



TAIJI LABORATORY  
FOR GRAVITATIONAL WAVE UNIVERSE



ICTP-AP  
International Centre  
for Theoretical Physics Asia-Pacific  
国际理论物理中心-亚太地区



中国科学院大学  
University of Chinese Academy of Sciences

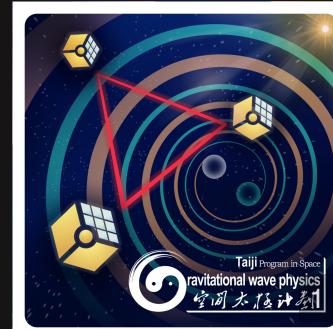
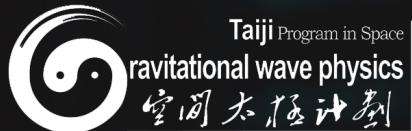
# 引力波数据探索：编程与分析实战训练营

第 2 部分 基于 Python 的数据分析基础  
数据分析可视化之 Matplotlib / Seaborn

主讲老师：王赫

ICTP-AP, UCAS

2023/12/07 & 2023/12/10





# 数据分析可视化理论基础

- 数据化思维与数据可视化
- 数据可视化：目的与原则
- 数据可视化：模式与对象
- 数据可视化：常用工具
- 数据可视化：基本处理流程



# 什么是数据可视化？

- 往年的双十一展示用大屏的数据可视化系统实时监控，通过这些数据对网站进行优化、对用户群体进行基本的分析等。



图片来源于网络

- 数据可视化**，是关于数据视觉表现形式的科学技术研究。其中，这种数据的视觉表现形式被定义为，一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。
- 它是一个处于不断演变之中的概念，其边界在不断地扩大。主要指的是技术上较为高级的技术方法，而这些技术方法允许利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。与立体建模之类的特殊技术方法相比，数据可视化所涵盖的技术方法要广泛得多。

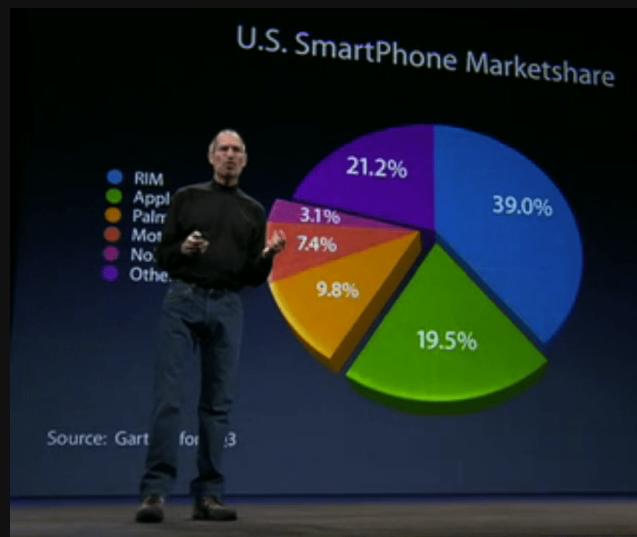




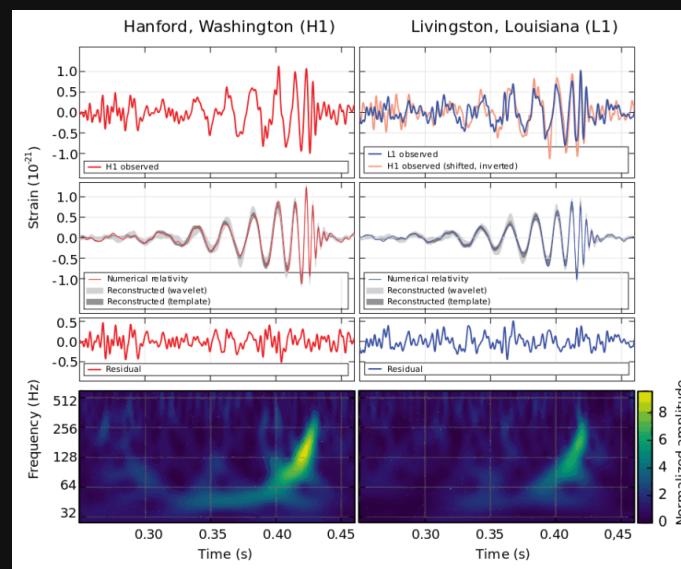
# 什么是数据可视化?

- **数据可视化**的巨大需求，带动了诸多领域应用
- 几乎所有涉及 Data 这个行业的岗位招聘，都会或多或少的涉及到 Data Visualization 能力的需求，甚至有不少是招专门的数据可视化分析师 (Data Visualization Analyst)
- **工业界**: 以高效性、建设性方式的讨论结果，理解运营管理与业务性能之间的关系，提供直接有效的决策依据。
- **学术界**: 在庞大繁杂的资源环境中，有目的的挖掘数据背后的特征信息，揭示不同思考角度下的潜在客观规律。

