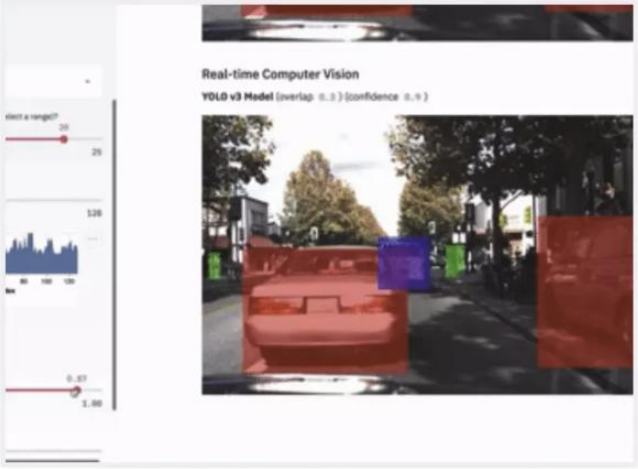
数据挖掘与商务分析

第6讲 模型应用开发与部署

主讲教师：肖升生

课程导入：模型应用开发与部署需求

- Turn Data Scripts into Web Apps
- Interactive
- No Frontend Experience Required
- Easy to Deploy


Real-time Computer Vision
YOLO v3 Model (overlap 0.3) (confidence 0.9)
elect a range?
25
128
0.87
1.00

Real time object detection

An image browser for the Udacity self-driving-car dataset with real-time object detection.

See on GitHub

讲授提纲

01 数据的探索性分析

02 可视化

03 Streamlit 工具介绍

04 模型开发与部署

05 动手实践

讲授提纲

01 数据的探索性分析

02 可视化

03 Streamlit 工具介绍

04 模型开发与部署

05 动手实践

什么是数据的探索性分析？

- EDA, Exploratory Data Analysis, **数据探索性分析**
  - **重点是可视化**
  - 聚类和异常检测被视为探索性技术
  - 在数据挖掘中聚类和异常检测

- 探索性分析中，我们重点关注：
  - 数据的描述性统计分析
  - 可视化

- 数据、模型可视化与工程化部署

示例：IRIS 数据集

- Many exploratory data techniques are nicely illustrated with the iris dataset.
  - Dataset created by famous statistician Ronald Fisher
  - 150 samples of three species in genus *Iris* (50 each)
    - *Iris setosa*
    - *Iris versicolor*
    - *Iris virginica*
  - Four attributes
    - sepal width
    - sepal length
    - petal width
    - petal length
  - Species is class label

Iris virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

汇总统计

Summary statistics

- location - mean, median
- spread - standard deviation, variance, range
- frequency and Mode
- percentiles, mean

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = \mathit{median}\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

讲授提纲

01 数据的探索性分析

02 可视化

03 Streamlit 工具介绍

04 模型开发与部署

05 动手实践

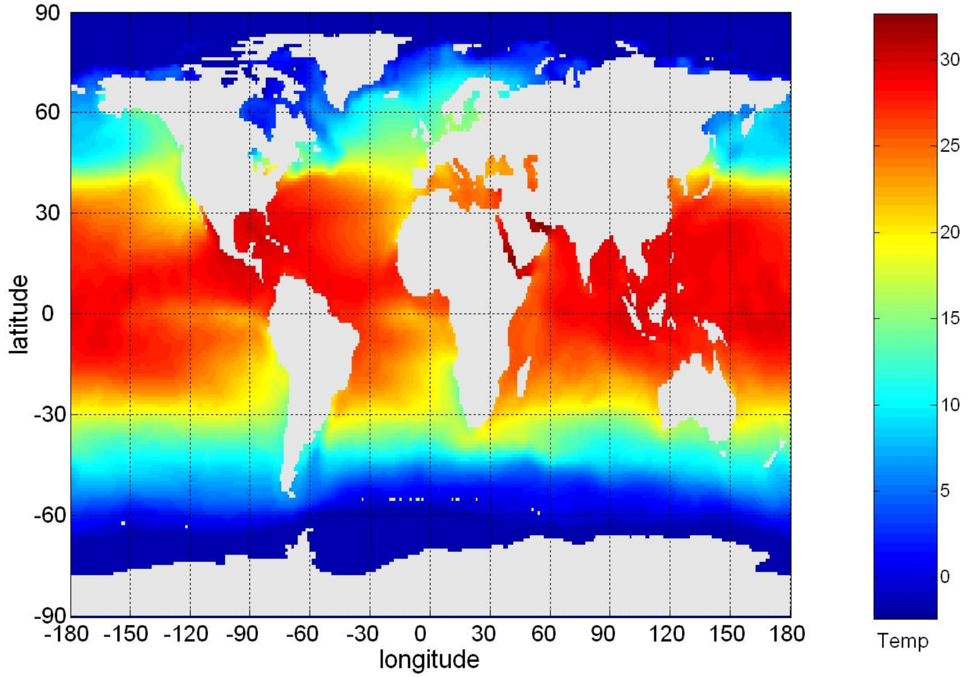
可视化

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- Visualization of data is one of the most powerful and appealing techniques for data exploration
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

示例：Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure



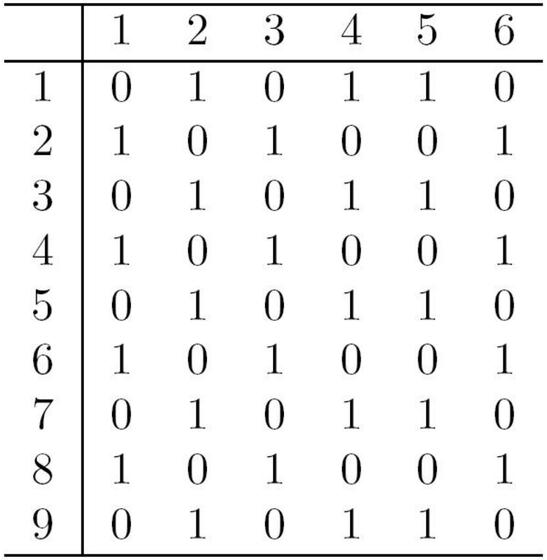
可视化元素

- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.

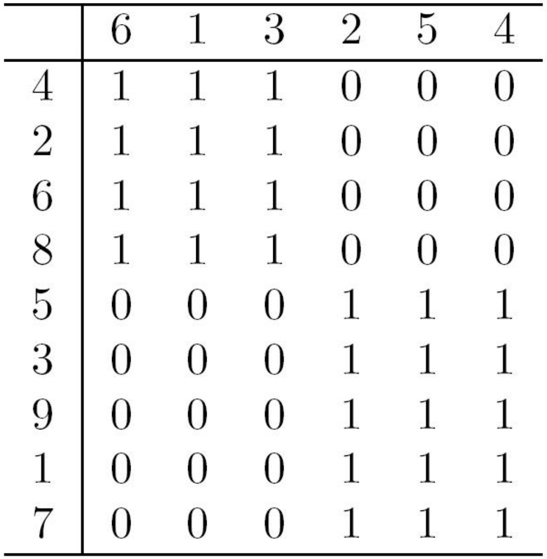
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

可视化：Arrangement

- Example:
- Arrangement is the placement of visual elements within a display Can make a large difference in how easy it is to understand the data



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0



	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

可视化：Selection

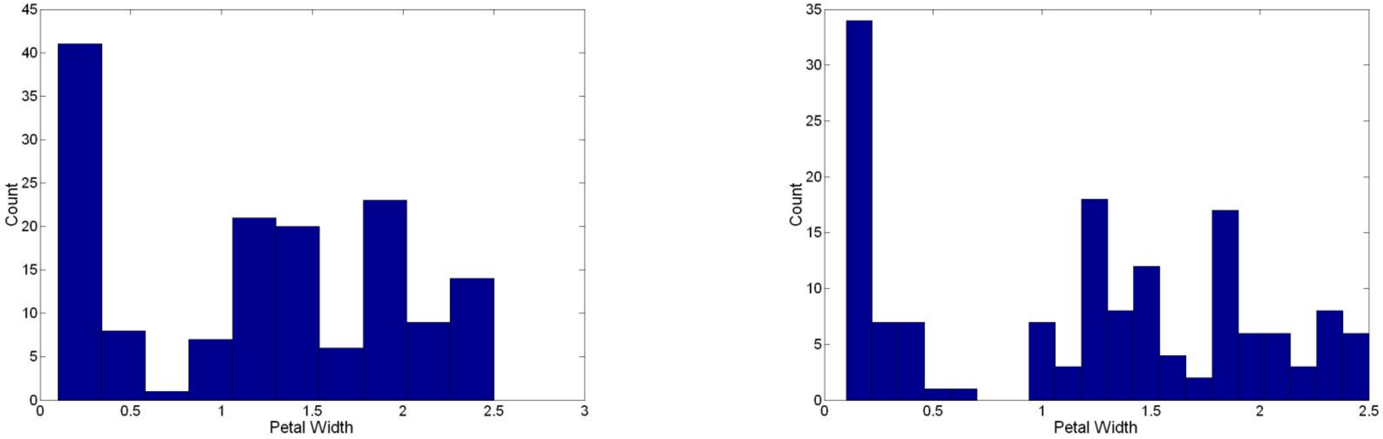
- Selection Is the elimination or the de-emphasis of certain objects and attributes

- Selection may involve the choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered

- Selection may also involve choosing a subset of objects
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas

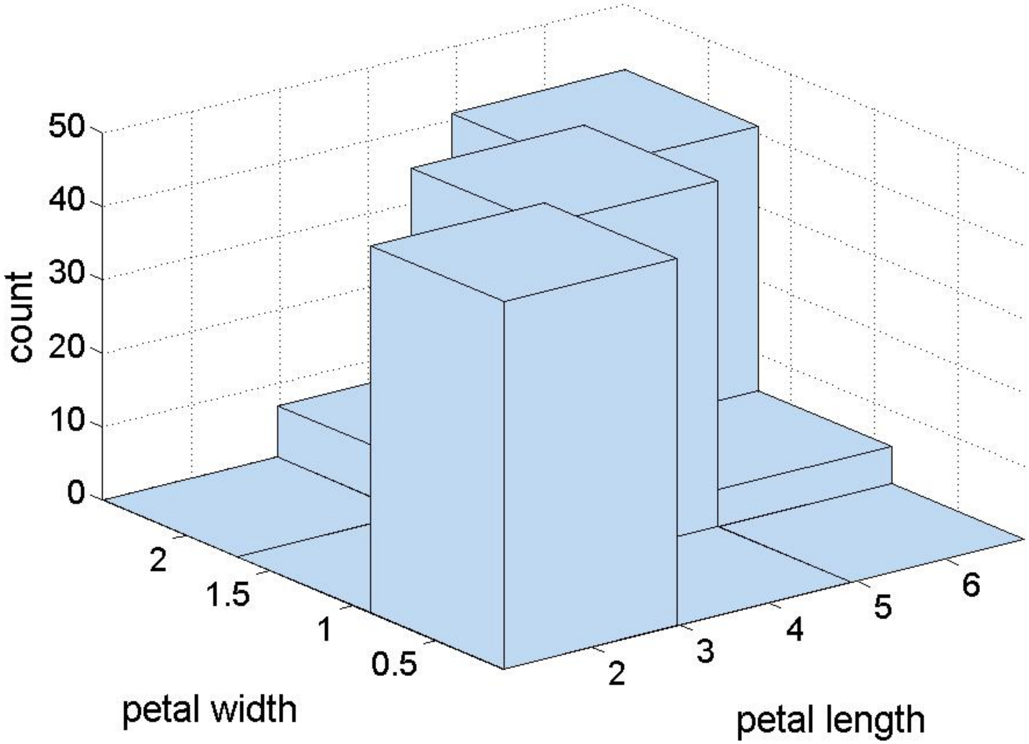
可视化技术: Histograms

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)



