библиотек с командной строки в терминале:
**conda install PKGNAME**

PKGNAME – имя устанавливаемого пакета



## Запуск приложений Spyder, Jupyter Notebook, Conda из пакета Anaconda

### Запуск Spyder

**Windows menu Start**
→ Anaconda3 (64-bit)
→ **Spyder** (anaconda3)

Запускается как самостоятельное приложение в своем окне

### Запуск Jupyter Notebook

**Windows menu Start**
→ Anaconda3 (64-bit)
→ **Jupyter Notebook** (anaconda3)

Запускается в окне браузера

### Запуск Conda

**Windows menu Start**
→ Anaconda3 (64-bit)
→ **Anaconda Powershell Prompt (anaconda3)**

Запускается терминал с командной строкой

**Подведем итоги:**

- Мы рассмотрели, что такое Data Mining и какие задачи можно решать с его помощью.
- Узнали, какое место занимает Data Mining в более общем процессе извлечения знаний KDD.
- Выяснили, зачем при анализе данных нужен язык Python.
- Установили пакет Anaconda Individual Edition на персональный компьютер.
- Научились запускать приложения Spyder, Jupyter Notebook, Conda.