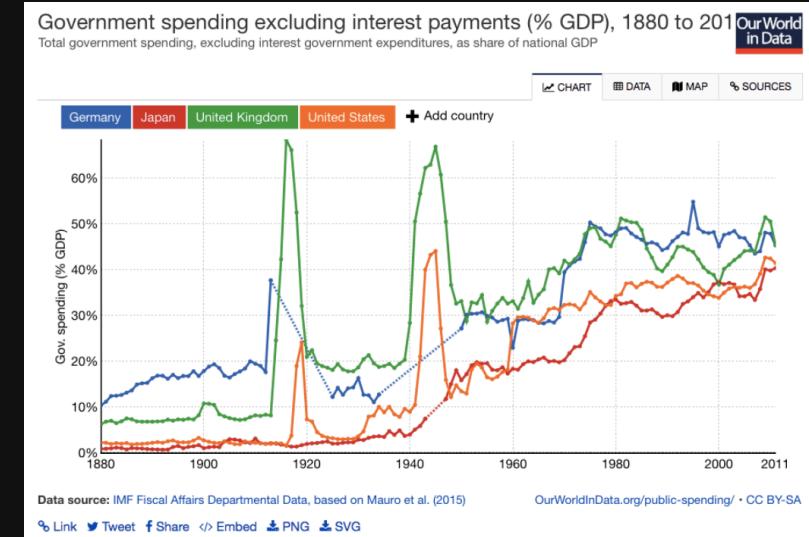
工业界



学术界



工商管理





# 数据化思维的重要性

- **数据化思维**是指根据数据来思考事物的一种思维模式，是一种量化的、重视事实、追求真理的思维模式。

	2	3	4	5	6	7
Batch 1	10	40	50	20	10	50
Batch 2	30	60	70	50	40	30

数据分析

数据



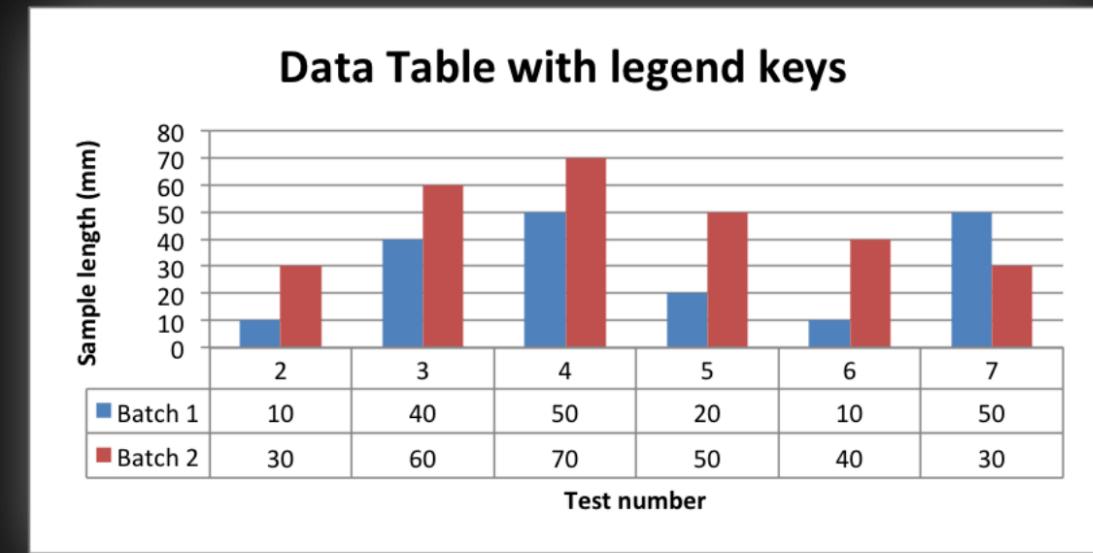
结论



# 数据化思维的重要性

- **数据化思维**是指根据数据来思考事物的一种思维模式，是一种量化的、重视事实、追求真理的思维模式。

	2	3	4	5	6	7
Batch 1	10	40	50	20	10	50
Batch 2	30	60	70	50	40	30



数据分析

数据



结论

数据分析

数据



结论

数据可视化

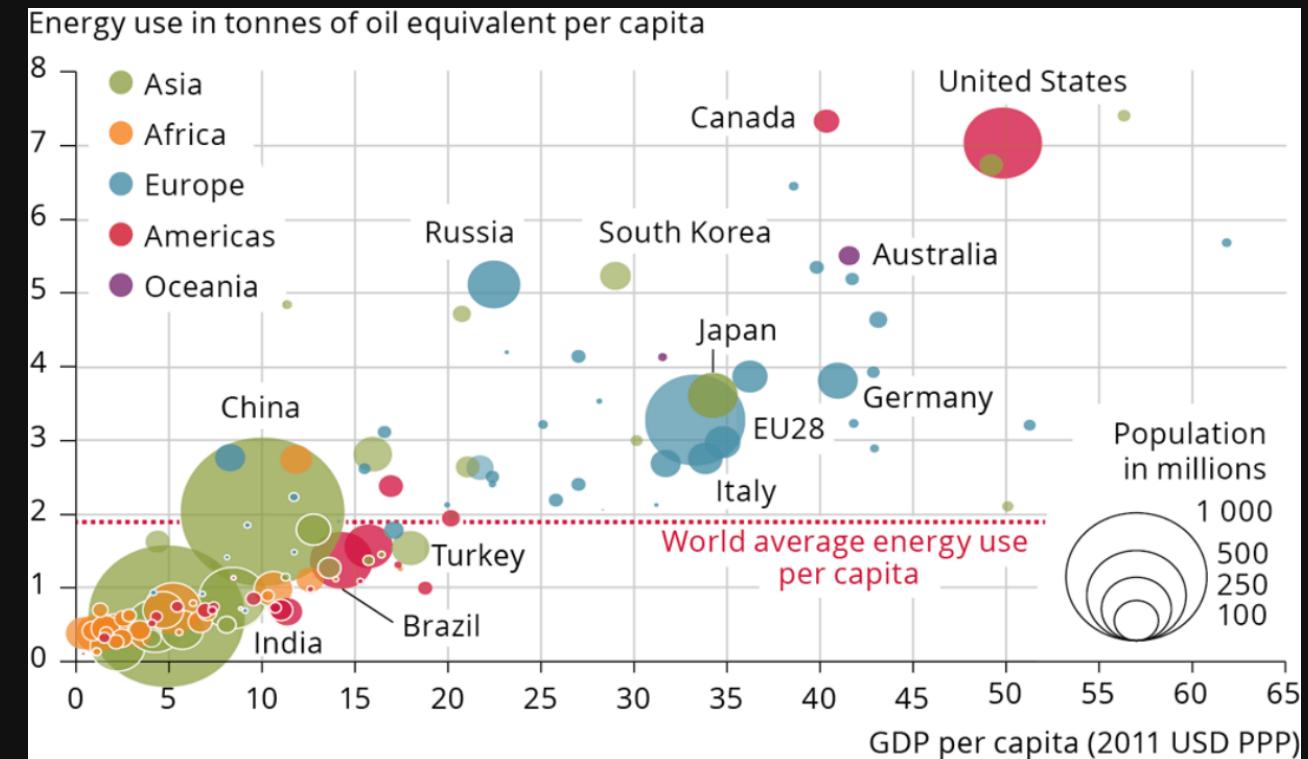


