CERN openlab

CERN SCIENCE FOR OPEN DATA

CERN Openlab Summer Student Lightning Talk

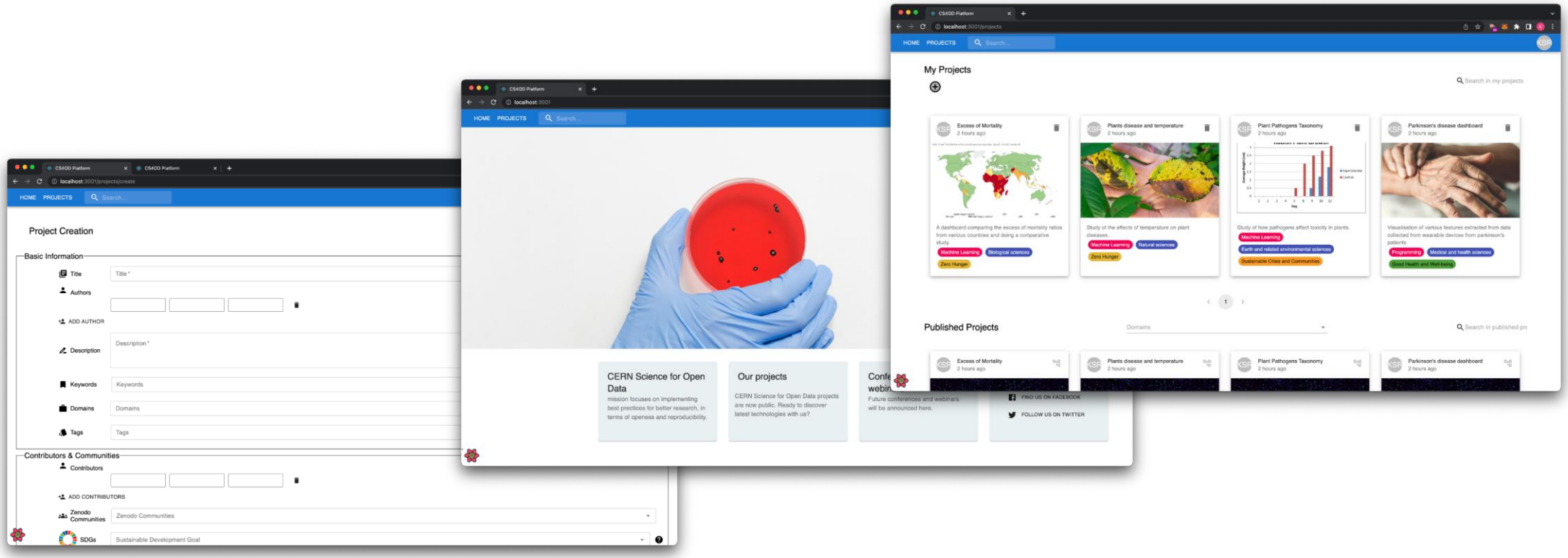
Kumar Saurabh Raj | Supervisors - Anna Ferrari, Paolo Tedesco

15/09/2022

What is CS4OD?

Introduction

CS4OD or CERN Science for Open Data is a web platform where researchers can share their work in a reproducible manner and collaborate effectively.



