posit::conf(2024)

Data Science Workflows With Posit Tools – Python Focus

August 12, 2024

Sam Edwardes & Michael Beigelmacher

Introduction

Data Science Workflows With Posit Tools — Python Focus

Logistics

- **Wifi:**
  - Network: Posit Conf 2024
  - Password: conf2024
- There are **gender-neutral bathrooms** located on levels 3, 4, 5, 6 & 7
- There is a **meditation/prayer room** is located in 503. Available Mon & Tues 7am - 7pm, and Wed 7am - 5pm.
- The **lactation room** is located in 509, same timings as above.
- Participants who do not wish to be **photographed have red lanyards**; please note everyone's lanyard colors before taking a photo and respect their choices.

Code of Conduct

- Everyone who comes to learn and enjoy the experience should feel welcome at posit::conf. Posit is committed to providing a professional, friendly and safe environment for all participants at its events, regardless of gender, sexual orientation, disability, race, ethnicity, religion, national origin or other protected class.
- The Code of Conduct and COVID policies can be found at Code of Conduct - Posit. Please review them carefully. You can report Code of Conduct violations in person, by email, or by phone. Please see the policy linked above for contact information.

Meet the team

Introduce yourself to your neighbors!

Howdy neighbour!

What you will learn

An opinionated end-to-end data science workflow

- Reading data
- Tidy data
- Data validation
- Automation
- Alerting
- Model development
- Model deployment
- Model alerting
- Application development and deployment
- Environment management
- Interoperability

We will use a combination of:

- **Open source** tools (developed by Posit and others)
- **Posit's professional products** (Workbench, Connect, and Package Manager)

Agenda

Time	Activity
09:00 - 10:30	● Introduction ● Environment setup ● Virtual environments ● Reading data
10:30 - 11:00	*Coffee break*
11:00 - 12:30	● Data validation ● Model training
12:30 - 13:30	*Lunch break*
13:30 - 15:00	● Model deployment ● Model monitoring
15:00 - 15:30	*Coffee break*
15:30 - 17:00	● Shiny app ● Better practices ● Wrap up

The sticky situation

I'm lost/need help

I'm done and ready to move along

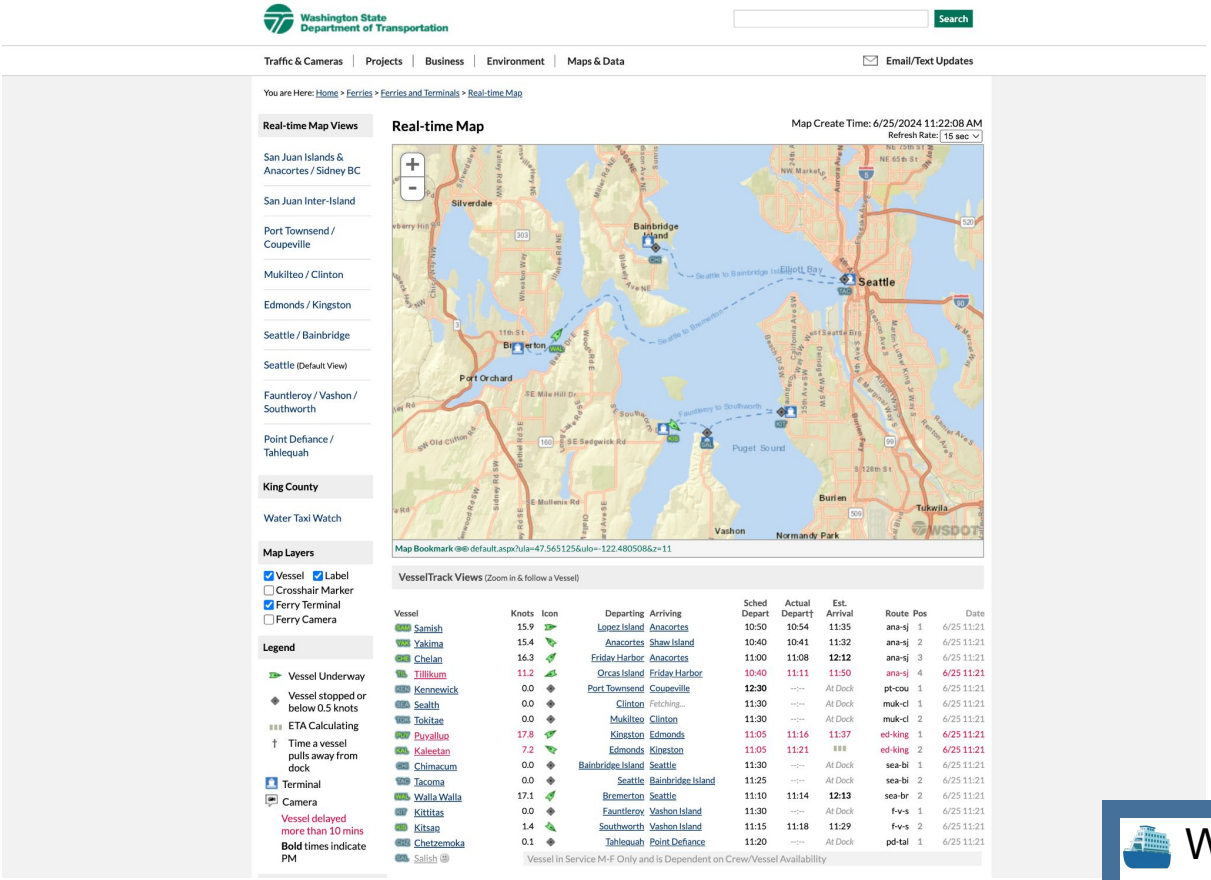
Put them up on the back of your laptop screen.