# 数据化思维与数据可视化

- 以高效性、建设性方式的讨论结果，理解运营管理与业务性能之间的关系，提供直接有效的决策依据
- 在庞大繁杂的资源环境中，有目的的挖掘数据背后的特征信息，揭示不同思考角度下的潜在客观规律

## 数据可视化的不同境界

- Level 1 有强烈的用可视化技术去呈现分析结果的欲望
- Level 2 针对具体的问题能够快速确定可视化实施方案
- Level 3 养成一种能用图说清楚的事绝不用文字的习惯
- Level 4 有意识地提升可视化结果的审美和自解释属性



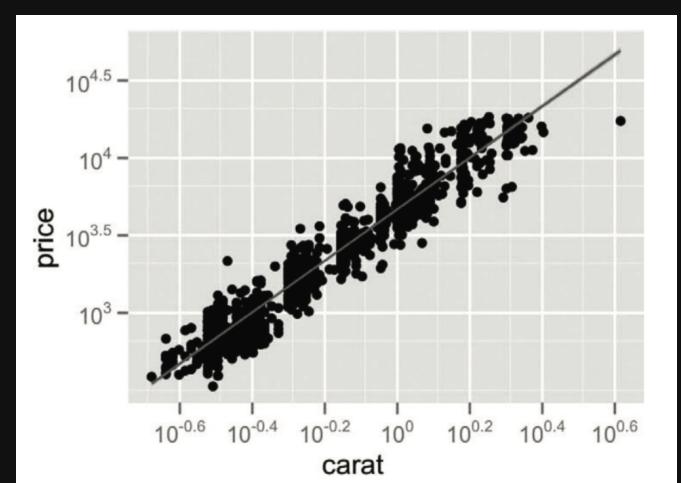
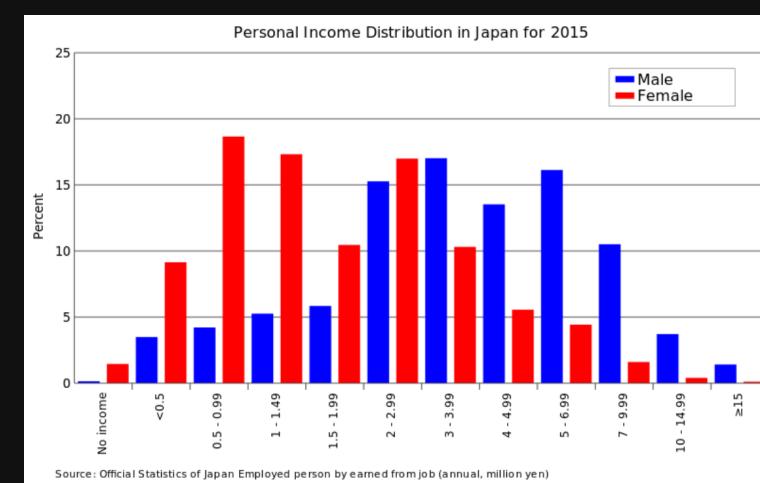


# 数据可视化：目的与原则

数据

解释数据 + 信息传递  
压缩信息 + 突出观点

- 比较两个数据的大小 (可以很直观的做比较，可视化建立起数据和人之间的桥梁)
- 校验两组数据的差异 (比如看两组数据是否来自同一个分布，数据可视化可以帮到你！)
- 统计随机变量的分布 (如分位数、 $3\sigma$ 区间等统计量的可视化，可以不言自明的获得信息！)