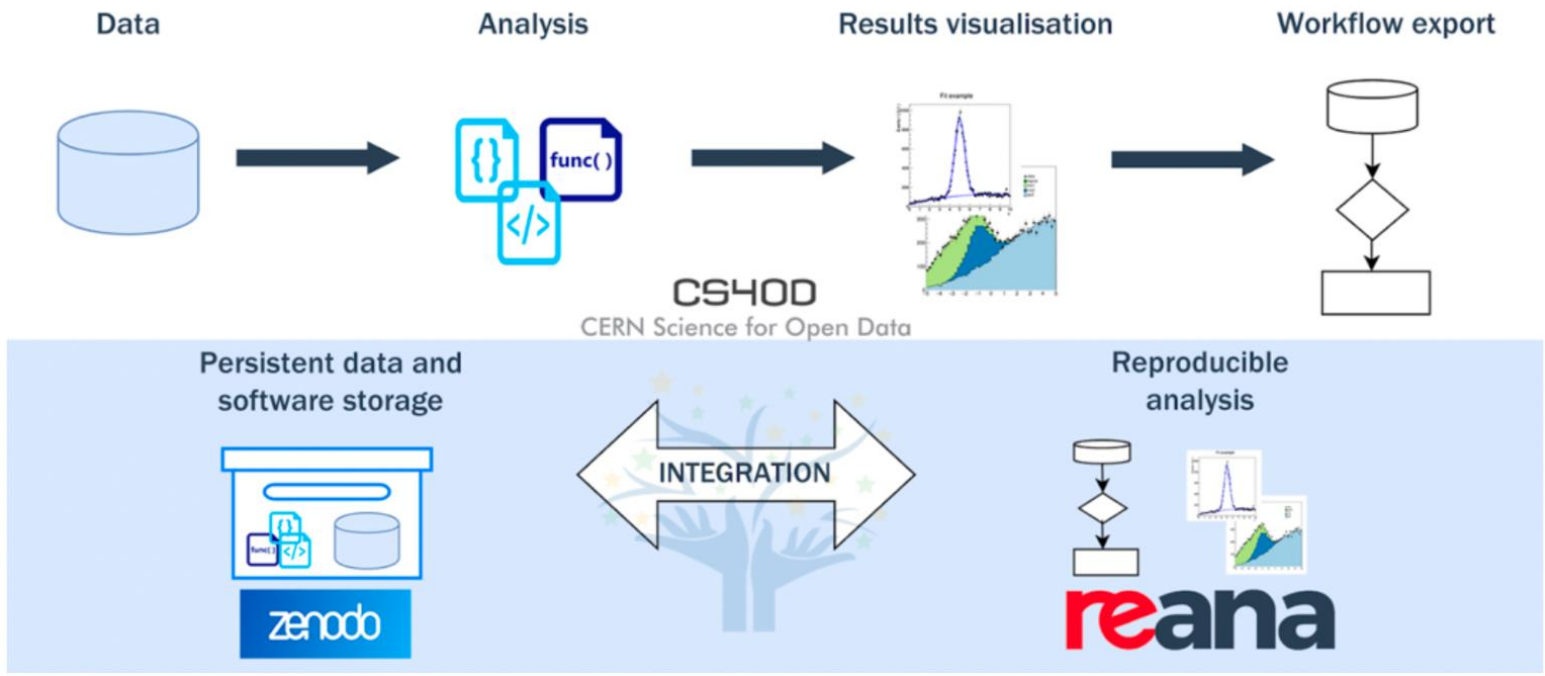
What is CS4OD?

Key Features / Highlights

- Access open research content - guides researchers to create, version and share their research work and leverage research content from all around the world to accelerate their discoveries.
- Create, Share and Fork research projects - while extending work done by others, starting from a fork speeds things up significantly.
- Publish research output persistently and rely on persistent research work from others - integrated with Zenodo to retrieve research content reliably
- Provides a systematic approach for general users to follow and guides them to research best practices.
- Builds on top of already existing open tools like REANA, SWAN and more which are powerful but remain inaccessible to most of the research community as their adoption in everyday research requires significant additional effort and a wide range of prerequisite knowledge.
- At a large scale, it brings forth a new and effective way for researchers from different fields and backgrounds to communicate and collaborate.

What is CS4OD?

Under the Hood

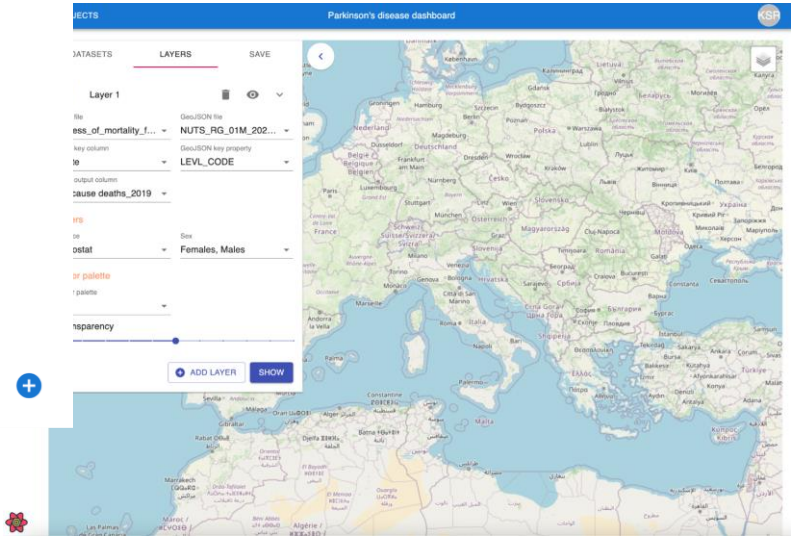


- Fills the gap between researchers with different background and cutting edge technology.
- Brings all stages of research, from data processing, analysis to visualisation on a single platform, abstracts away the complexities of using these tools individually.

