Python Tools to Deploy Your Machine Learning Models Faster 🛩️

Jeff Hale

gradio

Streamlit

FastAPI

Welcome Aboard!

Pilot: Jeff Hale

- linkedin.com/in/-jeffhale/
- twitter@discdiver
- jeffhale.medium.com/

Preflight check

What’s one thing you’re hoping to learn today?

Planes

gradio

Streamlit

FastAPI

Flight plans

- Test flights for each plane
  - Hello world
  - Show data
  - Plot
  - Model inference
- Cruising altitude & turbulence (pros & cons)
- Grab your luggage (takeaways)
- Disembark

Flight plan materials

- GitHub repo: https://github.com/discdiver/dsdc-deploy-models
- Slides: https://www.slideshare.net/JeffHale6/getting-a-job-in-data-science-what-and-how-to-learn

Gradio

Ultralight

New, quick to fly, experimental

Gradio Demo 1: Hello world

Gradio #1: Hello world

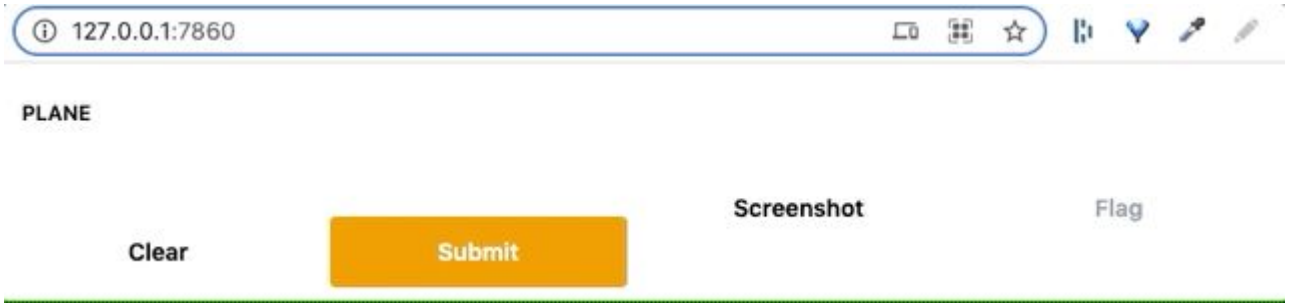
```
import gradio as gr


def hello(plane):
    return f"I'm an ultralight {plane} 🛩️"


iface = gr.Interface(
    fn=hello,
    inputs=['text'],
    outputs=['text']
).launch()
```

Gradio #1: Hello world

- *pip install gradio*
- *python gradio_hello.py*



Gradio Demo 2: Show data

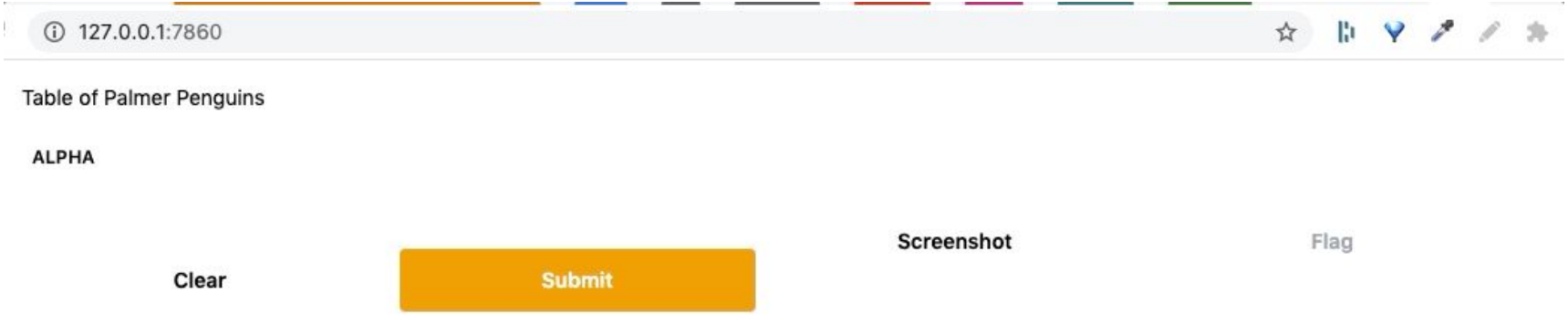
Gradio #2: Show me the data!

```
def show_pens(alpha):
    return pd.read_csv(
      'https://raw.githubusercontent.com/mwaskom/seaborn-data/master/penguins.cs
v')

iface = gr.Interface(
    fn=show_pens,
    inputs=['text'],
    outputs=[gr.outputs.Dataframe()],
    description="Table of Palmer Penguins"
).launch(share=True)
```