可视化技术: Two-Dimensional Histograms

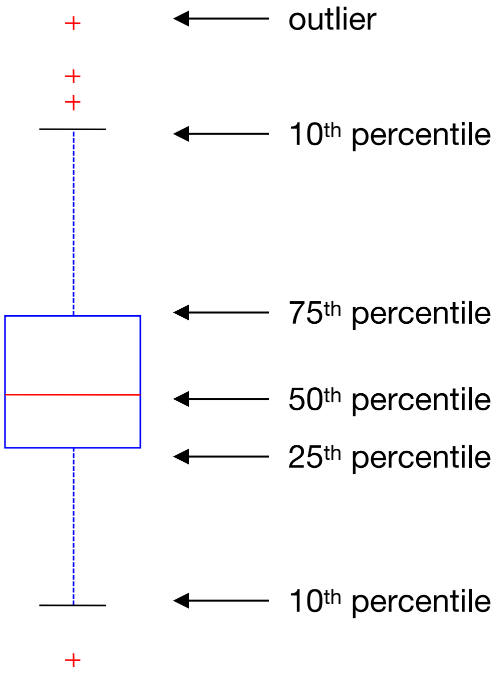
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



可视化技术: Box Plots

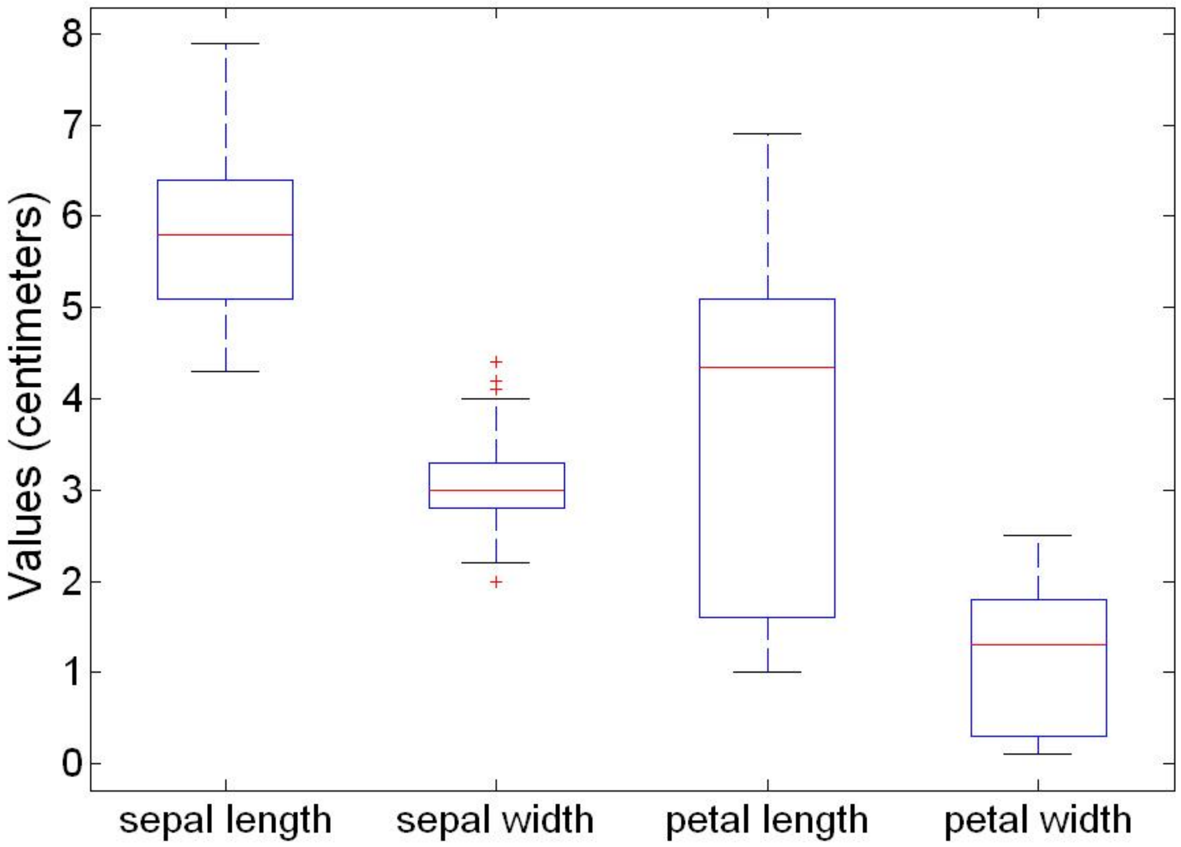
Box Plots

- Invented by J. Tukey
- Another way of displaying the distribution of data
- Following figure shows the basic part of a box plot