Asking Questions

We are using **GitHub Discussions**

https://github.com/posit-conf-2024/ds-workflows-python/discussions

Washington State Ferry Delays Project



https://wsdot.com/ferries/vesselwatch/default.aspx

https://i.pinimg.com/originals/c9/8b/3a/c98b3a997df52b6c8ad681590557c6bc.jpg

🚢 WSF is the largest operating public ferry system in the US! How cool is that?

🐳 21 ferries across Puget Sound and the greater Salish Sea.

Washington State Ferry Delays Project

Question

Can we predict when ferries will be delayed, and for how long?

Our Approach

Use historical delay, vessel, and weather data to create a model that will predict the duration of delays!

Project Objective:

Provide users with a self-service tool that predicts the duration of a delay.

Project Requirements:

- 🤖 Automate the pipeline
- 🔄 Project is easy to maintain and iterate upon
- 📁 Work is reusable by other teams, even if they don't use Python

Washington State Ferry Delays Project

This project/workshop has three major modules:

Understanding data
Reading data
Data validation
Saving data

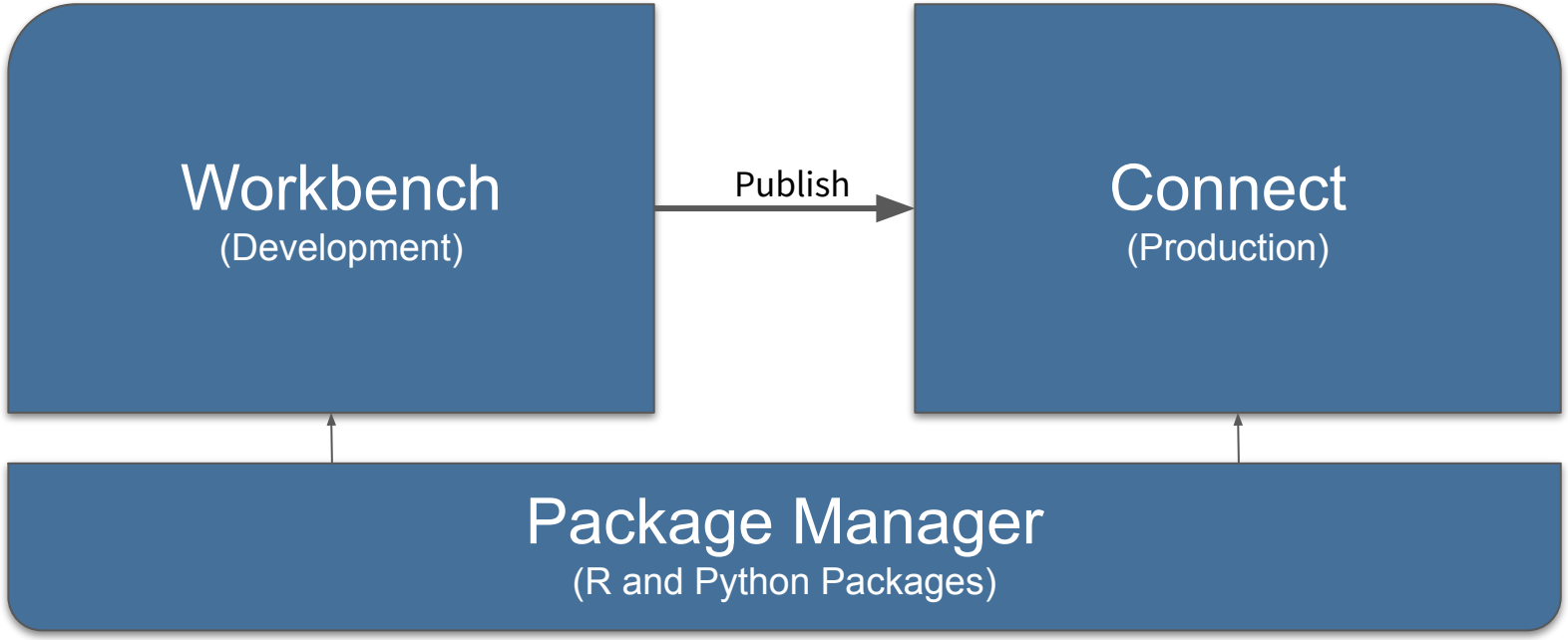
Data modelling
Model development
Model deployment
Model monitoring