Gradio #2: Show me the data!

- *pip install gradio pandas seaborn*
- *python gradio_pandas.py*
- Glitch only allows 10 cells to show initially or fails silently as of Nov. 16, 2021 TK



Gradio Demo 3: Plotting

Gradio #3: Plot it

- Plotly doesn’t work as of 2.8.7 (targeted for 2.9) ☹️
- You can use Matplotlib as of Gradio 2.8.2

Gradio #3: Plot it

```
def plot_pens(place_holder):
    """scatter plot penguin chars using matplotlib"""
    df_pens = pd.read_csv(
"https://raw.githubusercontent.com/mwaskom/seaborn-data/master/penguins.csv"
    )
    fig = plt.figure()
    plt.scatter(x=df_pens["bill_length_mm"],
y=df_pens["bill_depth_mm"])
    return fig
```

Gradio #3: Plot it

```
iface = gr.Interface(
    fn=plot_pens,
    layout="vertical",
    inputs=["checkbox"],
    outputs=["plot"],
    title="Scatterplot of Palmer Penguins",
    description="Let's talk pens. Click to see a plot.",
    article="Talk more about Penguins here, shall we?",
    theme="peach",
    live=True,
).launch()
```

Scatterplot of Palmer Penguins

Let's talk pens. Click to see a plot.

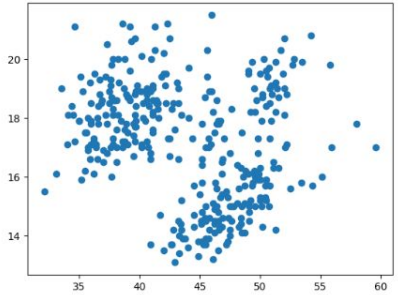
place holder

Clear

Submit

Output

0.3s



Flag

Talk more about Penguins here, shall we?

