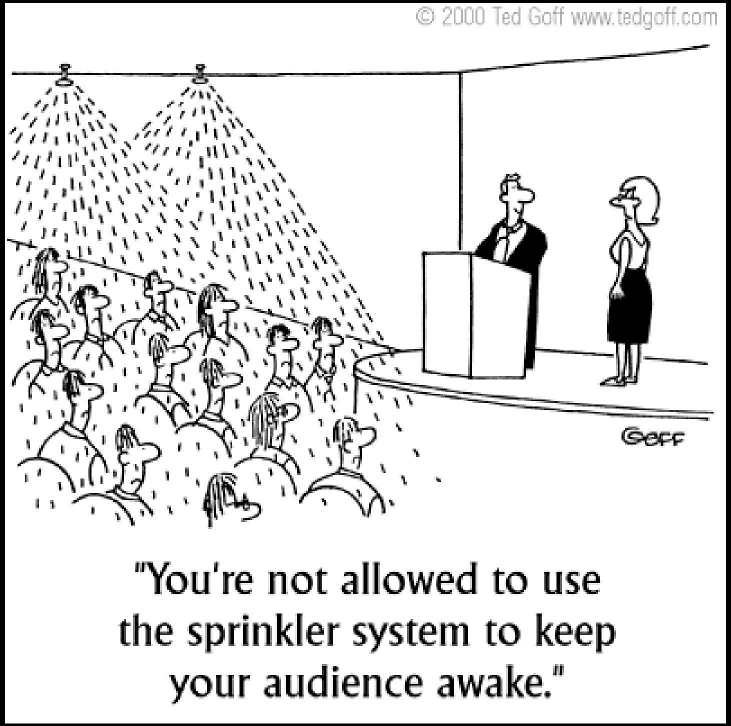
Empowering Data Enthusiasts with Open-Source Tools

Eduard Maievskyi, Ph.D.

www.linkedin.com/in/eduard-maievskyi-002618107

VanLUG, Burnaby Public Library (Metrotown) Branch
2024-06-15

Let’s have a great time together!



"You're not allowed to use the sprinkler system to keep your audience awake."

Outline

- **Introduction**
  - Who am I?
  - Why Linux???
  - The Open-Source Advantage
  - “This is the way” or why I prefer Python
  - SQL
- **Open-Source IDEs: Your Data Workspace**
  - What’s an IDE?
  - Top Open Source IDE Choices for Data Enthusiasts/Pythonists (personal opinion)
  - Virtual Environments!!!
- **Essential Python Libraries for Data Professionals**
  - Data Manipulation and Analysis (pandas, numpy)
  - Data Visualization (matplotlib, seaborn, plotly, D3.js)
  - Mathematical Modelling, Machine Learning, TSF, Deep Learning (scipy, scikit-learn, statsmodels, fbprophet, tensorflow, keras, pytorch, etc)
- **Conclusions**

Introduction

Who am I?