Data presentation
Dashboard development

Environment setup

Login into Posit Workbench and Connect

Posit Team


Workbench
(Development)
Publish
Connect
(Production)
Package Manager
(R and Python Packages)

Access your tools

WIFI credentials:

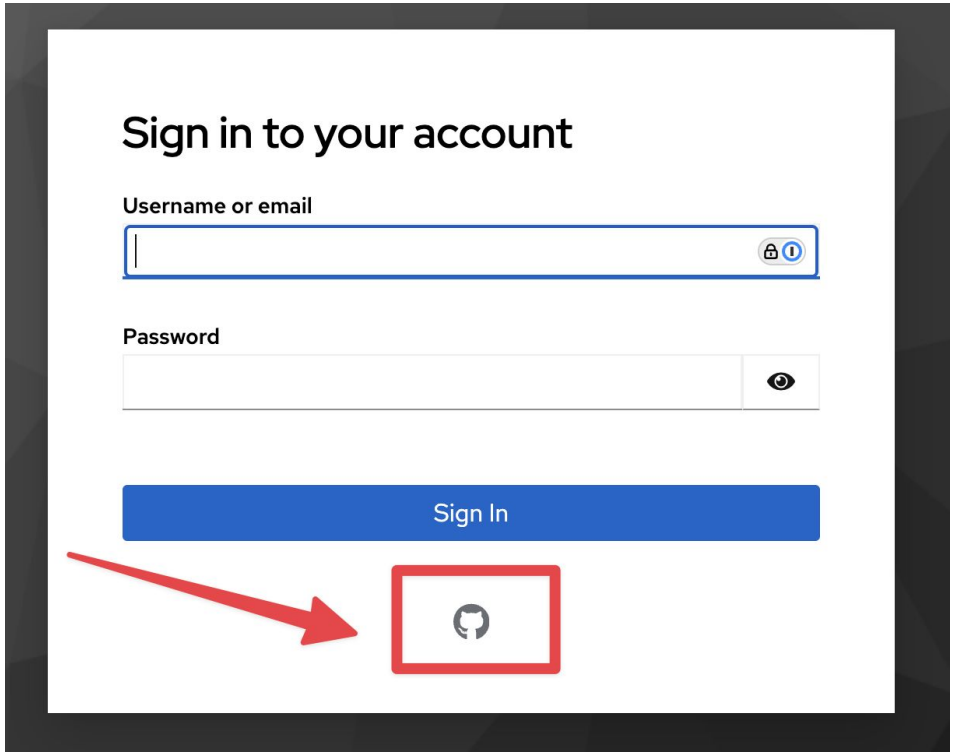
- Network: `Posit Conf 2024`
- Password: `conf2024`

Project landing page (bookmark this):

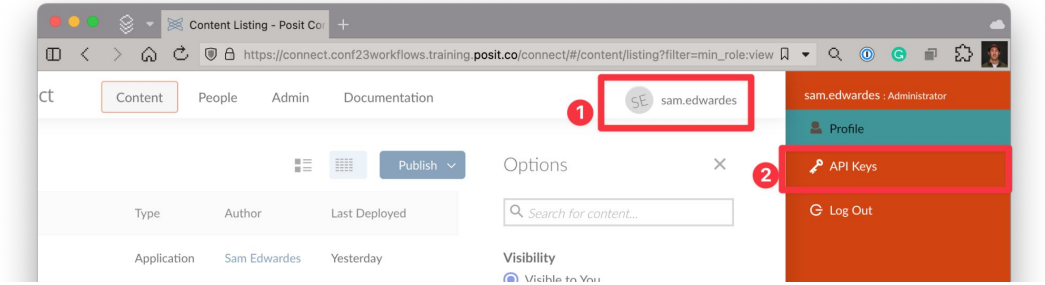
- https://github.com/posit-conf-2024/ds-workflows-python