# 数据可视化：目的与原则

数据

设计

**解释数据 + 信息传递  
压缩信息 + 突出观点**

- 比较两个数据的大小 (可以很直观的做比较, 可视化建立起数据和人之间的桥梁)
- 校验两组数据的差异 (比如看两组数据是否来自同一个分布, 数据可视化可以帮到你! )
- 统计随机变量的分布 (如分位数、 $3\sigma$ 区间等统计量的可视化, 可以不言自明的获得信息! )

**Simpler is almost always better!**

高效的表达 = 清晰的思路 = 有价值的工作

- 尽量用低维度表达
- 要避免视觉噪声 (标记、注释、颜色) 切忌喧宾夺主!
- 确保各个元素清晰





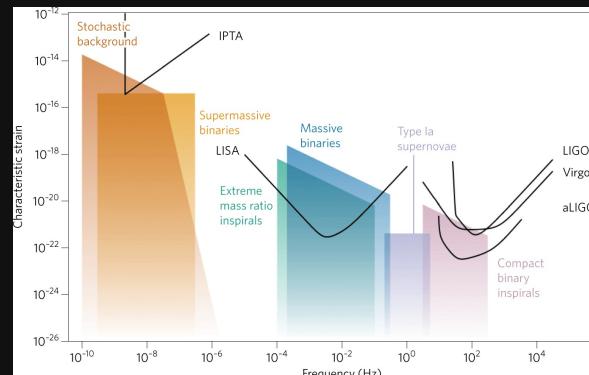
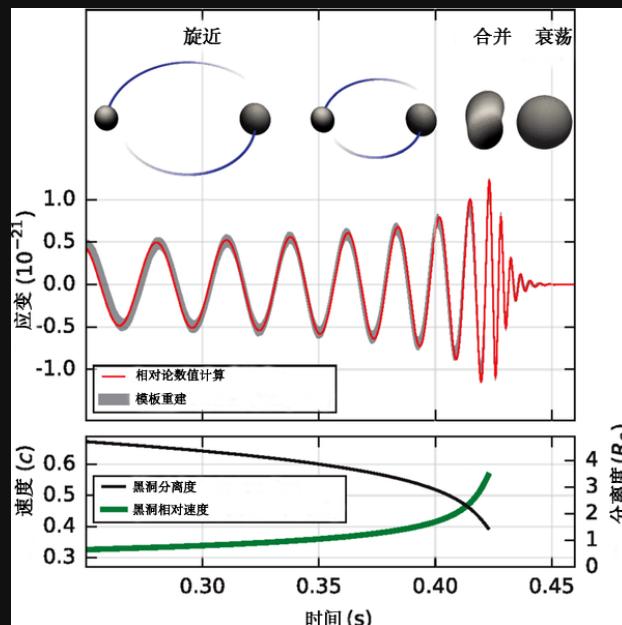
# 数据可视化：目的与原则

数据

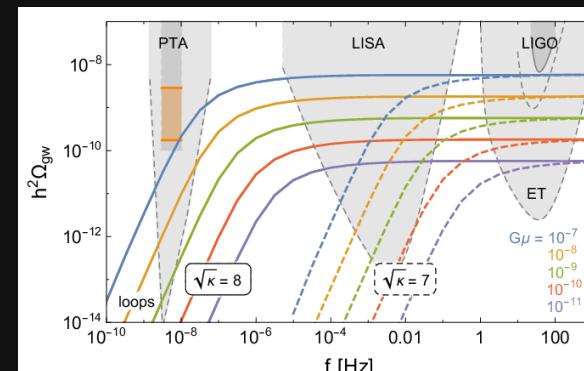
设计

解释数据 + 信息传递  
压缩信息 + 突出观点

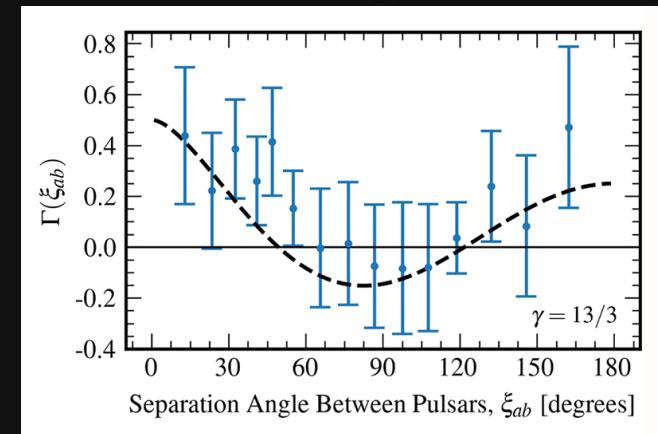
- 比较两个数据的大小 (可以很直观的做比较, 可视化建立起数据和人之间的桥梁)
- 校验两组数据的差异 (比如看两组数据是否来自同一个分布, 数据可视化可以帮到你! )
- 统计随机变量的分布 (如分位数、 $3\sigma$ 区间等统计量的可视化, 可以不言自明的获得信息! )



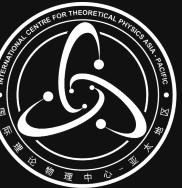
Lommen, A. Pulsar timing for gravitational wave detection.  
Nat Astron 1, 809–811 (2017).



Stochastic gravitational-wave background from metastable cosmic strings -  
Buchmuller, Wilfried et al - arXiv:2107.04578 CERN-TH-2021-107DESY 21-101



G. Agazie et al., "The NANOGrav 15 yr data set: Evidence for a gravitational-wave background," *Astrophys. J., Lett.* **951**, L8 (2023).