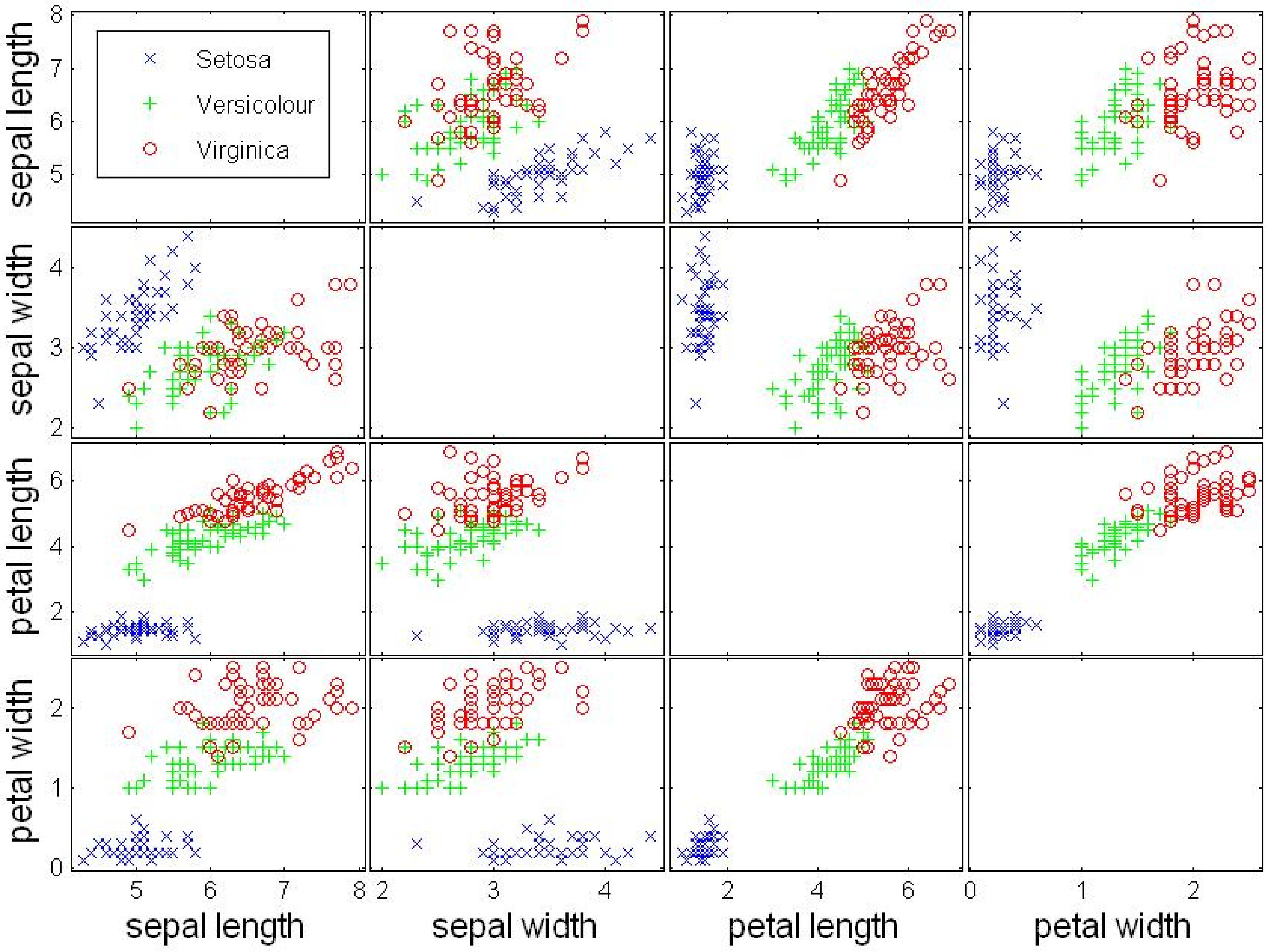


Box Plots 示例

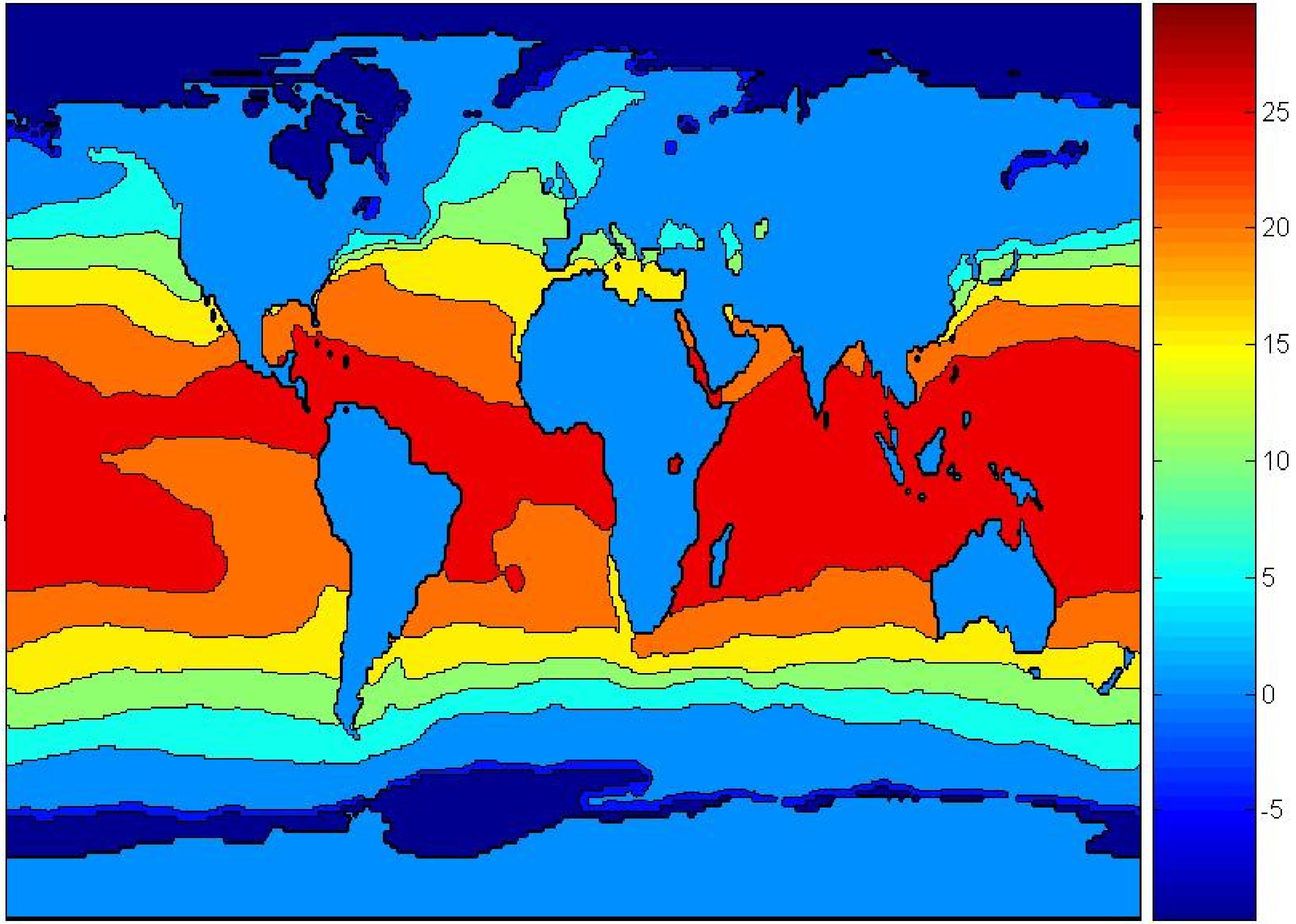
- Box plots can be used to compare attributes



可视化技术: Scatter Plot



可视化技术: Contour Plot



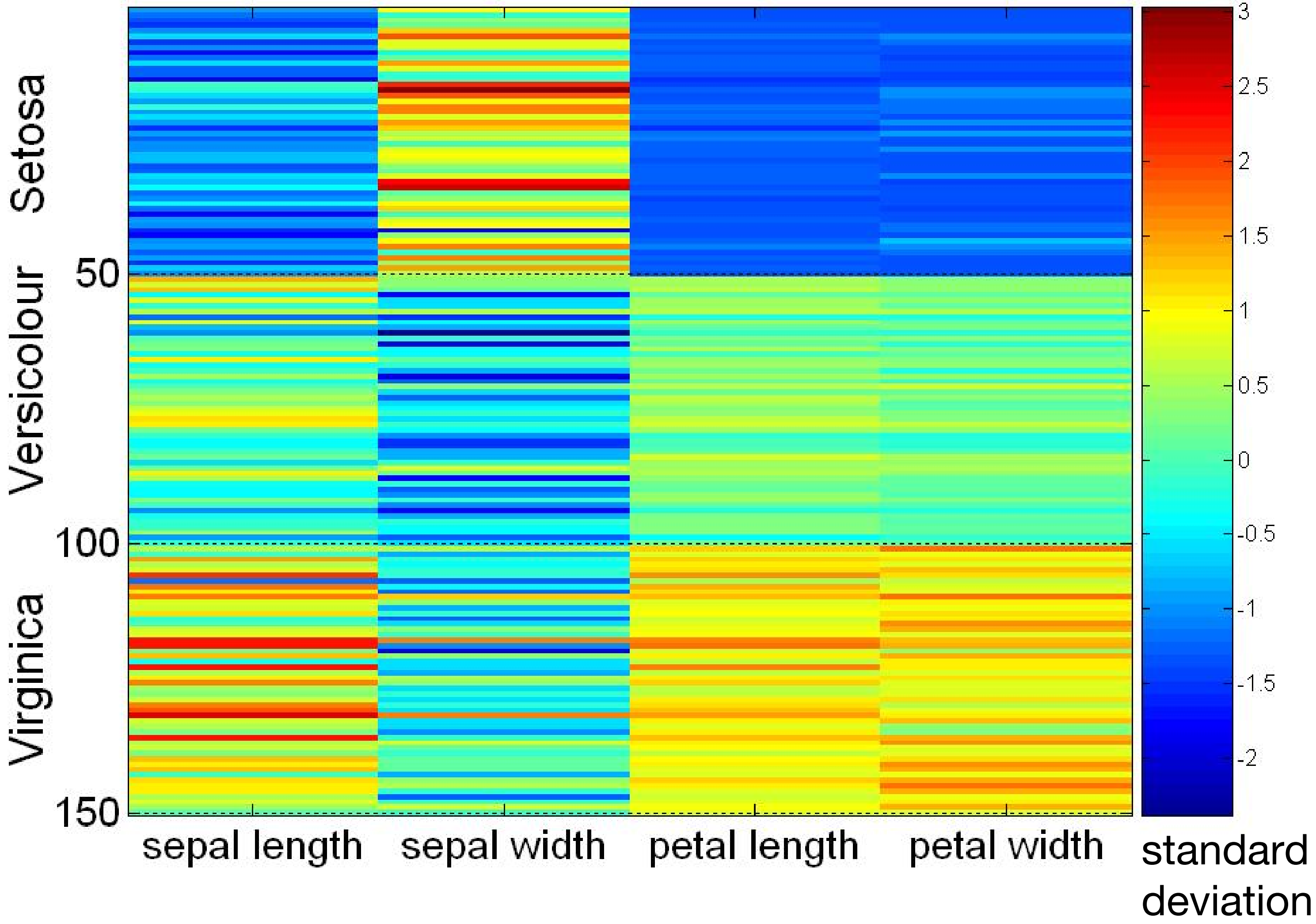
Example: SST Dec, 1998

Celsius

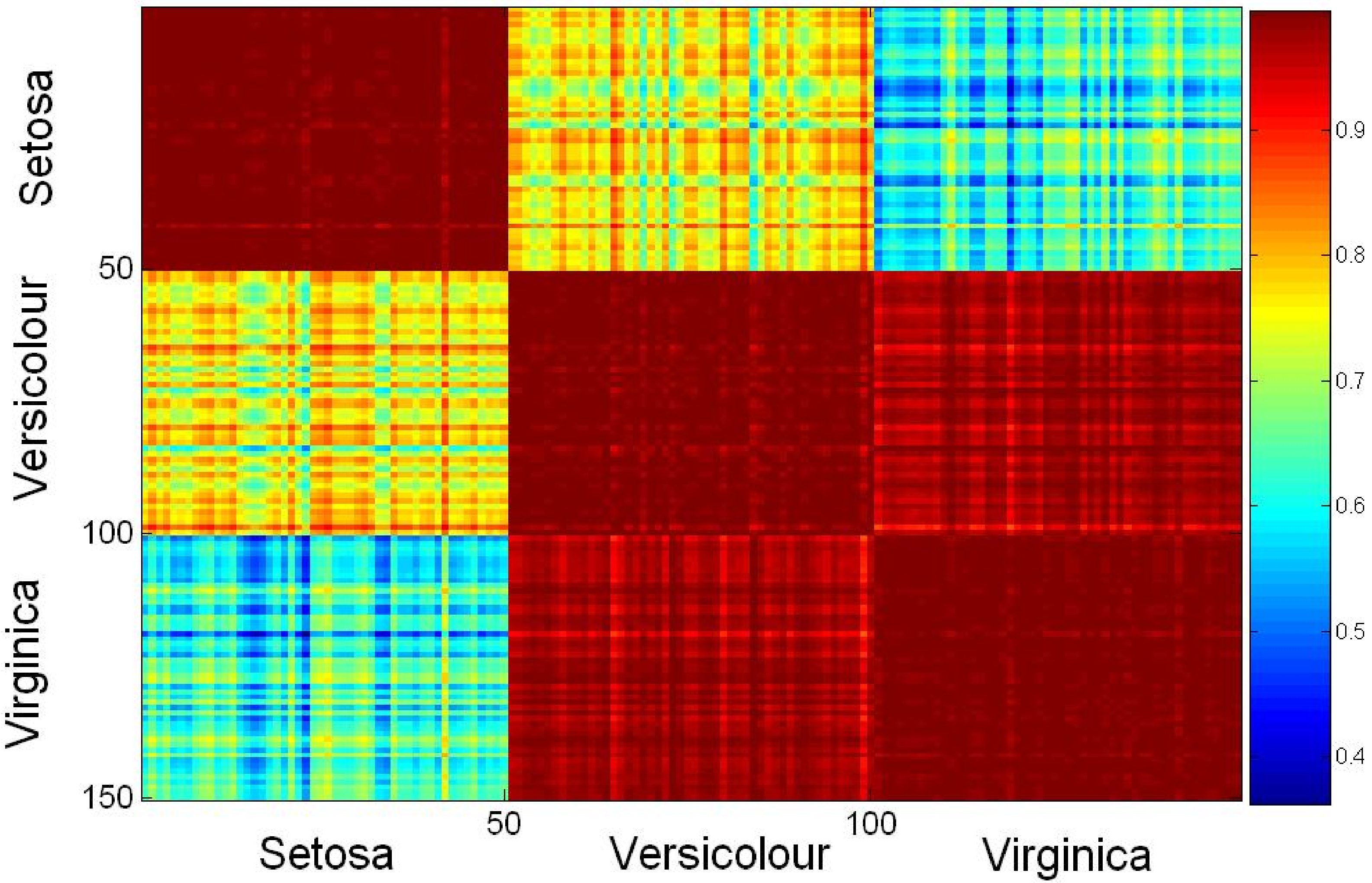
可视化技术: Matrix Plots

- Matrix plots
  - Can plot the data matrix
  - This can be useful when objects are sorted according to class
  - Typically, the attributes are normalized to prevent one attribute from dominating the plot
  - Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects
  - Examples of matrix plots are presented on the next two slides