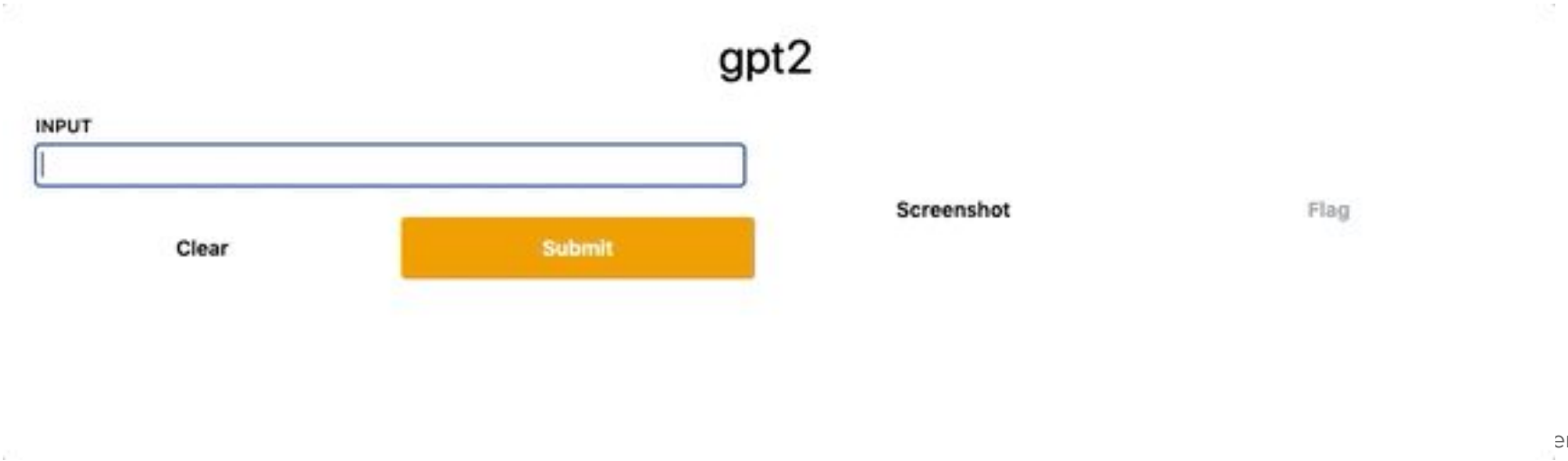
view the api • built with gradio

Gradio Demo 4: Model inference

Gradio #4: Model inference

```
import gradio as gr

gr.Interface.load('huggingface/gpt2').launch()
```



Gradio #4: Model inference - prettier

```
gr.Interface.load(
    "huggingface/gpt2",
    title="Storytelling with GPT2",
    css="""
        body {background: rgb(2,0,36);
              background: linear-gradient(
              180deg,
              rgba(2,0,36,1) 0%,
              rgba(7,51,99,1) 70%,
              rgba(6,3,17,1) 100%);}
        .title {color: white !important;}
        """,
).launch()
```

Gradio #4: Model inference - prettier

Gradio Data API - One Click!

Response:

```
{
  "data": [ Union[str, number] ],
  "durations": [ float ], // the time taken for the prediction to complete
  "avg_durations": [ float ] // the average time taken for all predictions so far (used to estimate the runtime)
}
```

Try it (live demo):

Python | cURL | Javascript

```
curl -X POST http://127.0.0.1:7860/api/predict -H 'Content-Type: application/json' -d '{"data": ["Hello World"]}'
```

Gradio Pros

- Quick demos for ML 🚀
- Built-in interpretability 🔎
- Auto-docs 📃
- Nice Hugging Face integration 🤗
- Bright future 😎

Gradio Cons

- Rough around the edges - early stage 🔪
- Not easy to customize elements ✨
- Single page only 1️⃣

Gradio - Hugging Face Acquisition

+

Abubakar Abid

Streamlit

Cessna Citation Longitude

Light, quick to takeoff, easy flying

Streamlit

Streamlit Demo 1: Hello world

Streamlit #1: Hello world

```
import streamlit as st

name = "Jeff"

st.title(f"Hello from Streamlit, {name}!")
```

- *pip install streamlit*
- *streamlit run streamlit_hello.py*

Hello from Streamlit!

Streamlit

Streamlit Demo 2: Show data

Streamlit #2: Show data

```
import streamlit as st
import pandas as pd

st.title("Streamlit with pandas")
show = st.checkbox("Show dataframe")
df_pens = pd.read_csv(
"https://raw.githubusercontent.com/mwaskom/seaborn-data/master/penguins.csv")

if show:
   df_pens
```

Streamlit #2: Show data

- *pip install streamlit pandas seaborn*
- *streamlit run streamlit_pandas.py*

Streamlit #2: Show data

Streamlit with pandas

☐ Show dataframe

Streamlit

Streamlit Demo 3: Plotting

Streamlit #3: Plotting

```
choice = st.radio("Select color", ["species", "island"])


fig = px.scatter_3d(
    data_frame=df_pens,
    x="bill_depth_mm",
    y="bill_length_mm",
    z="body_mass_g",
    color=choice,
    title="Penguins in 3D!",
)
fig
```