# 数据可视化：模式与对象

表现形式

- 一般来说，数据可视化的表现形式（模式），有三种：交互式、交互呈现式和呈现式。
  - 呈现式：用于展示/讲述，服务于群体（本课程的内容）
  - 交互式：用于引导/发现，服务于个体

分类	用户交互	图像生成	目标受众	传媒媒介
交互式	客户完全掌控数据	实时	少量个体	网络交互软件
交互呈现式	客户部分掌控数据	实时	少量群体	网络自助设备
呈现式	客户不能掌控数据	预先	大量群体	视频演示文稿

<https://data.cardifgravity.org/waveform-fitter/>





# 数据可视化：模式与对象

表现形式

- 一般来说，数据可视化的表现形式（模式），有三种：交互式、交互呈现式和呈现式。
    - 呈现式：用于展示/讲述，服务于群体（本课程的内容）
    - 交互式：用于引导/发现，服务于个体

分类	用户交互	图像生成	目标受众	传媒媒介
交互式	客户完全掌控数据	实时	少量个体	网络交互软件
交互呈现式	客户部分掌控数据	实时	少量群体	网络自助设备
呈现式	客户不能掌控数据	预先	大量群体	视频演示文稿



<https://iphysresearch-visiblegwevents-test-firebase-4v0f1o.streamlit.app/>

**demo\_page.py**

```
src > tl_gan > demo_page.py ...
1 import streamlit as st
2 import numpy as np
3 import pandas as pd
4 import gan_model
5
6 st.write("""
7 # Welcome to Streamlit
8
9 The best way to explore and share machine learning results.
10
11 This is a **Rideshare Example***
12 """
13
14
15
```

**SSH: demo-shadow** Python 3.7.3 64-bit (insight): conda Spaces: 4 UTF-8 LF Python

**self-driving-app.py**

```
src > self-driving-app.py ...
166 image = image[1, :, 1, 43] # BGR -> RGB
167 return image
168
169 # Run the YOLO model to detect objects.
170 def detect(image, confidence_threshold, overlap_threshold):
171     # Load the YOLO model. Because this is cached it will only happen once.
172     get_cacheloader_useTrue()
173     net = loadNetwork(configPath, weightsPath)
174     net.setInput(blob)
175     output_layer_names = [output_layer_name[i][0] + 1 for i in net.getUnconnectedOutLayers()]
176     net.setInput(blob)
177     net.setOutput(*output_layer_names)
178
179     # Suppress detections in case of too low confidence or too much overlap.
180     boxes, confidences, class_ids = [], [], []
181     H, W = image.shape[1:2]
182
183     for output in net.getOutputs():
184         for detection in output:
185             scores = detection[5:]
186             classID = np.argmax(scores)
187             confidence = scores[classID]
188             if confidence > confidence_threshold:
189                 box = detection[0:4] + np.array([H, W, H])
190                 center_x, center_y, width, height = box.astype("int")
191                 x_s, y_s = int(center_x - (width / 2)), int(center_y - (height / 2))
192                 box = np.append([x_s, y_s, int(width), int(height)])
193                 confidences.append(confidence)
194                 class_ids.append(classID)
195
196     indices = cv2.dnn.NMSBoxes(boxes, confidences, confidence_threshold, overlap_threshold)
197
198     # Map from YOLO labels to Udacity labels.
199     UDACITY_LABELS = {
200         0: "pedestrian",
201         1: "biker",
202         2: "car",
203         3: "truck",
204         5: "truck",
205         7: "truck",
206         9: "trafficlight"
207     }
208
209     xmin, ymin, ymax, labels = [], [], [], []
210     for idx in indices:
211         # Loop over the indexes we are keeping
212         if len(indices) > 0:
213             # Loop over the indexes we are keeping
```

**Streamlit**

# Welcome to Streamlit

The best way to explore and share machine learning results.

This is a Rideshare Example

**Ground Truth**

Human-annotated data (frame 304)

**Real-time Computer Vision**

YOLO v3 Model (overlap 0.0 ) (confidence 0.6)