Visualizations

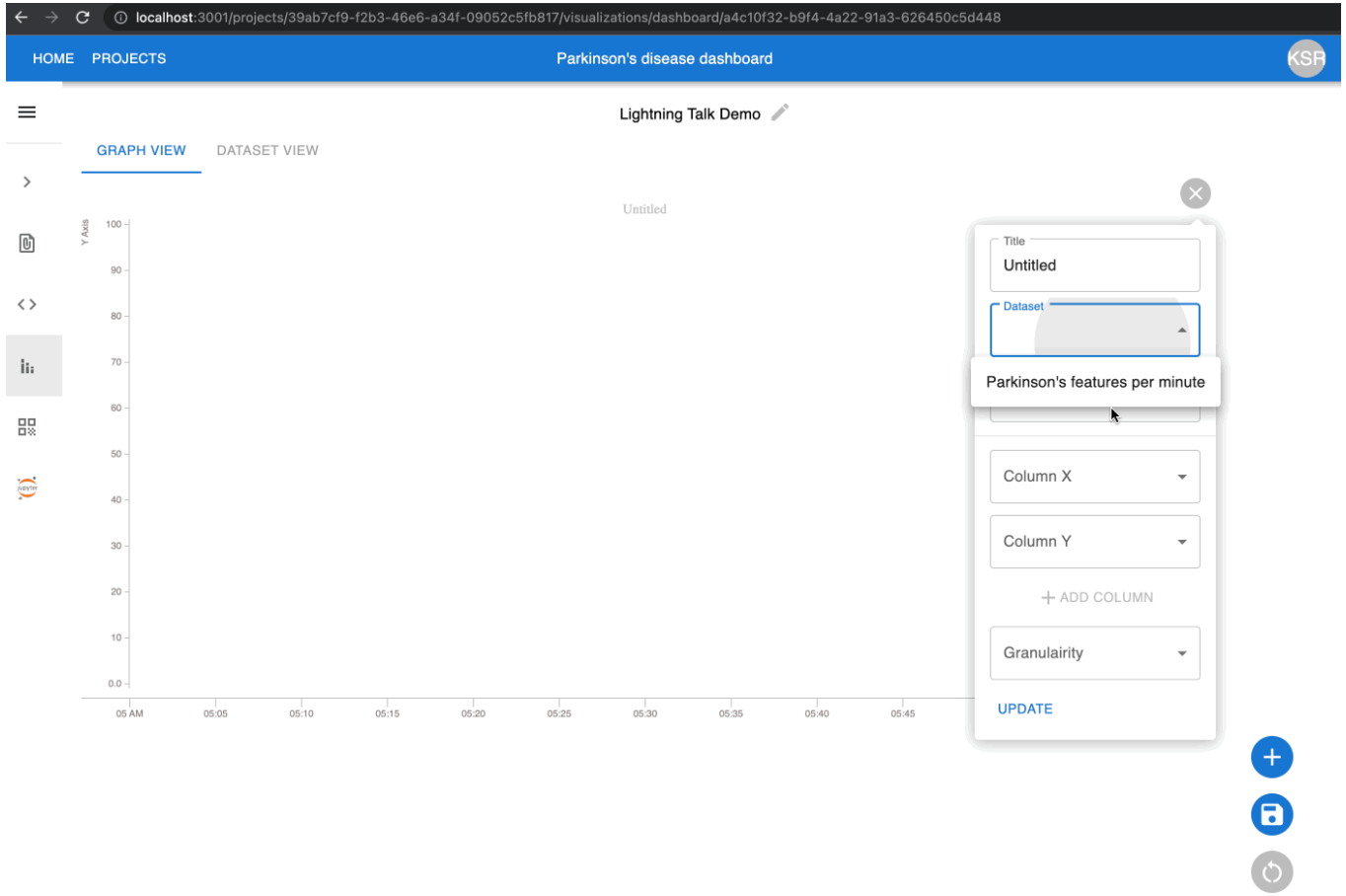
Support for new types of visualizations

- Visualizations in CS4OD are a quick way to understand datasets better and draw insights. However, the visualization capabilities of CS4OD were limited.
- Over 20 new ways to visualize data added
- support for categorical, time-series, cross-sectional and geo-location data