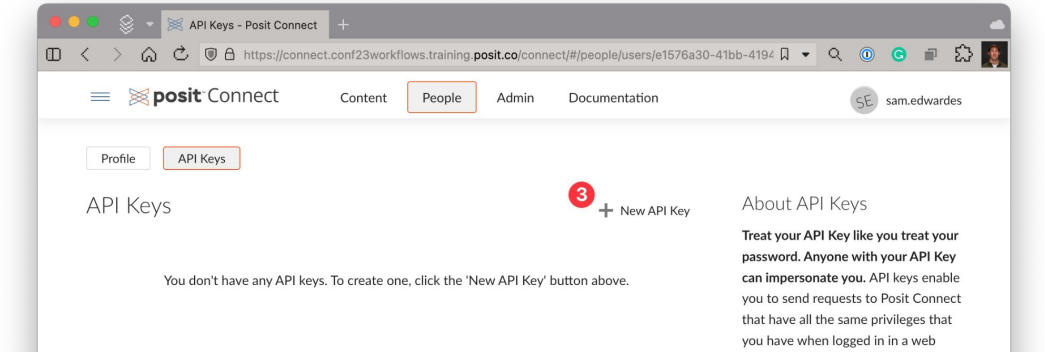
Workbench // Setup

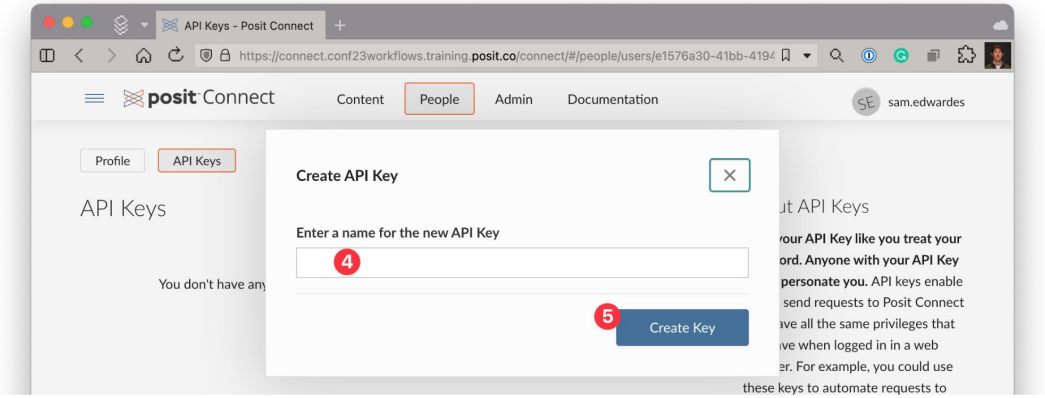
1. Sign into Workbench with GitHub
2. Start a VS Code session, name it “Setup”
3. Open a project from Git: Press **cmd** + **shift** + **p** (Mac) or **ctrl** + **shift** + **p** (Windows), then type “Git: Clone”. Enter this URL: https://github.com/posit-conf-2024/ds-workflows-python.git. Then press “Open”.
4. Open the terminal (“Terminal: Create New Terminal”), and then run `/bin/bash init.sh`. Wait for the script to finish running.



Connect // Create an API key







- You can name the API key anything you want, for example “workbench”.
- Save the API key that is generated to somewhere you can find again.
- **Remember to save your API Key! We will need it in a few minutes.**

Workbench // Secrets

Enter your personal secrets in the ~/.bashrc file ("File: Open File...").

- CONNECT_API_KEY
- WSDOT_ACCESS_CODE

```
# ------------------------------------------------------------
# Added for Posit Worskshop
# ------------------------------------------------------------

# Secrets
export CONNECT_API_KEY='xxx'
export WSDOT_ACCESS_CODE='xxxx'
```

Click here to get your WSDOT_ACCESS_CODE: https://wsdot.wa.gov/traffic/api/

Project File Structure

GitHub Repo:
https://github.com/posit-conf-2024/ds-workflows-python

Repo layout:

```
.
├── LICENSE.md
├── materials
│   ├── 01-reading-data
│   ├── 02-data-exploration-and-validation
│   ├── 03-model-training
│   ├── 04-model-monitoring
│   ├── 05-shiny-app
│   ├── 06-bonus-stuff
│   └── README.md
└── README.md
```