Data Matrix 示例1



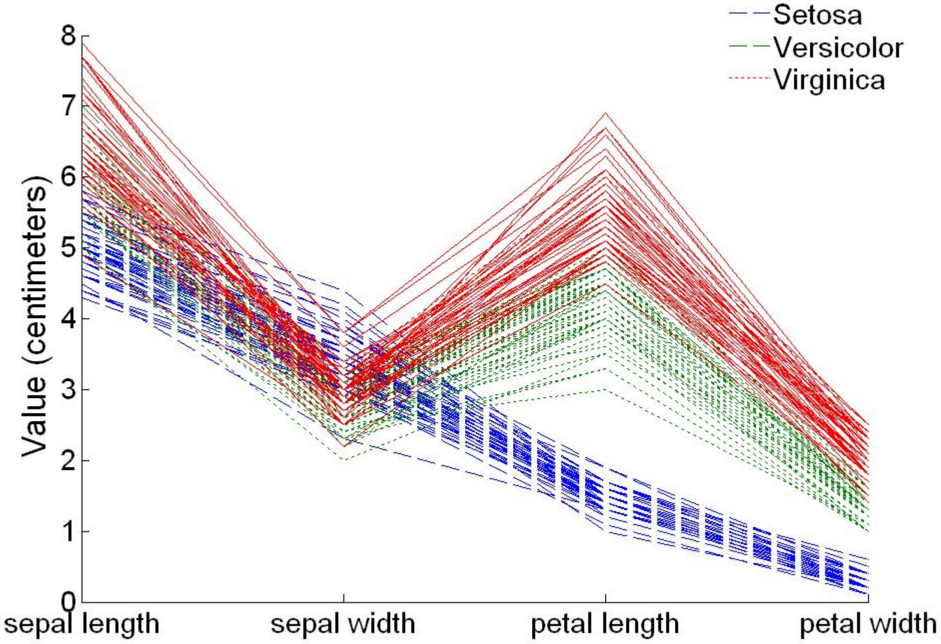
Data Matrix 示例2

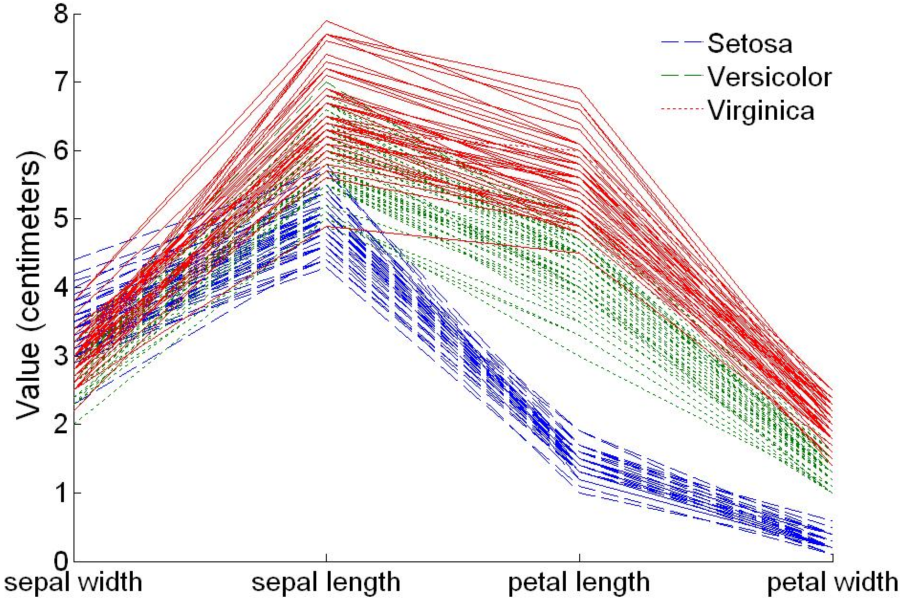


可视化技术: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots 示例





其他可视化技术

Star Plots

- Similar approach to parallel coordinates, but axes radiate from a central point
- The line connecting the values of an object is a polygon

Chernoff Faces

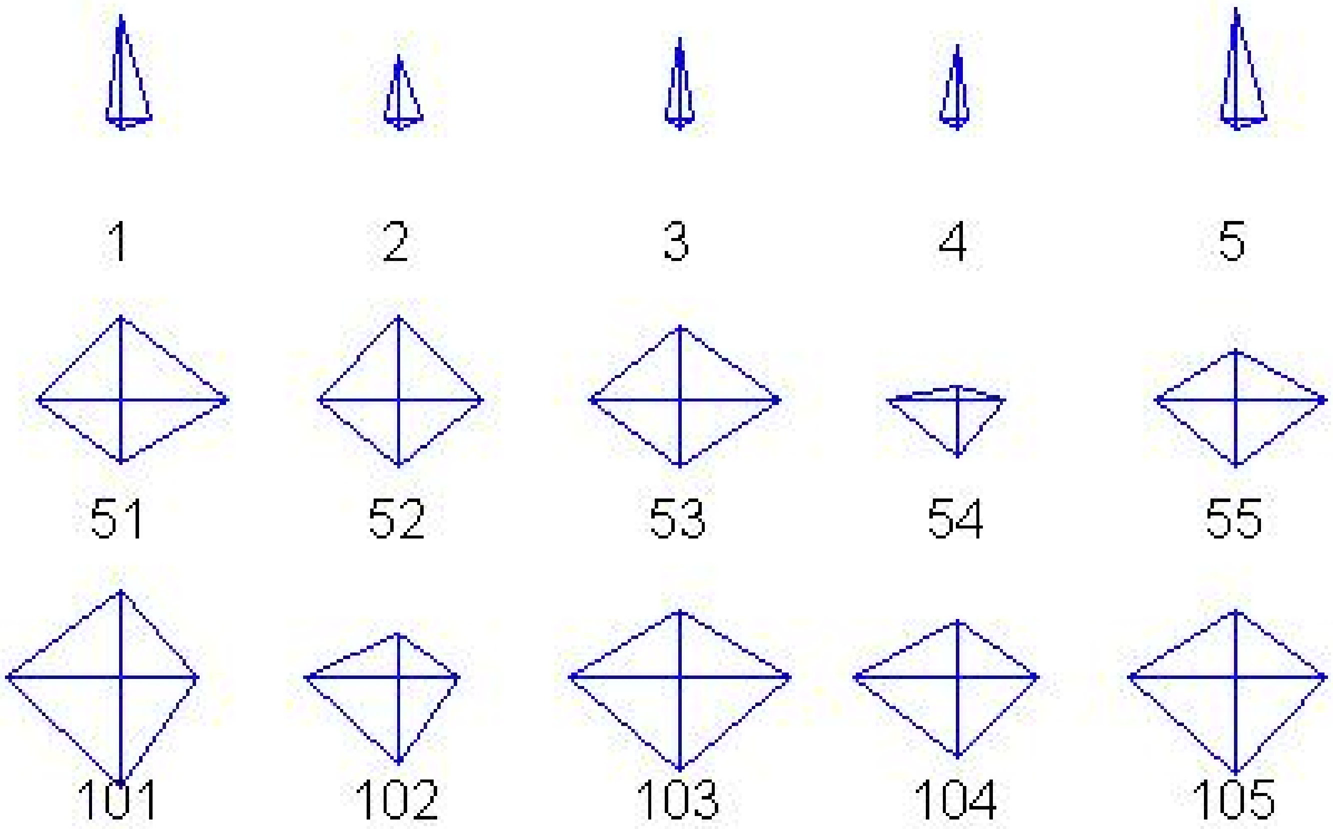
- Approach created by Herman Chernoff
- This approach associates each attribute with a characteristic of a face
- The values of each attribute determine the appearance of the corresponding facial characteristic
- Each object becomes a separate face
- Relies on human's ability to distinguish faces

示例: Star Plots

Setosa

Versicolour

Virginica

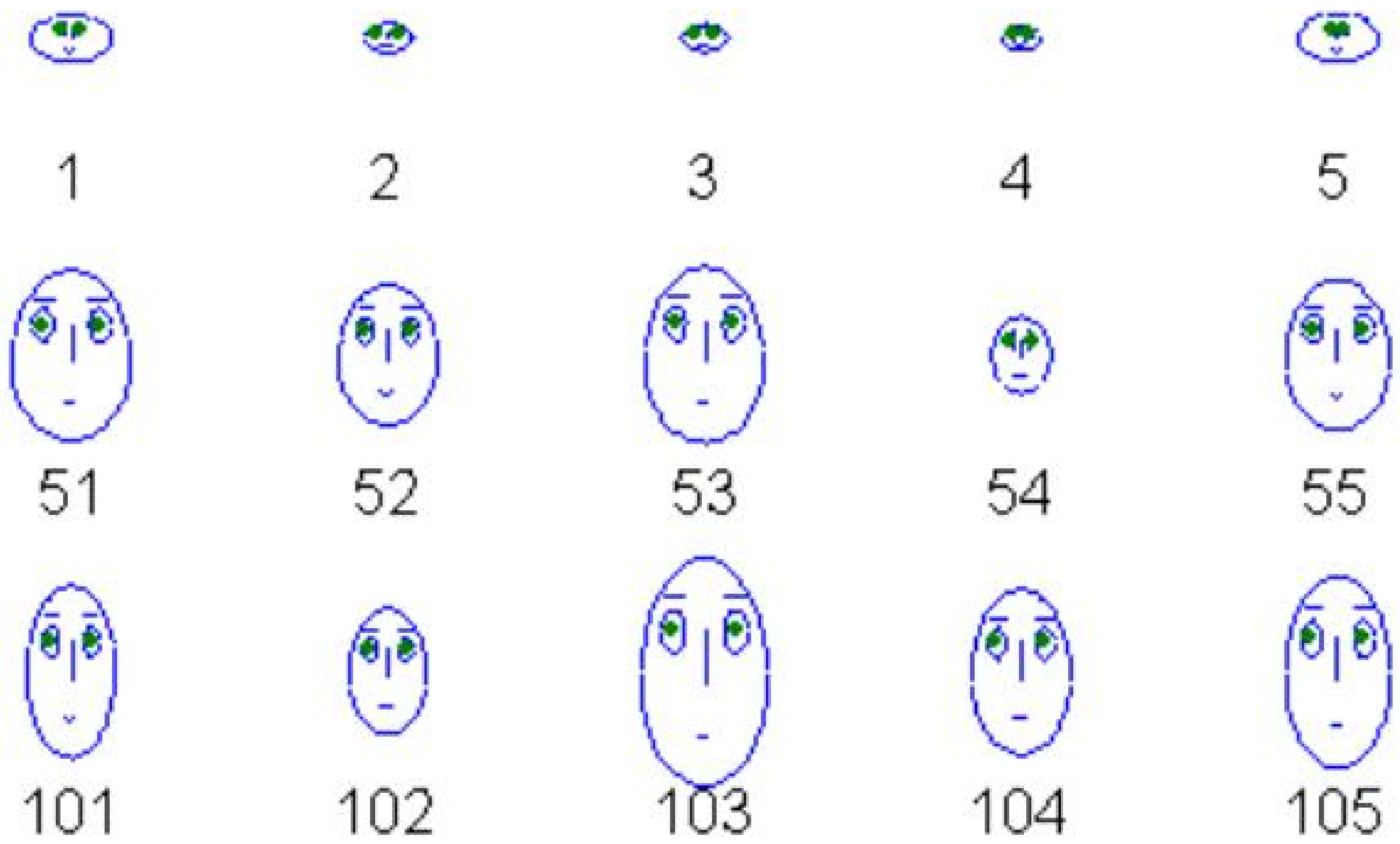


示例: Chernoff Faces

Setosa

Versicolour

Virginica



可视化原则

- ACCENT Rules

  - Apprehension
  - Clarity
  - Consistency
  - Efficiency
  - Necessity
  - Truthfulness