Visualizations

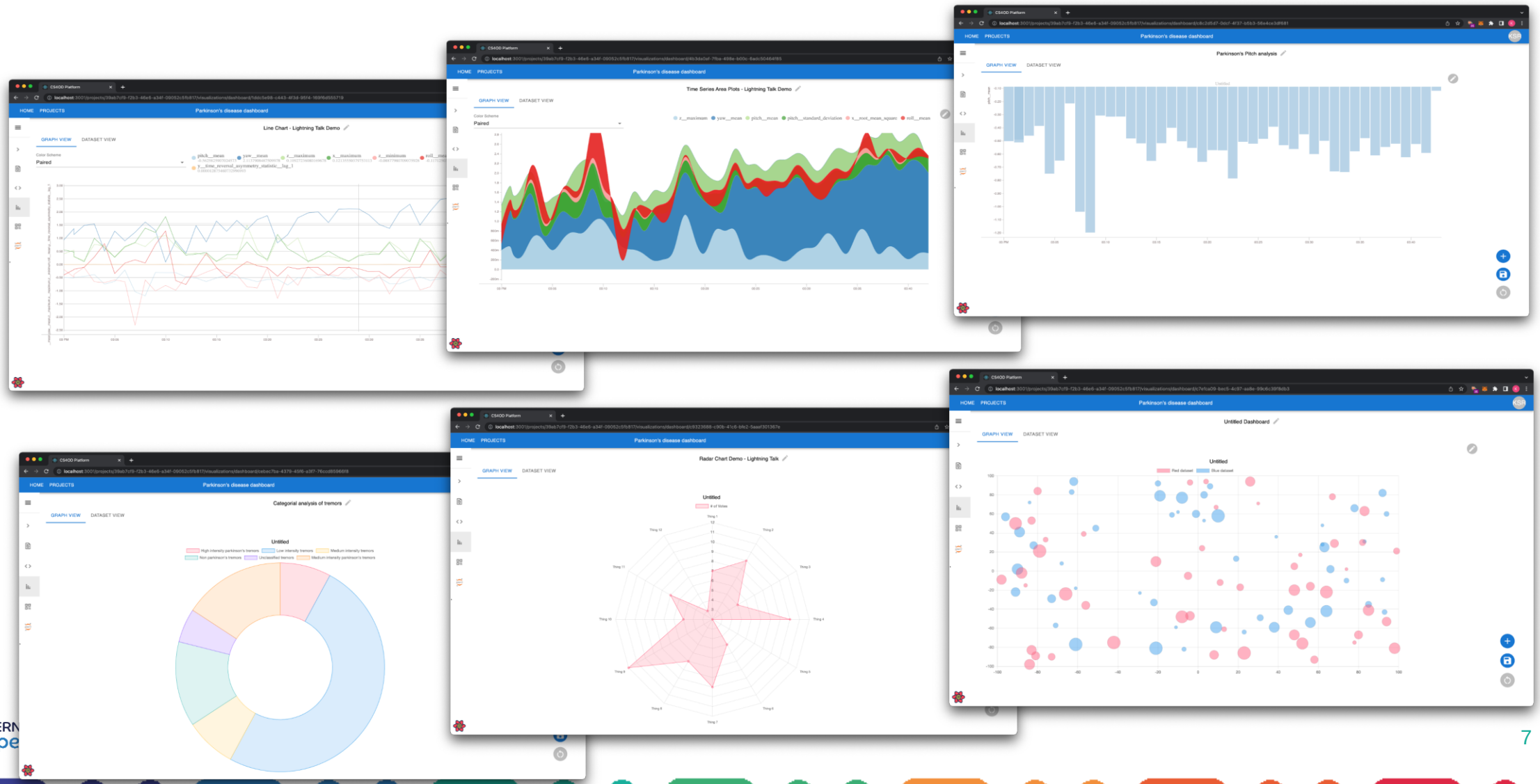
How to create one?



- The setup parses the selected dataset and allows selections of columns to plot.
- Allows selection of granularity (more on this later…)
- Select multiple Y-cols to stack graphs of multiple values against an axis
- Add multiple graphs to a dashboard
- View the dataset corresponding to the graphs in the dataset tab.

Visualizations

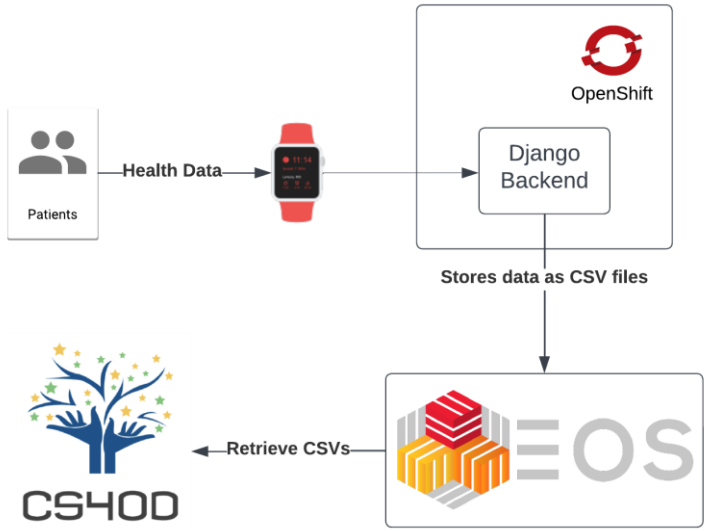
Some interesting visualizations



Use Case: Parkinson's Disease Visualization

- Providing visualizations tools to help caregivers understand the development of the Parkinson's disease during the patient's daily life (between visits).
- Real data was collected from Kuranos team and features were extracted.
- The nature of the data presented unique challenges.