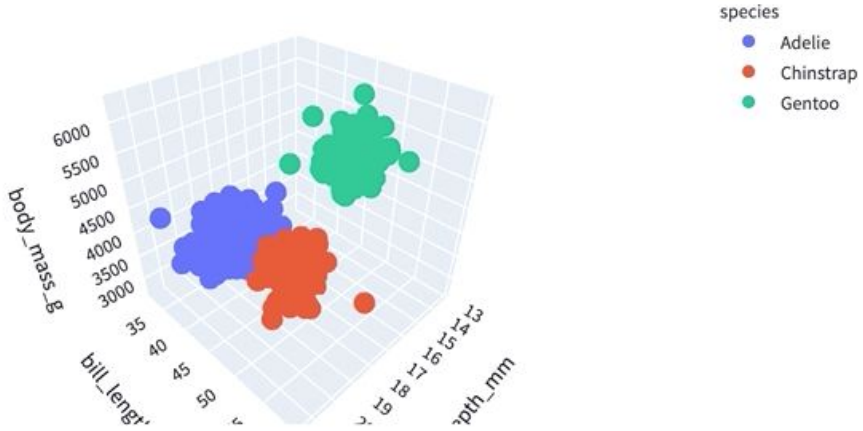
Streamlit #3: Plotting

Select color

species

island

Penguins in 3D!



Streamlit

Streamlit Demo 4: Model Inference

Streamlit #4: Model inference

```
import streamlit as st
from transformers import pipeline


st.header("GPT-2 Stories")
input_text = st.text_area("Enter your text here:")


generator = pipeline("text-generation", model="gpt2")
output = generator(input_text, max_length=100, )
output[0]["generated_text"]
```

Streamlit #4: Model inference

GPT-2 Stories

Enter your text here:

ok friends, let's talk

ok friends, let's talk

Riot - I'm on the board now. Please explain your message.

No one likes news on my channel.

Please put your email in the contact form. - Yes I have a problem

Streamlit #4: Model inference - prettier

GPT-2 Stories

Enter your text here:

ok friends, let's talk

ok friends, let's talk

Riot - I'm on the board now. Please explain your message.

No one likes news on my channel.

Please put your email in the contact form. - Yes I have a problem

Streamlit #4: Model inference - prettier

```
st.header("Story time")
st.image("https://cdn.pixabay.com/photo/2017/07/12/19/03/highway-2497900_960_720.jpg")
col1, col2 = st.columns(2)

with col1:
    input_text = st.text_area("Enter your text here:")
    with st.spinner("Generating story..."):
        generator = pipeline("text-generation", model="gpt2")
        if input_text:
            generated_text = generator(input_text, max_length=60)
            st.success("Here's your story:")
            generated_text[0]["generated_text"]

with col2:
    st.header("I'm in another column")
```

Streamlit #4: Model inference - prettier

Story time

Enter your text here:

ok smarty, here we are

Here's your story:

ok smarty, here we are.

There's also the recent news of IBM's (IBM) (IBM Inc) (IBM) (IBM) (IBM MS) (IBM) smartwatches and the new BlackBerry (R) in this week's Best

I'm in another column

Streamlit Serving Options

- Serve from Streamlit's servers for free: bit.ly/st-6plots
- Serve from Hugging Face or Heroku for free
- Pay Streamlit for more/better
- Host elsewhere

Streamlit Pros

- Quick websites for many Python use cases 🐍
- Many intuitive interactive widgets ✅
- Nice hosting options 🎁
- Thoughtful docs 📄
- Strong development cadence & team 💪

Streamlit Cons

- Some customizations cumbersome ✨
- Single page only (without hacks) 1️⃣

Streamlit Snowflake Acquisition

Recent acquisition by Snowflake

Image Source: Snowflake, Public domain, via Wikimedia Commons

FastAPI

Boeing 737

Commercial grade, fast, smart!

FastAPI

FastAPI Demo 1: Hello world

FastAPI #1: Hello world

```python
import uvicorn
from fastapi import FastAPI

app = FastAPI()

@app.get("/")
def home():
    return {"Hello world": "How's it going?"}

if __name__ == "__main__":
    uvicorn.run("fastapi_hello:app")
```

FastAPI #1: Hello world

pip install fastapi uvicorn

python fastapi_hello.py

Returns json

{"Hello world":"How's it going ?"}

FastAPI: Async

FastAPI #1: Hello world

```
@app.get("/not-async")
def home():
    return dict(zip(range(10), range(10)))
```