讲授提纲

01 数据的探索性分析

02 可视化

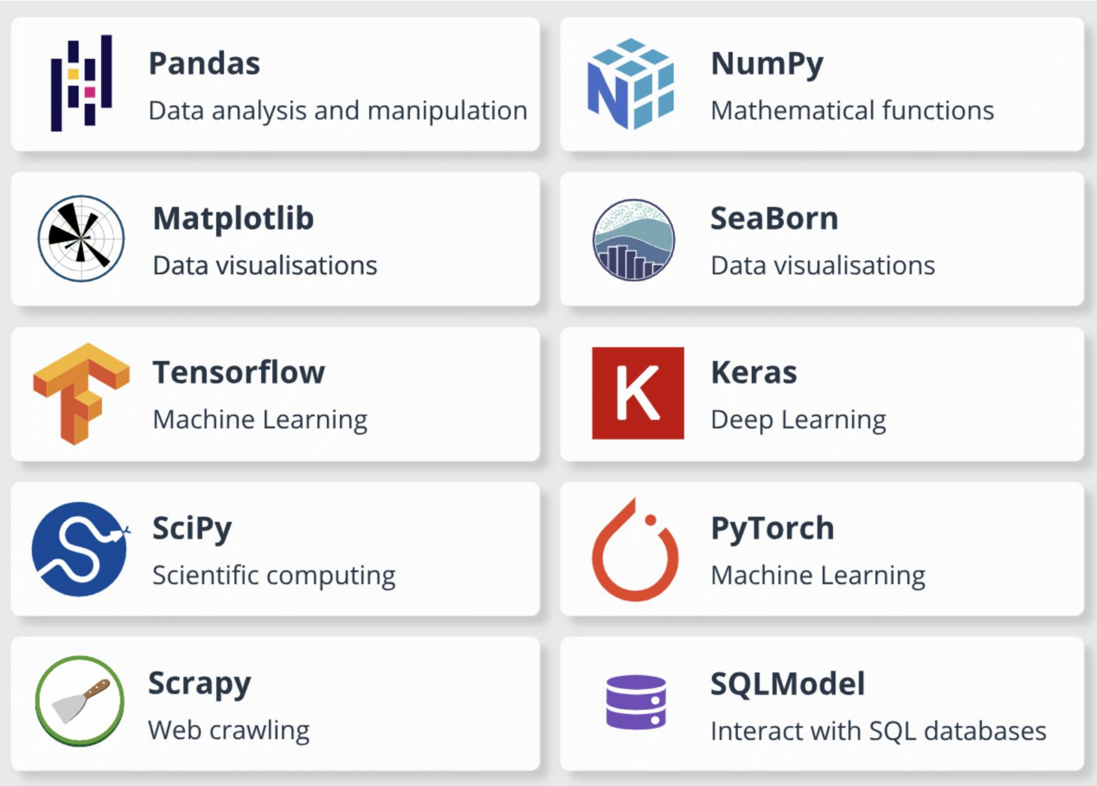
03 Streamlit 工具介绍

04 模型开发与部署

05 动手实践

Python 工具

- Python offers several libraries for analyzing, manipulating data, and developing interfaces to facilitate the creation of data analysis applications



Pandas
Data analysis and manipulation

NumPy
Mathematical functions

Matplotlib
Data visualisations

SeaBorn
Data visualisations

Tensorflow
Machine Learning

Keras
Deep Learning

SciPy
Scientific computing

PyTorch
Machine Learning

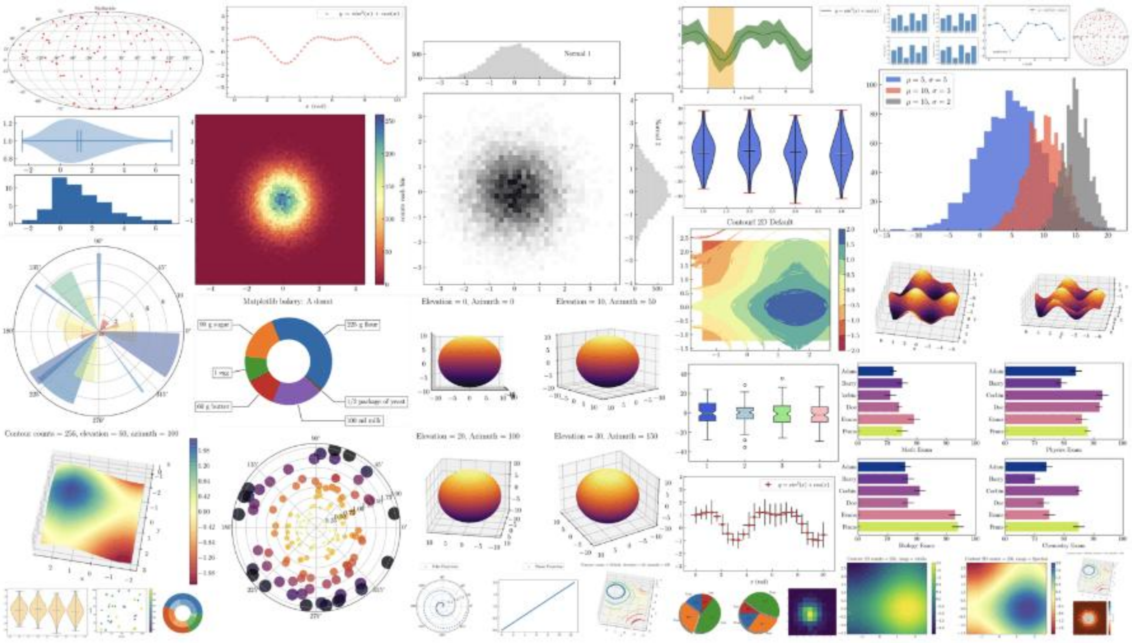
Scrapy
Web crawling

SQLModel
Interact with SQL databases

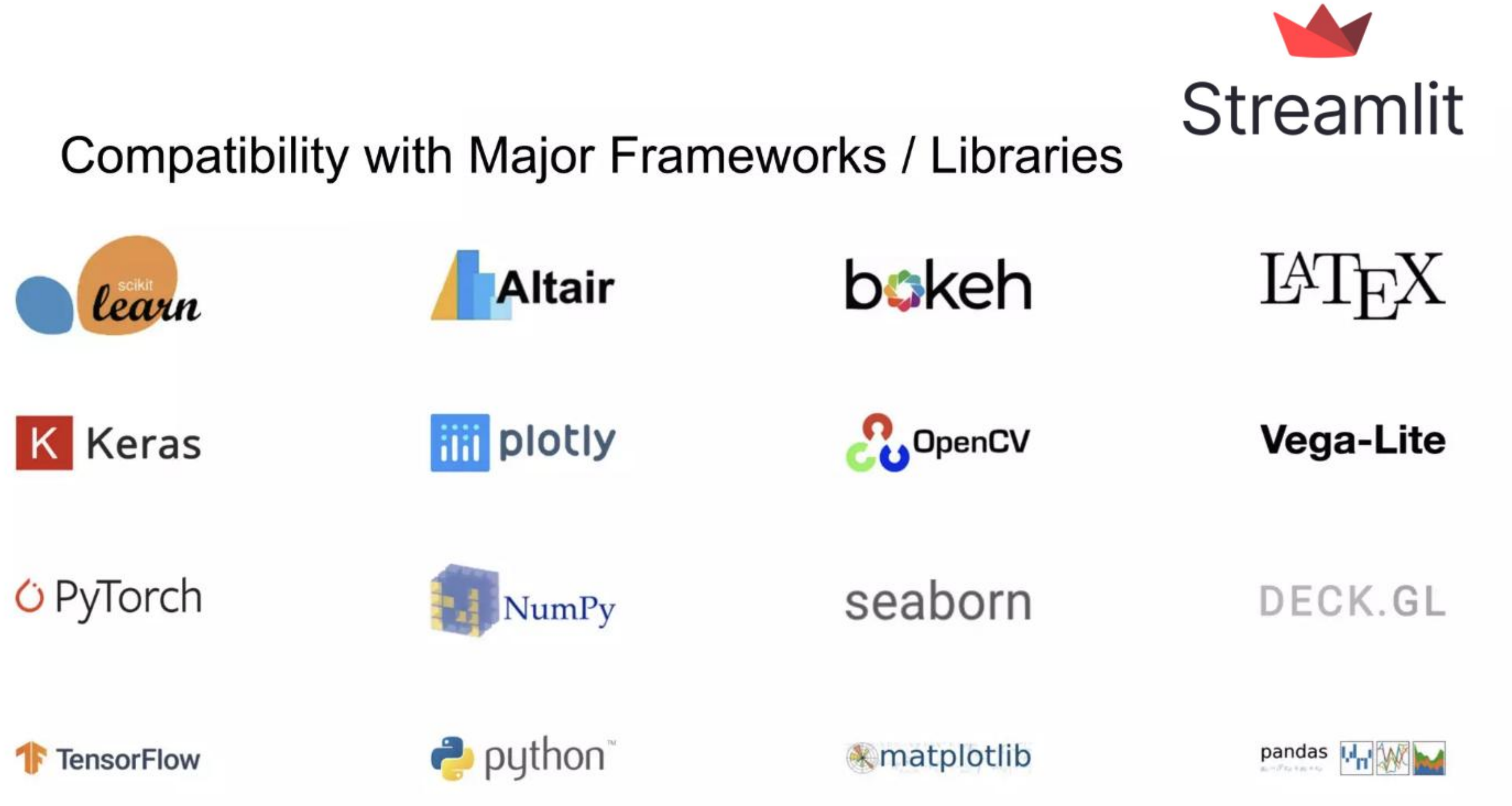
Python 工具: Matplotlib

- **Matplotlib is used for**

  - Creating static, animated, and interactive visualisations
  - Produces publication-quality figures
  - Has a wide variety of graphs and plots



为什么选择 Streamlit?