# 数据可视化：模式与对象

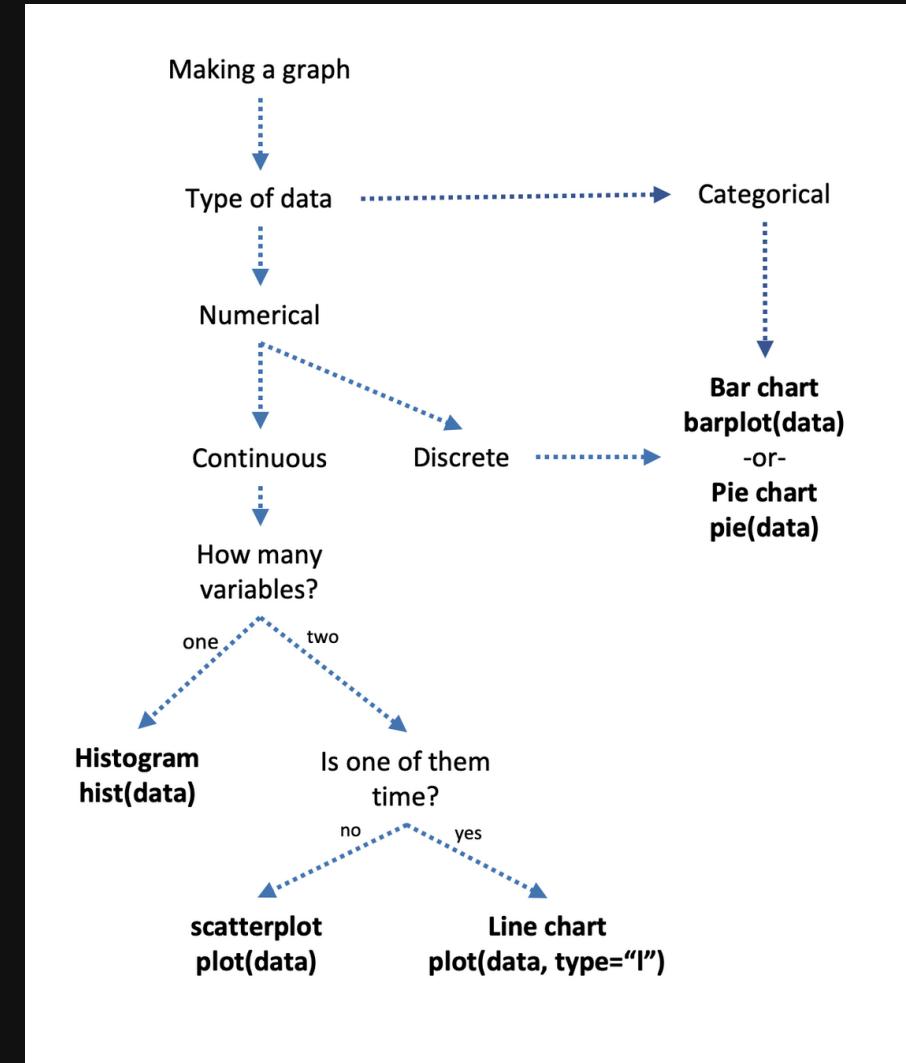
表现形式      数据类型

- 一般来说，数据可视化的数据（对象）有三类概念：（本课程的重点）
  - 定量数据（Quantitative）：连续、离散
  - 定类数据（Categorical）：城市、品类等
  - 定序数据（Ordinal）：尺码、态度等

	离散数据	连续数据
有序数据	尺码、自然数	温度、经纬度
无序数据	形状、类别	方向、色彩

## Tips

- 面对某一列特征，心中要先掂量一下：
  - 这是离散的？连续的？
  - 然后再掂量是不是定序的？





# 数据可视化：模式与对象

表现形式

数据类型

- 定量数据 (Quantitative)

位置 > 长度/角度 > 面积 > 体积 > 密度 > 颜色

上述为优选关系，体现了低维度优先的原则。

- 定序数据 (Ordinal)

位置 > 密度 > 颜色 > 连接 > 包含

其中的密度可以通过疏密程度来体现；

颜色主要是通过深浅体现，避免视觉噪声；

连接可以用箭头等从属关系来体现有序性。

- 定类数据 (Categorical)

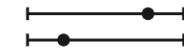
位置 > 颜色 > 连接 > 包含 > 形状

表现类之间的关系，确保元素清晰。

Channels: Expressiveness Types and Effectiveness Ranks

④ **Magnitude Channels: Ordered Attributes**

Position on common scale



Position on unaligned scale



Length (1D size)



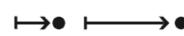
Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



④ **Identity Channels: Categorical Attributes**

Spatial region



Color hue



Motion



Shape



▲ ↑ Effectiveness ↓ ▼

[ Same ] ↑ Least ↓



# 数据可视化：模式与对象

表现形式

数据类型

- 定量数据 (Quantitative)

位置 > 长度/角度 > 面积 > 体积 > 密度 > 颜色

上述为优选关系，体现了低维度优先的原则。

- 定序数据 (Ordinal)