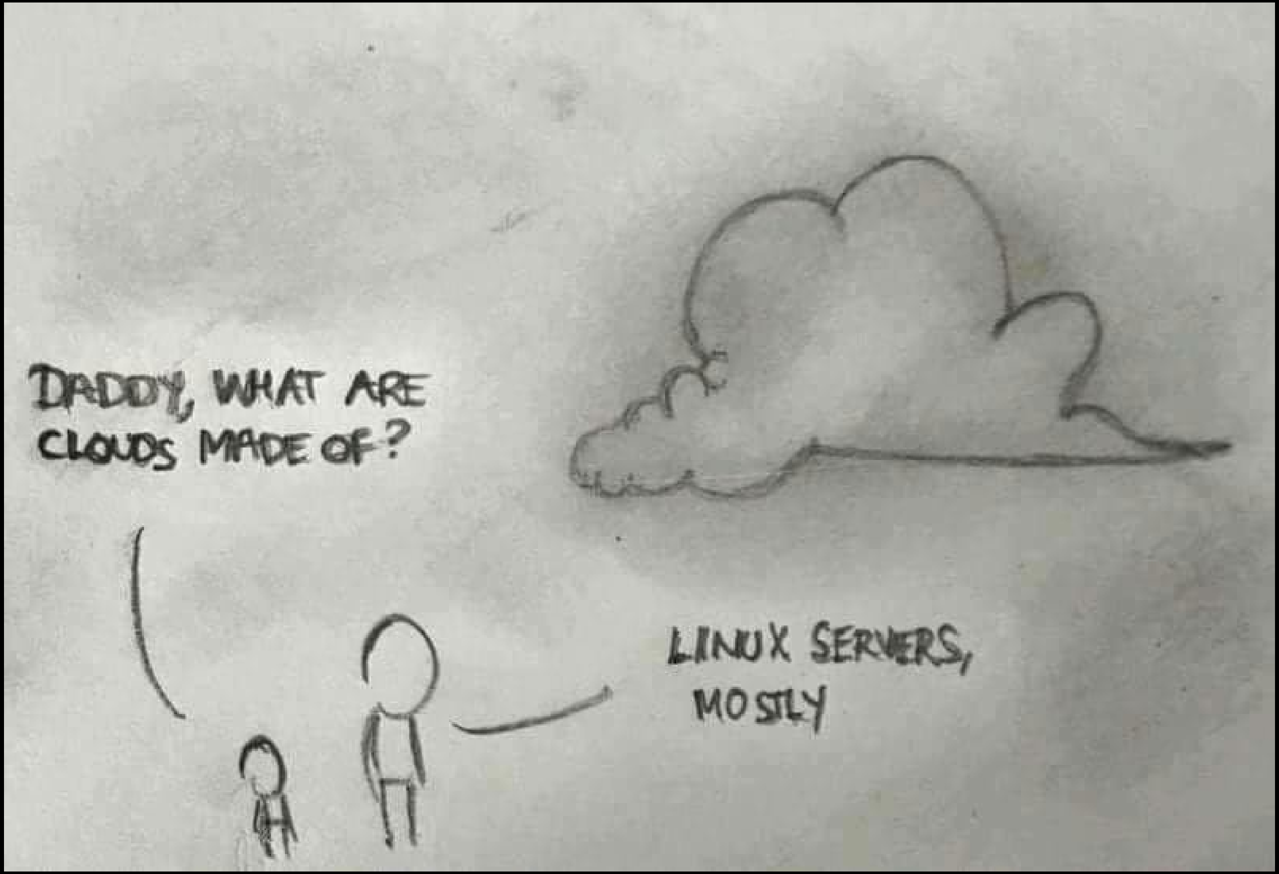
Genius, billionaire,
playboy, philanthropist.

-Who am I? -Eduard Maievskyi...

- 🇺🇦 **B. Sc. in Physics (June 2011)**
  - V. N. Karazin Kharkiv National University, Kharkiv, Ukraine
- 🇺🇦 **M. Sc. In Physics (May 2012)**
  - V. N. Karazin Kharkiv National University, Kharkiv, Ukraine
- 🇵🇱 **Ph. D. in Physics (October 2012 – July 2017)**
  - W. Trzebiatowski Institute of Low Temperature and Structure Research Polish Academy of Sciences in Wroclaw,Poland
- 🇵🇱 **Data Scientist (July 2017 -July 2018)**
  - Physiolution Polska sp. z o.o., Wrocław, Poland
- 🇳🇱 **Data Scientist (September 2018 – April 2019)**
  - Yieldr Labs DI B.V., Amsterdam, Netherlands.
- 🇳🇱 **Data Scientist (April 2019 – October 2019)**
  - Hotelchamp B.V., Amsterdam, Netherlands
- 🇳🇱 **Data Scientist (December 2019 - March 2020)**
  - Belvilla Services B.V., Amsterdam, Netherlands
- 🇵🇱 **Research Assistant (March 2020 – December 2021)**
  - W. Trzebiatowski Institute of Low Temperature and Structure Research Polish Academy of Sciences in Wroclaw,Poland
- 🇵🇱 **Vice President, BI/Back-End Developer (January 2022 - February 2024)**
  - BNY, Wrocław, Poland
- 🇨🇦 **Lead Software Developer (April 2024 - present)**
  - Quantum World Technologies Inc. + Virtusa + BNY, Vancouver, BC