Streamlit

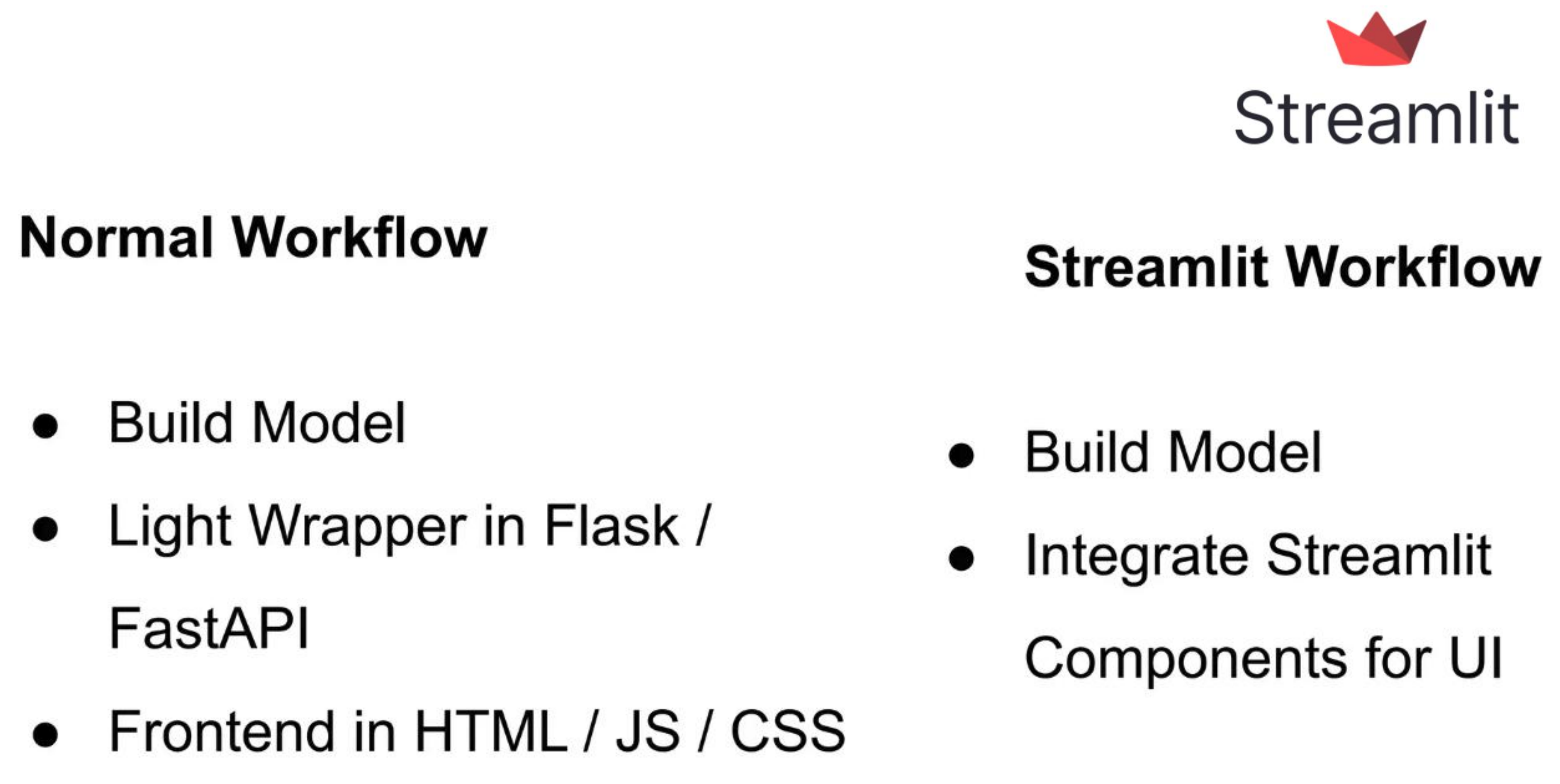
Compatibility with Major Frameworks / Libraries

如何使用？

Streamlit

- Creating an Interface for Machine Learning
- Visualizing Data

模型交互形式



Streamlit

Normal Workflow

- Build Model
- Light Wrapper in Flask / FastAPI
- Frontend in HTML / JS / CSS

Streamlit Workflow

- Build Model
- Integrate Streamlit Components for UI

模型交互形式：示例

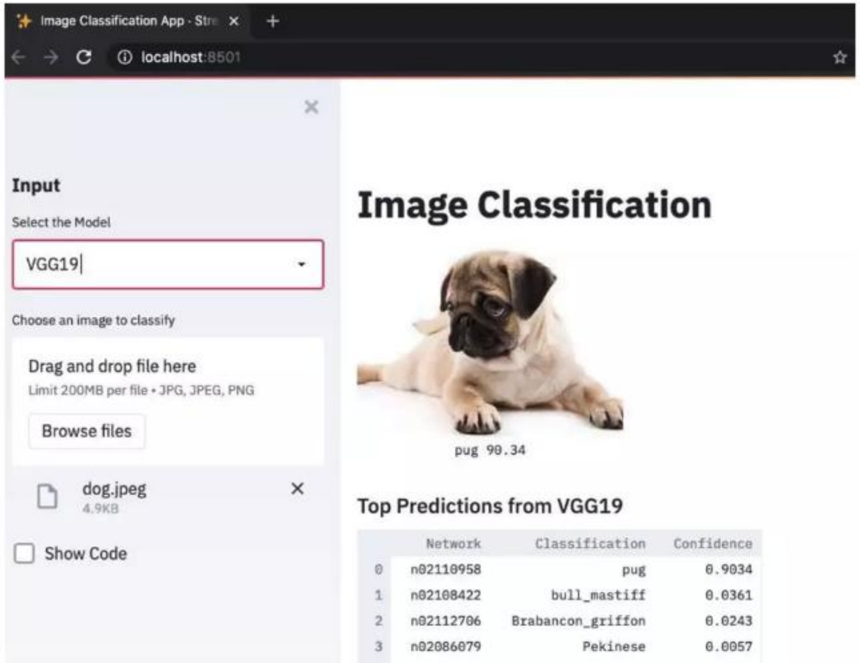
Streamlit

- Interface for Image Net

  Image Classification

- 3 Lines of Streamlit Magic ✨

- Similar to CLI Parser



Input

Select the Model

VGG19

Choose an image to classify

Drag and drop file here

Limit 200MB per file • JPG, JPEG, PNG

Browse files

dog.jpeg
4.9KB

□ Show Code

Image Classification

pug 90.34

Top Predictions from VGG19

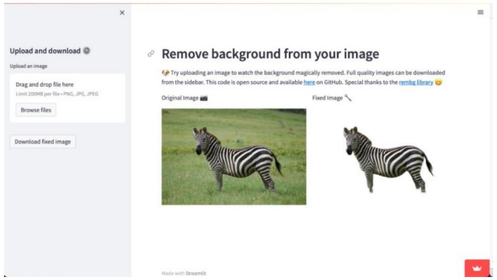
	Network	Classification	Confidence
0	n02110958	pug	0.9034
1	n02108422	bull_mastiff	0.0361
2	n02112706	Brabancon_griffon	0.0243
3	n02086079	Pekinese	0.0057

数据可视化

- Notebooks
- Presentations
- Scripts
- Code
- Reports in Tableau
- Web Apps?

Examples Gallery

- **There are several templates and applications created by the community**
  - https://streamlit.io/gallery



Upload and download

Upload an image

Drag and drop file here

Limit 200MB per file • PNG, JPG, JPEG

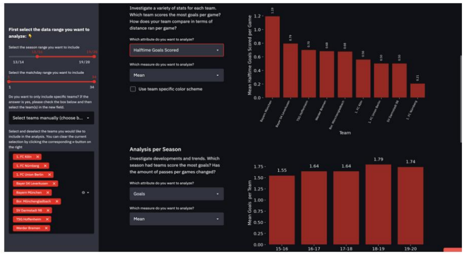
Browse files

Download fixed image

Remove background from your image

Try uploading an image to watch the background magically removed. Full quality images can be downloaded from the sidebar. This code is open source and available here on GitHub. Special thanks to the rembg library

Background Removal



First select the data range you want to analyze:

Select the season range you want to include: 15/16 – 19/20 (13/14 … 19/20)

Select the matchday range you want to include: 1 – 34 (1 … 34)

Do you want to only include specific teams? If the answer is yes, please check the box below and then select the team(s) in the new field.

Select teams manually (choose b...

Select and deselect the teams you would like to include in the analysis. You can clear the current selection by clicking the corresponding x-button on the right

1. FC Köln, 1. FC Nürnberg, 1. FC Union Berlin, Bayer 04 Leverkusen, Bayern München, Bor. Mönchengladbach, SV Darmstadt 98, TSG Hoffenheim, Werder Bremen

Investigate a variety of stats for each team. Which team scores the most goals per game? How does your team compare in terms of distance ran per game?

Which attribute do you want to analyze? Halftime Goals Scored

Which measure do you want to analyze? Mean

☐ Use team specific color scheme

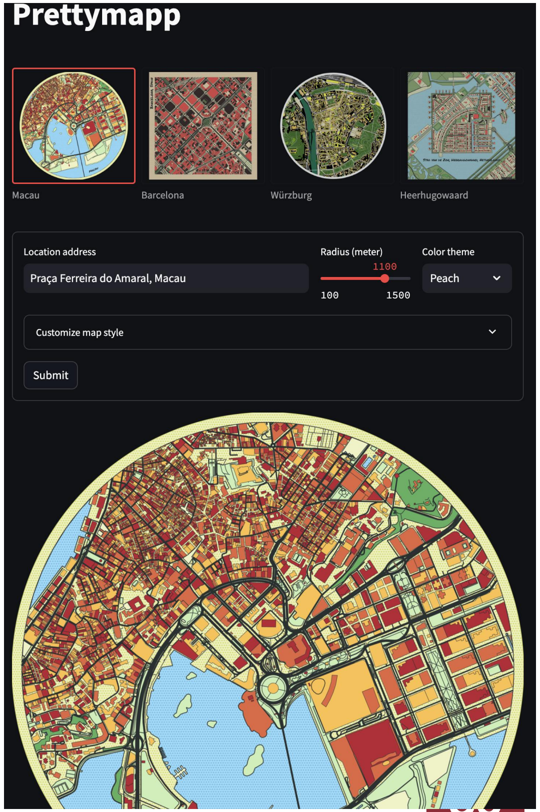
Analysis per Season

Investigate developments and trends. Which season had teams score the most goals? Has the amount of passes per games changed?