Traditional visualization approach



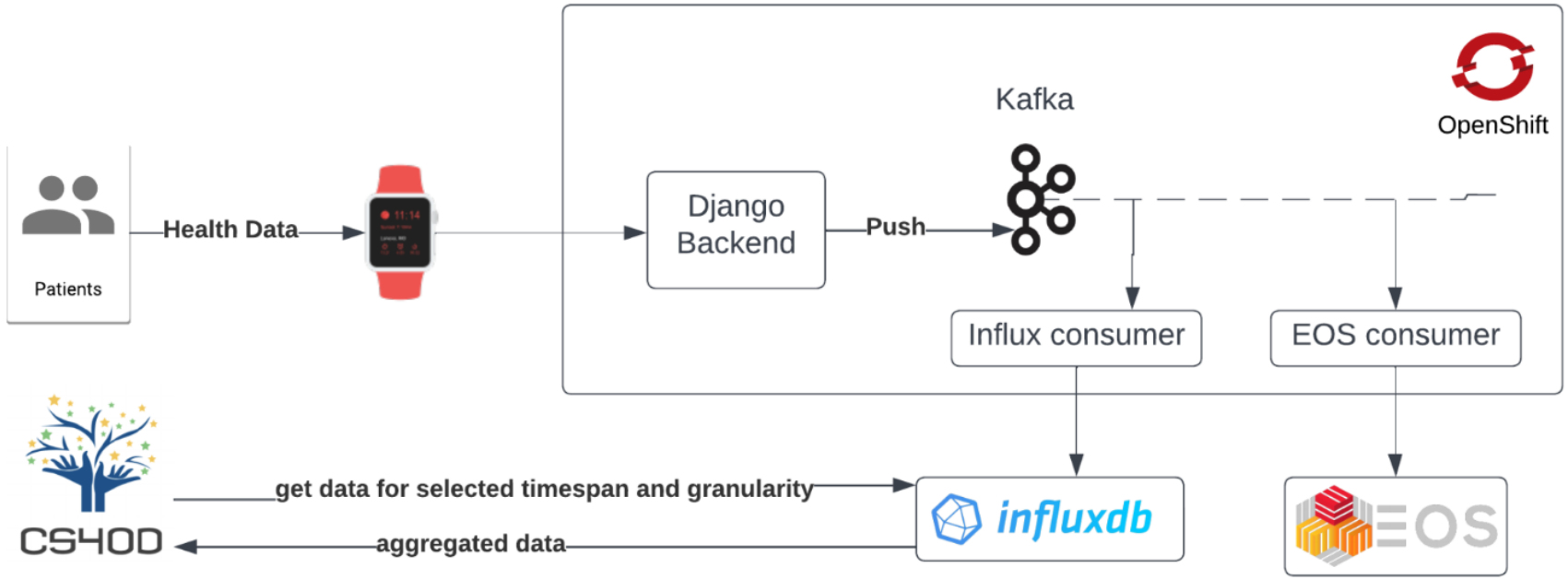
- Health data is collected from a smart watch at a sampling data between 20 to 50 Hz. Hence 20 to 50 data points are generated in a second.
- This time series data is continuously being generated, every second, minute and hour.
- To plot the data, CS4OD downloads the entire CSV file then plots it.
- For even moderately large files, frontend crashes trying to parse and visualize it.

Use Case: Parkinson's Disease Visualization

- However all the data might not be required to plot the visualizations!!!..

Enter Time Series Databases

- If the database has 50 records in every seconds and the user is trying to plot a time-range spanning over an year, then the data aggregated for each day might suffice. Based upon the granularity level the user selects, aggregation over different window sizes is performed to reduce the size of



Future Work

- Adding support for different types of databases.
- Add more advanced and customizable visualizations.

Thank you

QUESTIONS?

kumar.saurabh.raj@cern.ch

kumarsaurabhraj.sr@gmail.com
Linkedin - https://www.linkedin.com/in/kumarsaurabhraj/