www.linkedin.com/in/eduard-maievskyi-002618107

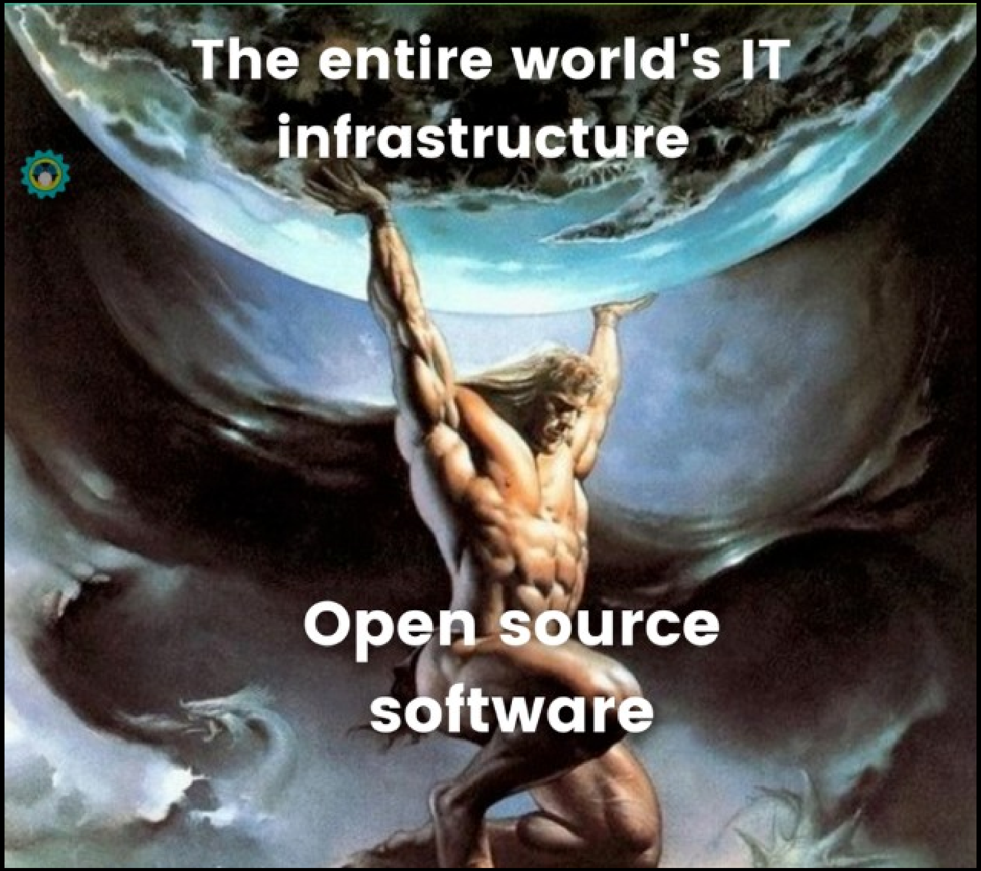
Why Linux?



Why Linux?

1. Open-source nature and customization flexibility
2. Range of applications and tools
3. Less security issues
4. Reliability
5. Community support and resources
6. Cost-effectiveness compared to proprietary software
7. Scalability for handling large amounts of data and traffic
8. Compatibility with hardware and software platforms
9. Continued growth and usage of Linux for servers

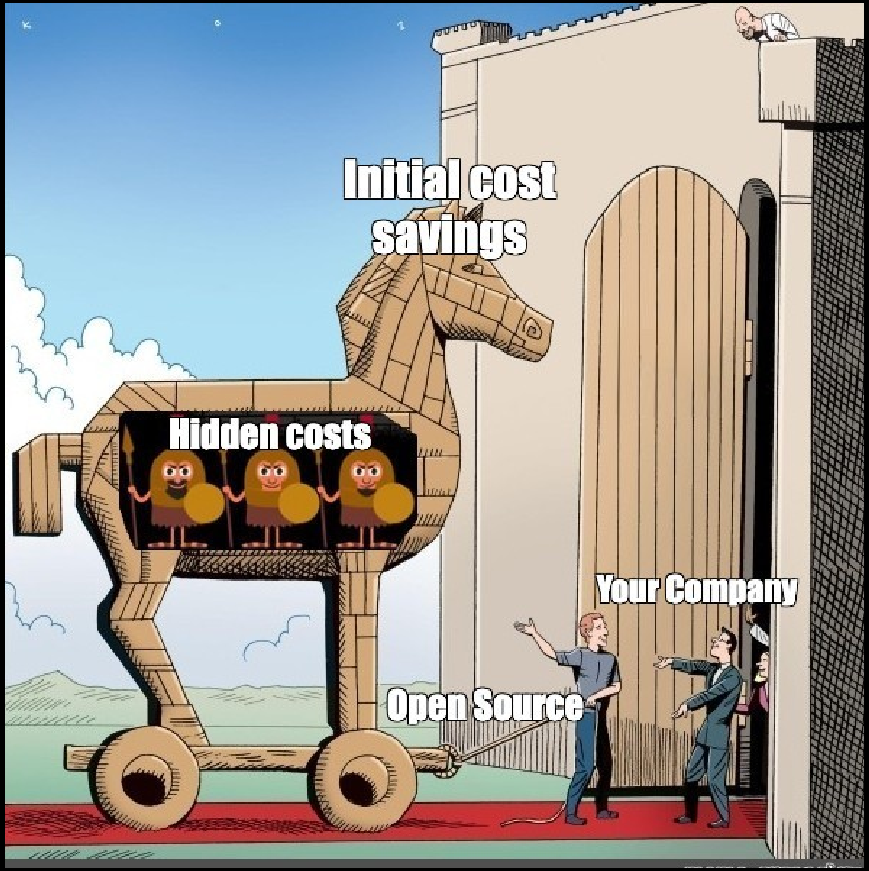
The Open-Source Advantage



The Open-Source Advantage

1. Flexibility and Agility
2. Ability to start small
3. Speed (especially with PRO support)
4. Enhanced security through transparency
5. Attracting talents
6. Powering the digital transformation
7. No vendor lock-in