Which attribute do you want to analyze? Goals

Which measure do you want to analyze? Mean

Bundesliga analyzer



Prettymapp

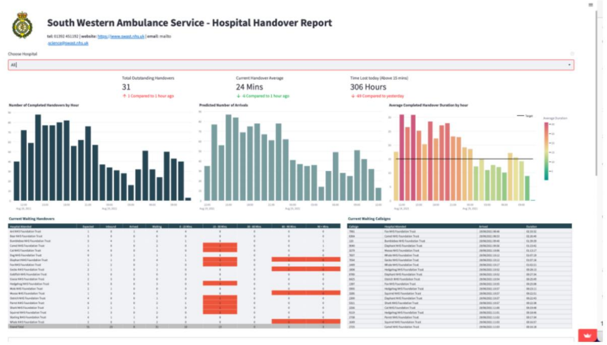
Location address: Praça Ferreira do Amaral, Macau

Radius (meter): 1100 (100 … 1500)

Color theme: Peach

Customize map style

Submit



South Western Ambulance Service - Hospital Handover Report

tel: 01392 451192 | website: https://www.swast.nhs.uk | email: mailto science@swast.nhs.uk

Choose Hospital: All

Total Outstanding Handovers: 31 (↑ 1 Compared to 1 hour ago)

Current Handover Average: 24 Mins (↓ -6 Compared to 1 hour ago)

Time Lost today (Above 15 mins): 306 Hours (↓ -69 Compared to yesterday)

Number of Completed Handovers by Hour
Predicted Number of Arrivals
Average Completed Handover Duration by hour
Target
Average Duration
Current Waiting Handovers
Current Waiting Callsigns

SWAST - Hospital Handover Report

讲授提纲

01 数据的探索性分析

02 可视化

03 Streamlit 工具介绍

04 模型开发与部署

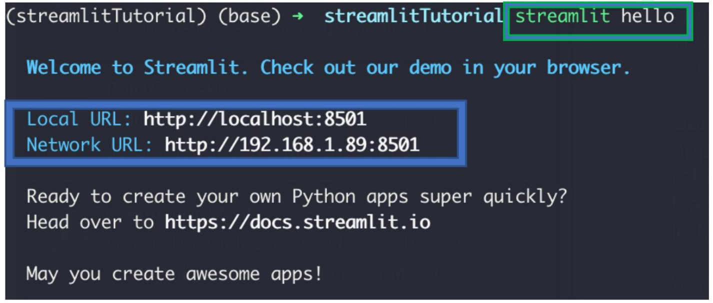
05 动手实践

Streamlit 安装

- Python 3.7 – Python 3.11
- Using a virtual environment is recommended
- Install Streamlit
  - pip install streamlit
- Test the installation
  - streamlit hello
- Launch your own application
  - streamlit run your_script.py [-- script args]
  - python -m streamlit run your_script.py

https://docs.streamlit.io/library/get-started/installation

使用启动

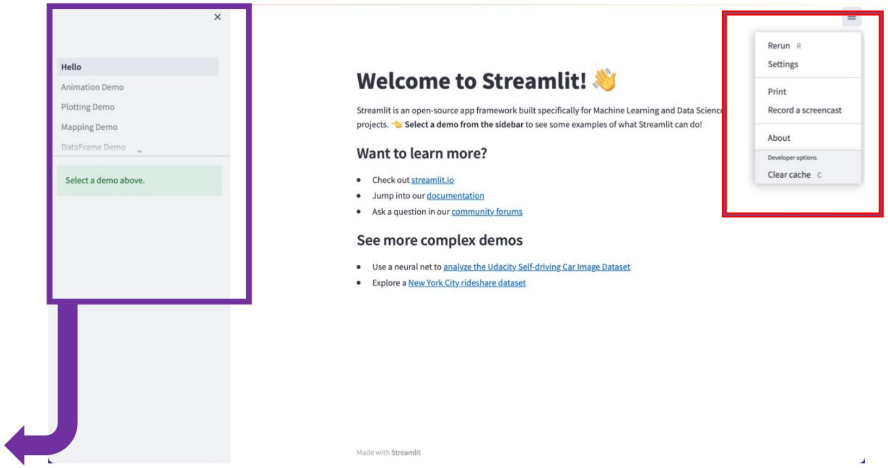


```
(streamlitTutorial) (base) ➜  streamlitTutorial streamlit hello

  Welcome to Streamlit. Check out our demo in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.89:8501

  Ready to create your own Python apps super quickly?
  Head over to https://docs.streamlit.io

  May you create awesome apps!
```

Command to start Streamlit

URL to reach the web server at port 8501

Hamburger menu



Welcome to Streamlit! 👋

Streamlit is an open-source app framework built specifically for Machine Learning and Data Scienc projects. 👈 **Select a demo from the sidebar** to see some examples of what Streamlit can do!

Want to learn more?

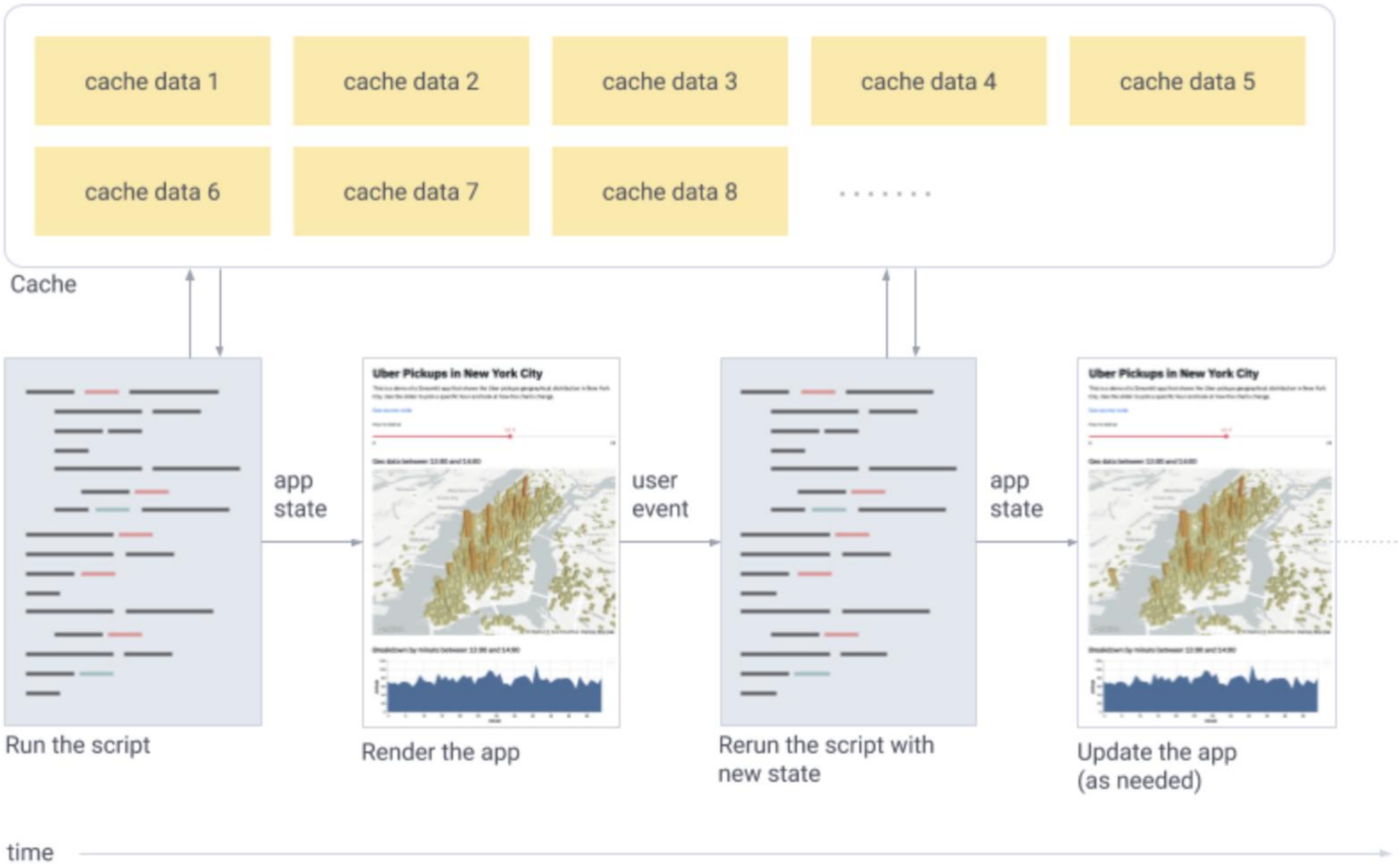
- Check out streamlit.io
- Jump into our documentation
- Ask a question in our community forums

See more complex demos

- Use a neural net to analyze the Udacity Self-driving Car Image Dataset
- Explore a New York City rideshare dataset

Sidebar with access to sample demos

Streamlit 运行机制



cache data 1

cache data 2

cache data 3

cache data 4

cache data 5

cache data 6

cache data 7

cache data 8

· · · · · · · ·

Cache

app state

user event

app state

Run the script

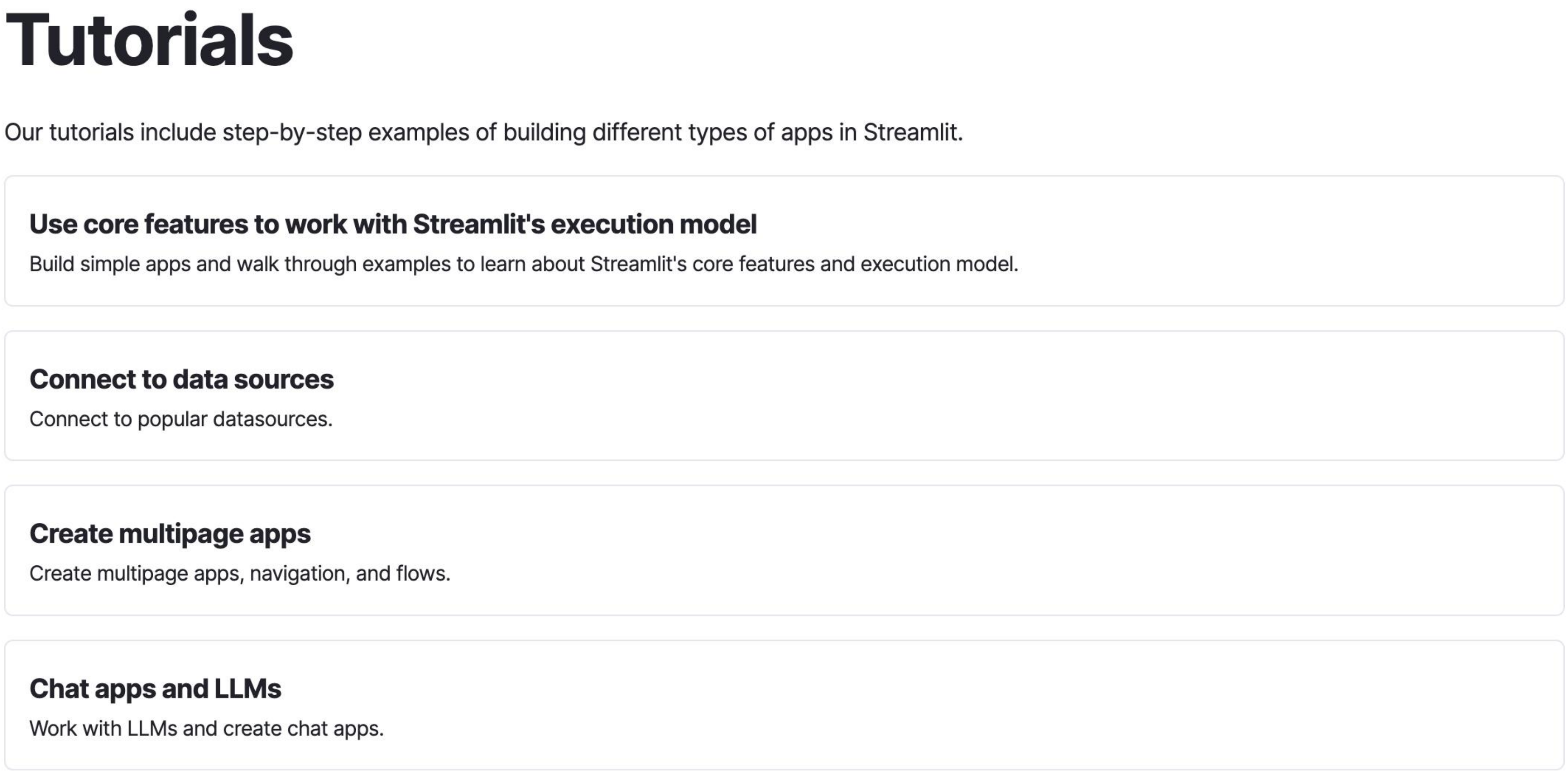
Render the app

Rerun the script with new state

Update the app (as needed)

time

基于Streamlit 的开发



Tutorials

Our tutorials include step-by-step examples of building different types of apps in Streamlit.