位置 > 密度 > 颜色 > 连接 > 包含

其中的密度可以通过疏密程度来体现；

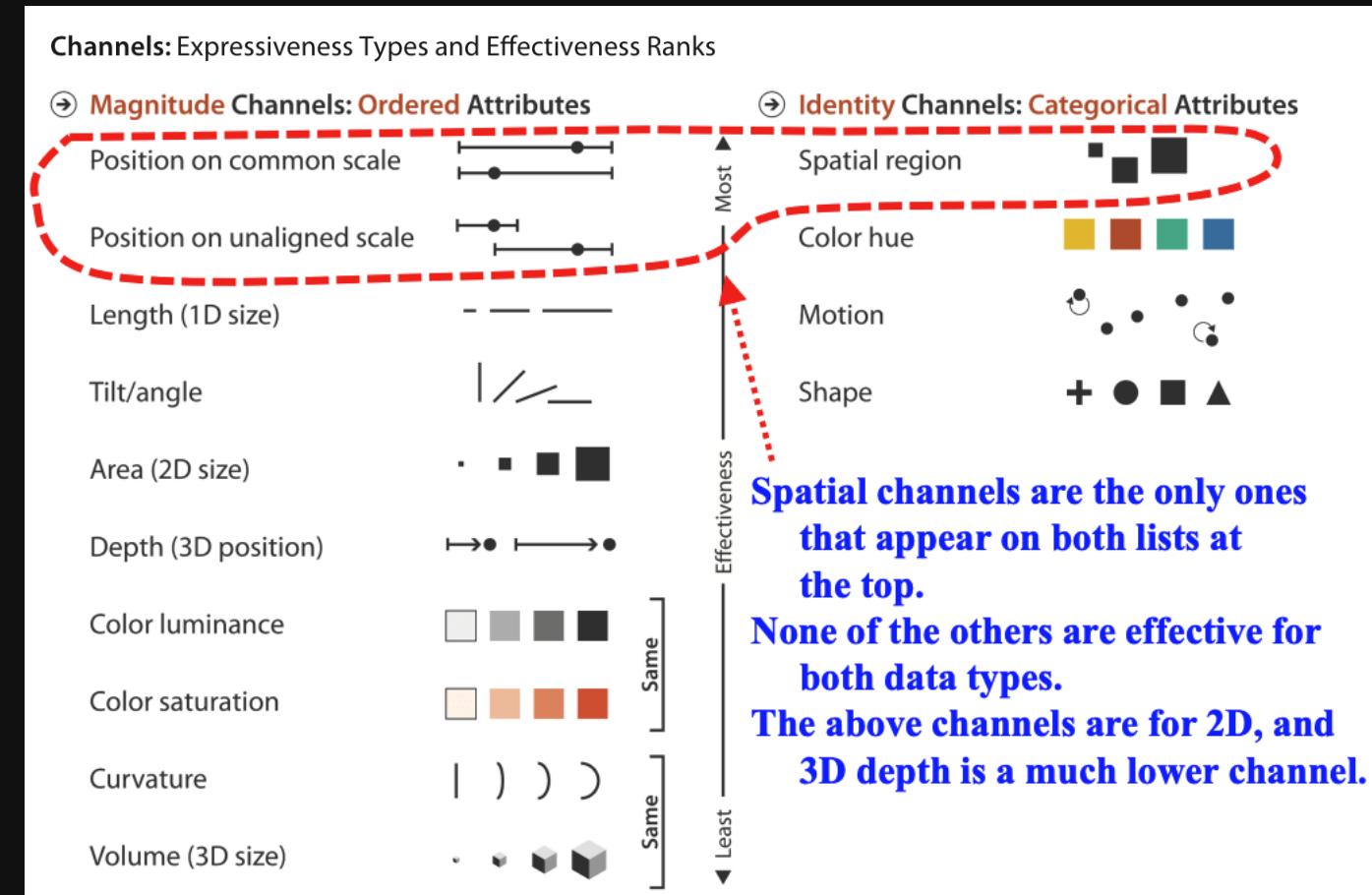
颜色主要是通过深浅体现，避免视觉噪声；

连接可以用箭头等从属关系来体现有序性。

- 定类数据 (Categorical)

位置 > 颜色 > 连接 > 包含 > 形状

表现类之间的关系，确保元素清晰。





# 数据可视化：模式与对象

## 表现形式

# 数据类型

- 定量数据 (Quantitative)

位置 > 长度/角度 > 面积 > 体积 > 密度 > 颜色

上述为优选关系，体现了**低维度优选**的原则。

- 定序数据 (Ordinal)

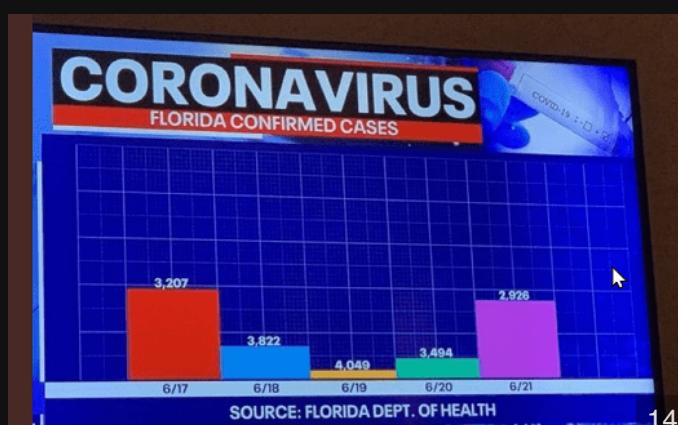
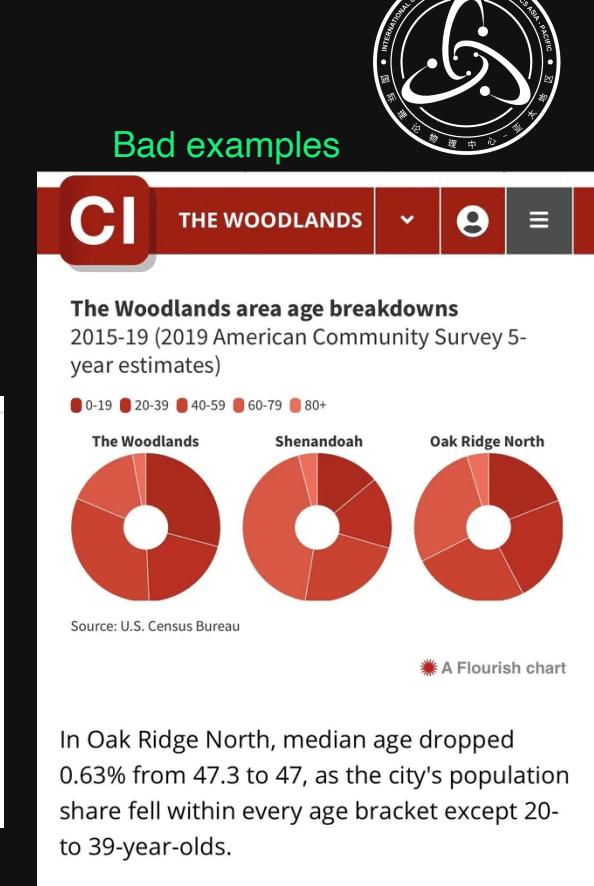
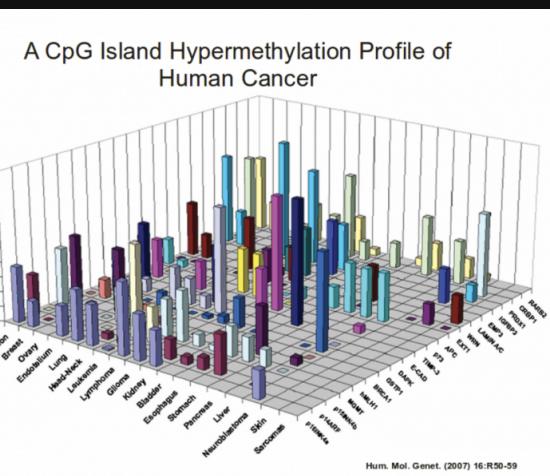
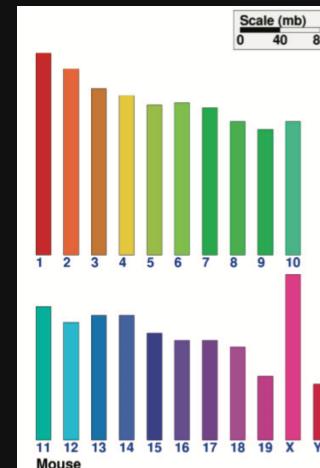
位置 > 密度 > 颜色 > 连接 > 包含