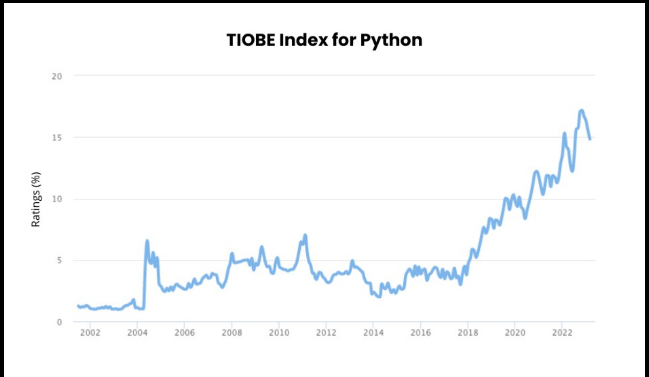
The Open-Source Disadvantage



“This is the way” or why I prefer Python

1. The most popular programming language
2. One of the most versatile coding languages
3. Gentle learning curve
4. Excellent standard and external libraries
5. A supportive and robust community
6. High demand
7. The Pythonic Way or the Zen of Python

https://peps.python.org/pep-0020/



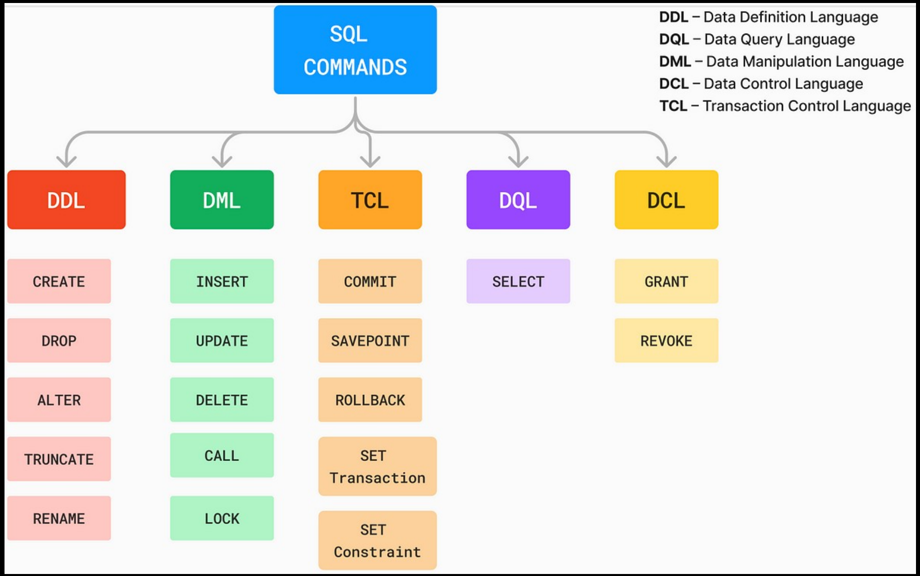
https://www.tiobe.com/tiobe-index/

Guido van Rossum
Creator of Python.
Python’s BDFL
(benevolent
dictator for life)
until July 2018

SQL

Structured Query Language (SQL) (pronounced S-Q-L; historically "sequel") is a domain-specific language used to manage data, especially in a relational database management system (RDBMS). It is particularly useful in handling structured data, i.e., data incorporating relations among entities and variables.

Introduced in the 1970s… https://en.wikipedia.org/wiki/SQL



SQL COMMANDS

DDL – Data Definition Language
DQL – Data Query Language
DML – Data Manipulation Language
DCL – Data Control Language
TCL – Transaction Control Language

DDL	DML	TCL	DQL	DCL
CREATE	INSERT	COMMIT	SELECT	GRANT
DROP	UPDATE	SAVEPOINT		REVOKE
ALTER	DELETE	ROLLBACK		
TRUNCATE	CALL	SET Transaction		
RENAME	LOCK	SET Constraint		

Open-Source IDEs: Your Data Workspace

What’s an IDE?