10ms

FastAPI #1: Hello world

```
@app.get("/yes-async")
async def home():
    return dict(zip(range(10), range(10)))
```

3ms

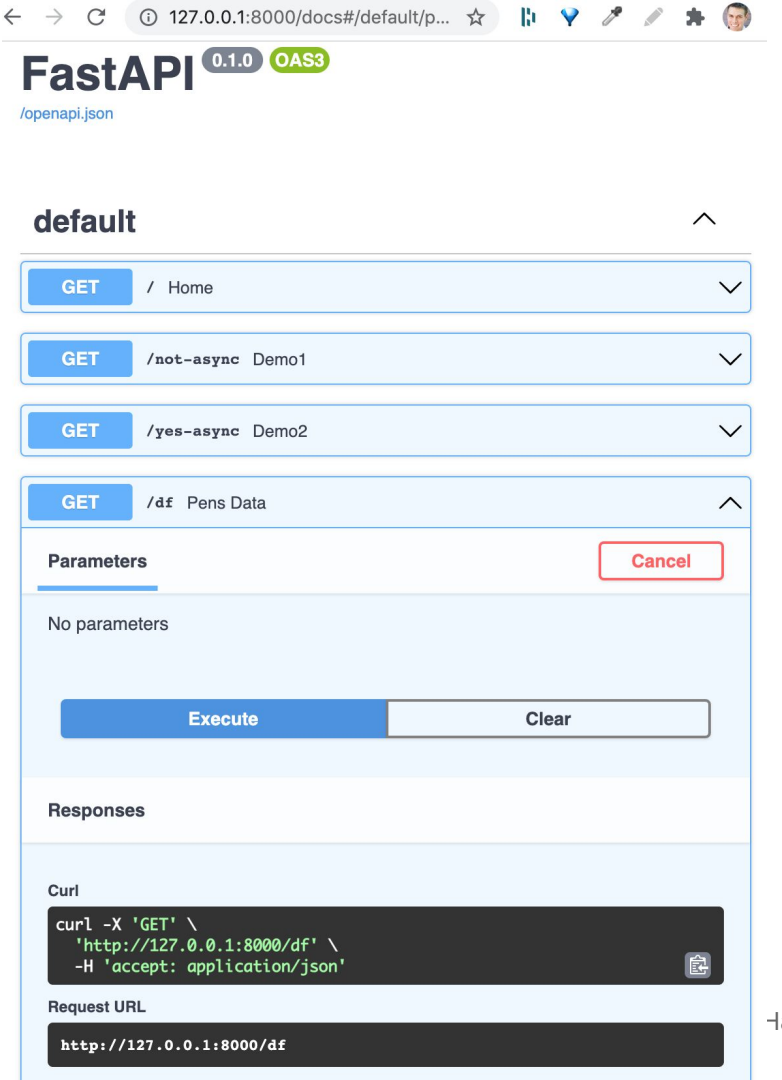
FastAPI

FastAPI Demo 2: Show data

FastAPI #2: Show me the data!

```
@app.get("/df")
async def pens_data():
    df_pens = pd.read_csv(
    "https://raw.githubusercontent.com/mwaskom/seaborn-d
    ata/master/penguins.csv")
    df_no_nans = df_pens.fillna(-1.01)
    return df_no_nans
```

Automatic docs



FastAPI #2: Show me the data!

Get back the whole DataFrame as json. Don’t need to convert. 😎

{"species":{"0":"Adelie","1":"Adelie"...