Use core features to work with Streamlit's execution model

Build simple apps and walk through examples to learn about Streamlit's core features and execution model.

Connect to data sources

Connect to popular datasources.

Create multipage apps

Create multipage apps, navigation, and flows.

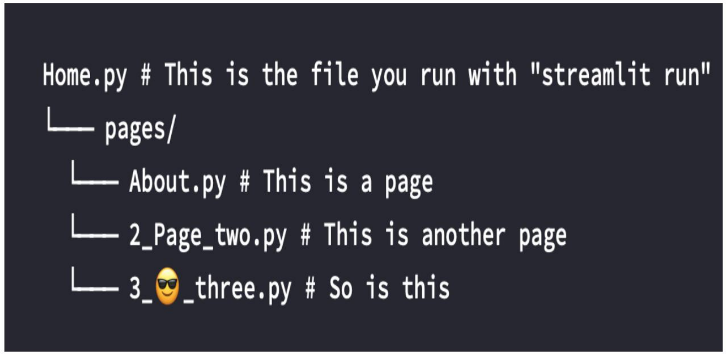
Chat apps and LLMs

Work with LLMs and create chat apps.

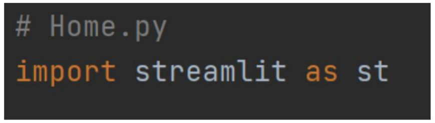
https://docs.streamlit.io/develop

基于Streamlit 的开发

- Before you develop your app, it's important to define the project *directory* structure
- You need to define an ***entrypoint file*** that represents the main page to show to the user
- Other additional pages should be placed in a sub-folder ***pages***
- Pages globally share the same Python modules



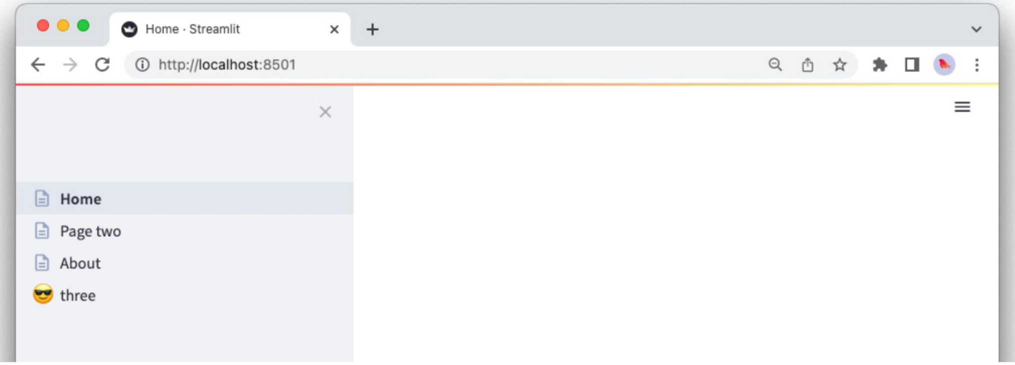
```
Home.py # This is the file you run with "streamlit run"
└── pages/
  └── About.py # This is a page
  └── 2_Page_two.py # This is another page
  └── 3_😎_three.py # So is this
```



```
# Home.py
import streamlit as st
```

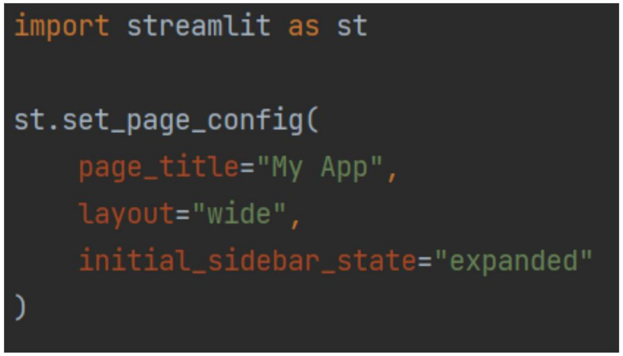
基于Streamlit 的开发 : Pages

- Pages are defined by files *.py* within the *"pages/"* folder
- File names are transformed into page names
- The order is given by the number preceding the title and/or by the alphabetical order of the title itself.
- The number used as a prefix in the file name is not interpreted as part of the title



Page 配置

- Set the default page configuration
  - ➢ st.set_page_config(page_title=None, page_icon=None, layout="centered", initial_sidebar_state="auto", menu_items=None)



```
import streamlit as st

st.set_page_config(
    page_title="My App",
    layout="wide",
    initial_sidebar_state="expanded"
)
```

⚠ *It must be the first Streamlit command and set only once!*

基于Streamlit 的开发 : Elements

- Widgets and elements specific to different types of activities and inputs
  - quickly integrate different features into your application
  - available through official documentation: https://docs.streamlit.io/library/api-reference
- Most significant categories:
  - Text elements
  - Input widgets
  - Layout
  - Visualization of data and graphs
  - Additional elements

Element Arguments

- The various elements can be integrated without special configurations
  - Personalization via certain arguments
- Some arguments are common to all (or most) of the elements:
  - ***label***: describes to the user the functionality of the element (e.g. the name of a clickable button)
  - ***label_visibility*:** determine label visibility (i.e., ”visible”, ”hidden”, ”collapsed”); the label should always be defined
  - ***disabled***: boolean flag to disable an element. Useful for making a widget available only if a certain condition occurs
  - ***use_container_width***: boolean flag to fit the size of the widget to that of the container it is part of
  - ***key:*** string or number to uniquely identify the widget. If omitted, it is generated based on content

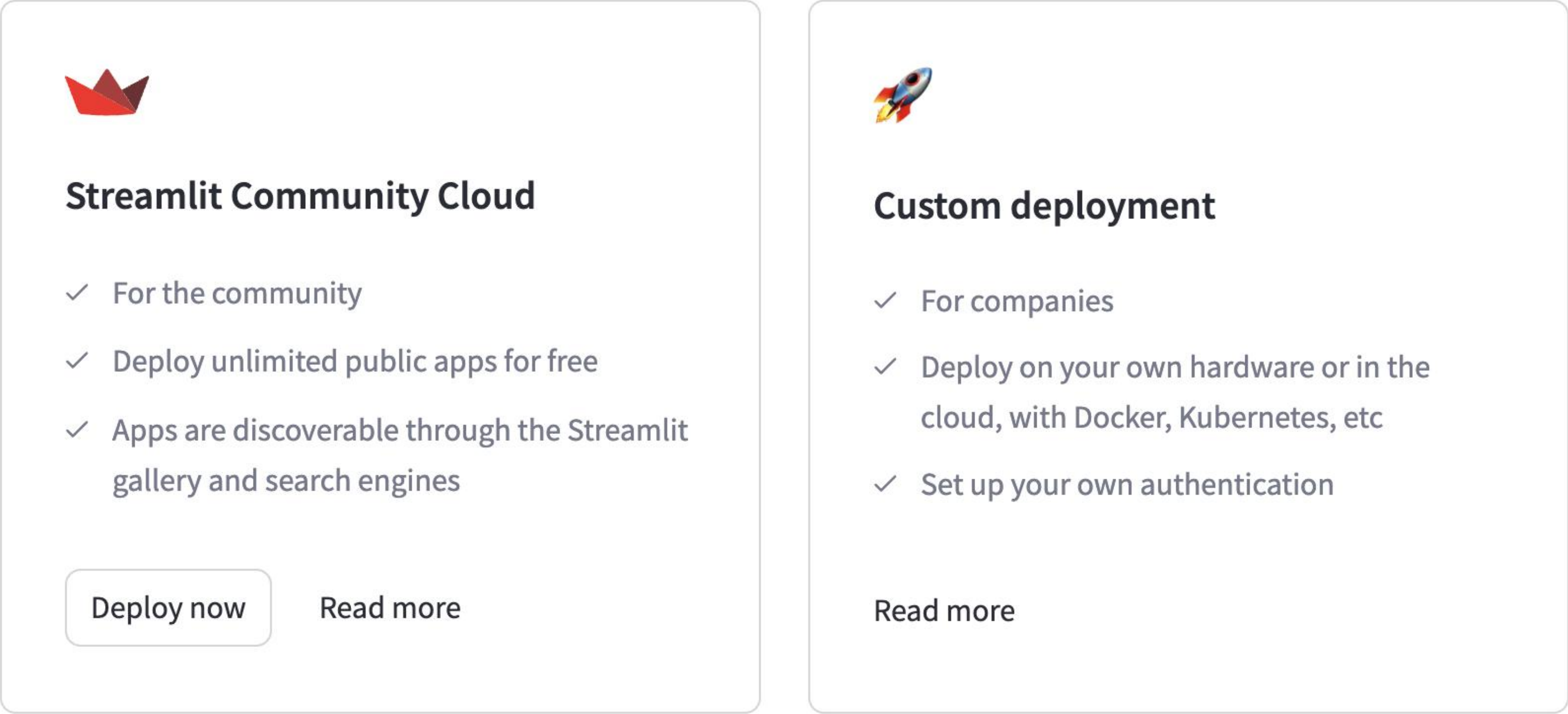
Different items cannot have the same key!

基于Streamlit 的部署

- Deploy your app and share it with your users
- There are three main processes:
  - Install Python, Streamlit, and other dependencies in your deployment environment
  - Securely handle your secrets and private information
  - Remote start your app (streamlit run)

https://docs.streamlit.io/deploy

部署选择



Streamlit Community Cloud

✓ For the community

✓ Deploy unlimited public apps for free

✓ Apps are discoverable through the Streamlit gallery and search engines

Deploy now   Read more

Custom deployment

✓ For companies

✓ Deploy on your own hardware or in the cloud, with Docker, Kubernetes, etc

✓ Set up your own authentication

Read more

snowflake

Docker

Kubernetes

讲授提纲

01 数据的探索性分析

02 可视化

03 Streamlit 工具介绍

04 模型开发与部署

05 动手实践

动手实践

- Streamlit Installation
- Streamlit: Elements
- Streamlit: Development
- Streamlit: Deployment

总结

01 数据的探索性分析

02 可视化

03 Streamlit 工具介绍

04 模型开发与部署

05 动手实践