An IDE, or Integrated Development Environment, enables programmers to consolidate the different aspects of writing a computer program.

IDEs increase programmer productivity by combining common activities of writing software into a single application: editing source code, building executables, and debugging.

Why you should use an IDE:

Enhanced Productivity:

Code Completion and Suggestions
Syntax Highlighting and Error Checking
Automated Refactoring

Powerful Debugging Tools

Project Management and Organization:

Project Structure Views
Build Automation

Collaboration and Version Control

Additional Benefits:

Code Templates
Customization
Extensions and Plugins

Top Open Source IDE Choices for Data Enthusiasts/Pythonists (personal opinion)

Spyder is a free and open source scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It features a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

Visual Studio Code was first announced on April 29, 2015 by Microsoft at the 2015 Build conference. A preview build was released shortly thereafter.

On November 18, 2015, the source code of Visual Studio Code was released under the MIT License and made available on GitHub. Extension support was also announced. On April 14, 2016, Visual Studio Code graduated from the public preview stage and was released to the web. Microsoft has released most of Visual Studio Code's source code on GitHub under the permissive MIT License, while the editor itself is distributed by Microsoft as proprietary freeware.

Virtual Environments!!!

Using virtual environments in the process of work with Python projects will help you avoid system pollution, sidestep dependency conflicts, minimize reproducibility issues, and dodge installation privilege lockouts.

https://realpython.com/python-virtual-environments-a-primer/#why-do-you-need-virtual-environments

Personally I prefer to use **pyenv-virtualenv** combination. You can read about it more here:

https://realpython.com/intro-to-pyenv/
https://github.com/pyenv/pyenv-virtualenv

Time for Demo

Essential Python Libraries for Data Professionals

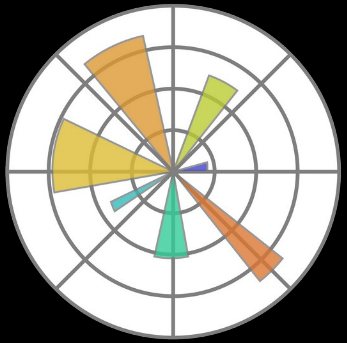
Data Manipulation and Analysis

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Time for Demo

Data Visualization



Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Data Manipulation and Analysis

Plotly is an open-source module of Python that is used for data visualization and supports various graphs like line charts, scatter plots, bar charts, histograms, area plots, etc. Plotly produces interactive graphs, can be embedded on websites, and provides a wide variety of complex plotting options.

D3.js is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of Scalable Vector Graphics, HTML5, and Cascading Style Sheets standards. It is the successor to the earlier Protovis framework.

https://observablehq.com/@d3/epicyclic-gearing?intent=fork

Time for Demo

Mathematical Modeling, Machine Learning, etc

SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.

SymPy is an open-source Python library for symbolic computation. It provides computer algebra capabilities either as a standalone application, as a library to other applications, or live on the web as SymPy Live or SymPy Gamma.

scikit-learn (formerly scikits.learn and also known as sklearn) is a free and open-source machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

https://scikit-learn.org/stable/user_guide.html

Mathematical Modeling, Machine Learning, etc

statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator. The results are tested against existing statistical packages to ensure that they are correct. The package is released under the open source Modified BSD (3-clause) license. The online documentation is hosted at statsmodels.org.

fbrophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

Mathematical Modeling, Machine Learning, etc

TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. It was developed by the Google Brain team for Google's internal use in research and production.

PyTorch is a machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, originally developed by Meta AI and now part of the Linux Foundation umbrella. It is recognized as one of the two most popular machine learning libraries alongside TensorFlow, offering free and open-source software released under the modified BSD license. Although the Python interface is more polished and the primary focus of development, PyTorch also has a C++ interface

Mathematical Modeling, Machine Learning, etc

Keras is an open-source library that provides a Python interface for artificial neural networks. Keras was first independent software, then integrated into the TensorFlow library, and later supporting more. "Keras 3 is a full rewrite of Keras as a low-level cross-framework language to develop custom components such as layers, models, or metrics that can be used in native workflows in JAX, TensorFlow, or PyTorch — with one codebase." Keras 3 will be the default Keras version for TensorFlow 2.16 onwards, but Keras 2 can still be used.

Time for Demo

Conclusion

Before we go 127.0.0.1 (home) let's make a pact. Let's commit to continuous learning, bold experimentation, and using every tool at our disposal to achieve our goals. The world is changing rapidly, and so can we. Let's make tomorrow a day of action, a day of growth, a day where we push our boundaries. The future is yours to create!