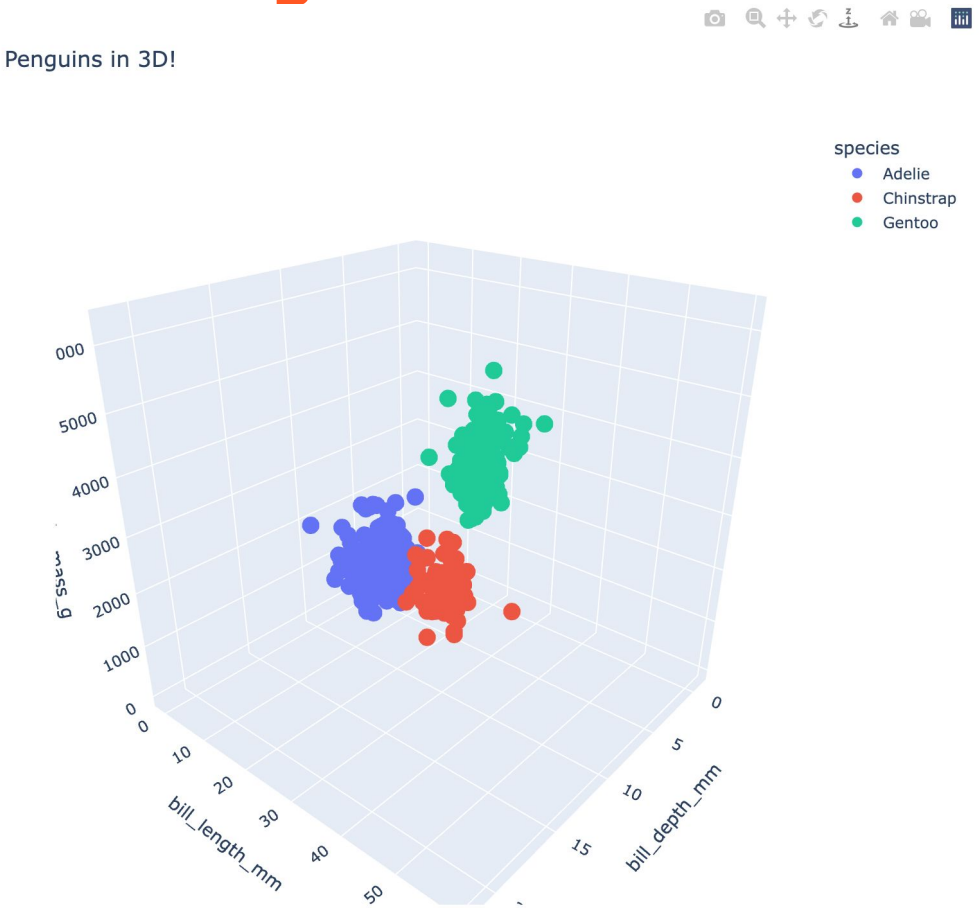
FastAPI

FastAPI Demo 3: Plotting

FastAPI #3: Plotting

```python
@app.get("/plot")
async def plot() -> HTMLResponse:
    """return a plotly plot"""
    fig = px.scatter_3d(
        data_frame=df,
        x="bill_depth_mm",
        y="bill_length_mm",
        z="body_mass_g",
        color="species",
        title="Penguins in 3D!",
    )
    return HTMLResponse(fig.to_html())
```

FastAPI #3: Plotting



FastAPI

FastAPI Demo 4: Model Inference

FastAPI #4: Model inference

Typing with mypy TK

FastAPI Pros

- Fastest Python API framework - async 🏎️
- Automatic API documentation ⭐
- Nice error messages ⚠️
- Extensive docs 📃
- Jinja templating ⚔️
- Nice test client 👌
- SQL Model integration ↔️