01 Reading data

Reading data from raw data sources

- Virtual environments with uv
- Use httpx to query external APIs
- Convert the data to tabular form using polars
- Save the data to a Postgres database

🖥️ Activity #1

☕ Coffee break

30 minute break

02 Data Exploration Tidying & Validation

Exploration, Tidying and validating raw data

- Read the raw data from SQL using polars
- Data exploration using polars
- Tidy the data using polars
- Validate the data using Pandera
- Write the validated data to SQL

🧑‍💻 Activity #2

03 Model building and Deployment

Model Operations & tasks

- Building a machine learning model
- Using the Vetiver package for MLOps
- Deploying model on Posit Connect

Lunch break

60 minute break

04 Model monitoring & Model Card

How to keep your model healthy

- Monitor model performance using vetiver
- Build and deploy a model card

🧑‍💻 Activity #4

☕ Coffee break

30 minute break

05 Shiny App

Sharing your work with others

- Create a shiny app so that non technical users can use your work

🧑‍💻 Activity #5

>

Better practises

<

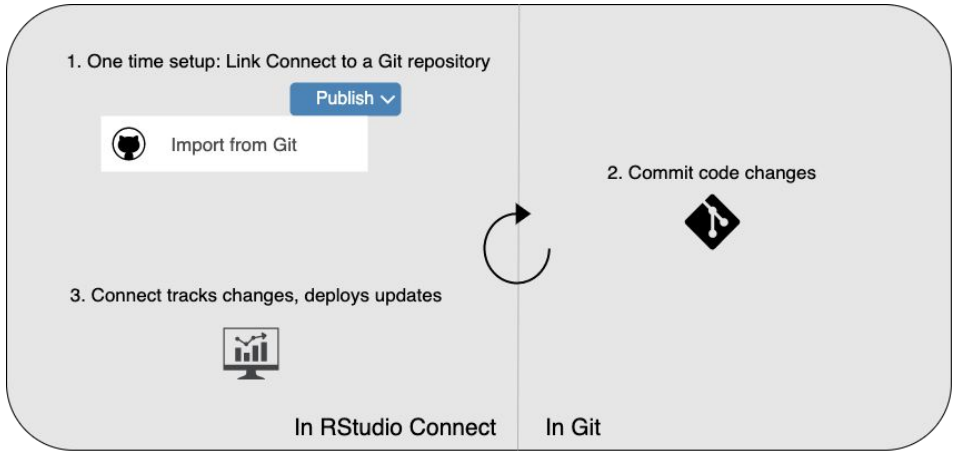
Connect: Integrate with version control

Connect lets you deploy directly from git repository. This works well for promoting content between dev, test, and production:

- https://docs.posit.co/connect/user/git-backed/
- https://solutions.posit.co/operations/code-promotion/

🧑‍💻 Activity

Deploy one of the outputs you created today using Git-Backed deployment.



Connect: Use the Connect API

Connect has an API that allows your to control almost everything programmatically. There are three primary ways to use it:

1. HTTP requests: https://docs.posit.co/connect/api/
2. posit-sdk (Python): https://github.com/posit-dev/posit-sdk-py/
3. connectapi (R): https://pkgs.rstudio.com/connectapi/

Ideas

- Publish via GitHub actions
- Update content from within another content item (e.g. press a button in Shiny to update a pin)
- Check out the cookbook for many more ideas! https://docs.posit.co/connect/cookbook/

🧑‍💻 Activity

Install the posit-sdk and get usage stats for your Shiny app: https://docs.posit.co/connect/cookbook/user-activity/.

>

Wrap up

<

Takeaways

- Understanding of a typical data science workflow in Python
- Introduction to pro tools
  - Workbench for writing code
  - Connect for sharing data products
  - Package Manager for hosting packages
- Introduction to open source tools
  - polars for working with tabular data
  - pydantic for data validation
  - vetiver for model deployment and monitoring
  - shiny for interactive apps
- Everything we learned today will always be available at https://github.com/posit-conf-2024/ds-workflows-python

How could we make this workflow better?

- A python package to encapsulate re-usable logic
- Deploy content programmatically
- Multiple models for better predictions
- Proactive monitoring based on vetiver metrics

Workshop Survey

Please go to https://pos.it/conf-workshop-survey

Your feedback is crucial! Data from the survey informs curriculum and format decisions for future conf workshops, and we really appreciate you taking the time to provide it.

posit™

Thank you.

Appendix