FastAPI Pros

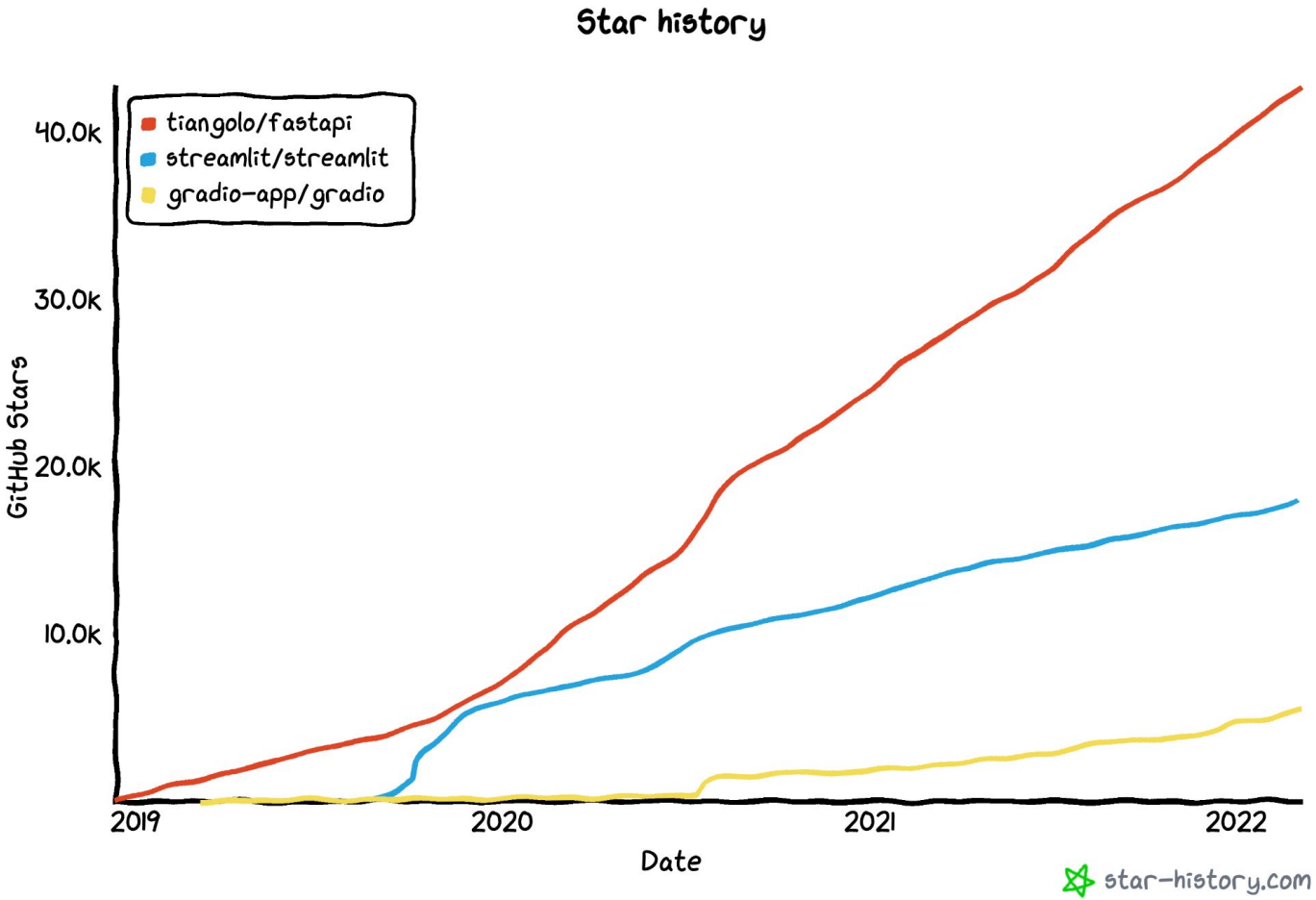
Sebastian Ramirez

FastAPI Cons

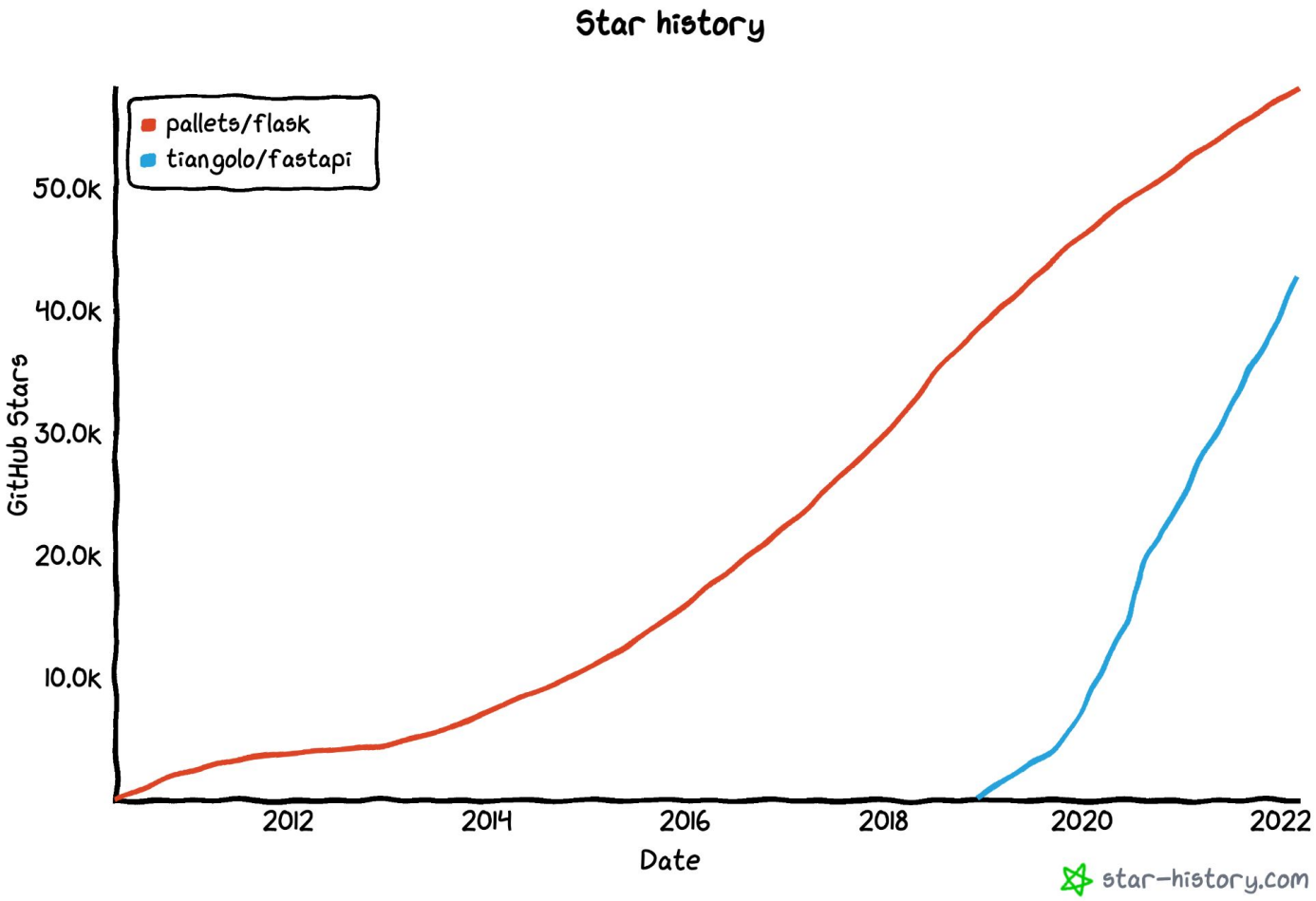
- Reliant on a single maintainer 😬
- Takes more code for HTML templating than Flask ⌨️

Summin' Up

	Web App	**Data API**
gradio	Yes	Yes
Streamlit	Yes	No
FastAPI	Yes (Jinja templates)	Yes



Source: https://star-history.t9t.io/#streamlit/streamlit&tiangolo/fastapi&gradio-app/gradio



Source: https://star-history.com/#tiangolo/fastapi&pallets/flask&Date

Grab Your Luggage (takeaways)

Use what you know, unless it doesn’t meet your needs.

Blank slate?

Learn what's popular, growing, and quick to get off the ground.

Streamlit

For single-page app that doesn't need custom styling.

Gradio for quick 🤗 models for fun.

FastAPI

FastAPI for serving data.

What to learn next?

(Newer to Python)

Streamlit

Disembark

Thank you for flying the deployment skies!

Jeff Hale

- linkedin.com/in/-jeffhale/
- @discdiver
- jeffhale.medium.com/

Versions used

- Gradio 2.8.7
- Streamlit 1.5.1
- FastAPI 0.75.0

Python Tools to Deploy Your Machine Learning Models Faster 🛩️

Jeff Hale

gradio

Streamlit

FastAPI