其中的密度可以通过疏密程度来体现；  
颜色主要是通过深浅体现，**避免视觉噪声**；  
连接可以用箭头等从属关系来体现有序性。

- 定类数据 (Categorical)

位置 > 颜色 > 连接 > 包含 > 形状

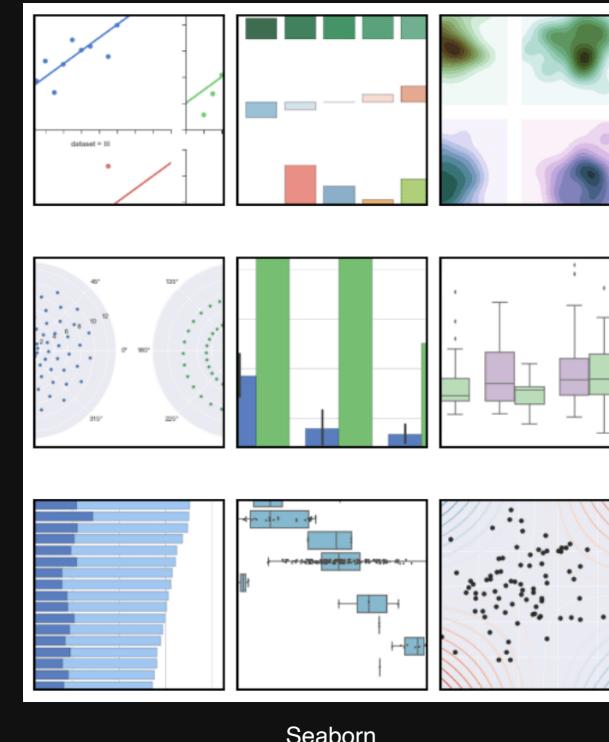
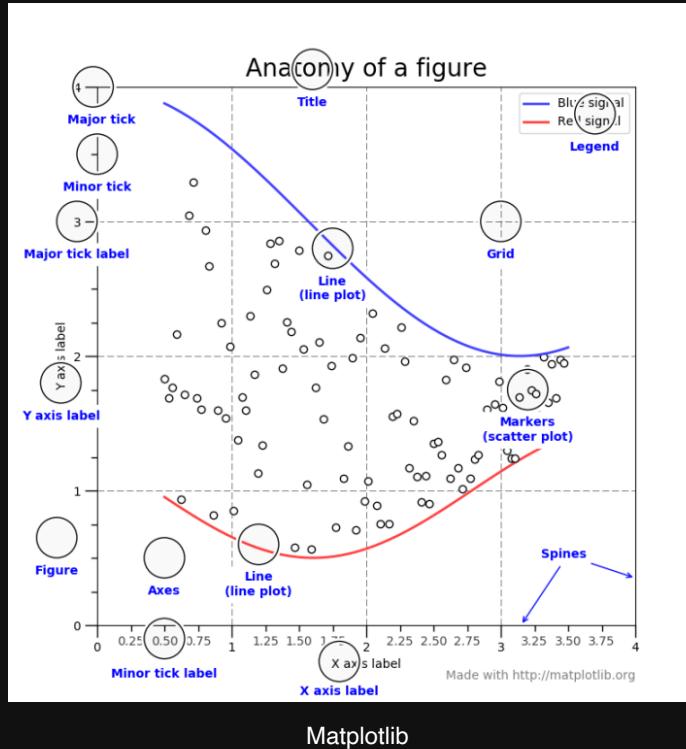
表现类之间的关系，确保元素清晰。



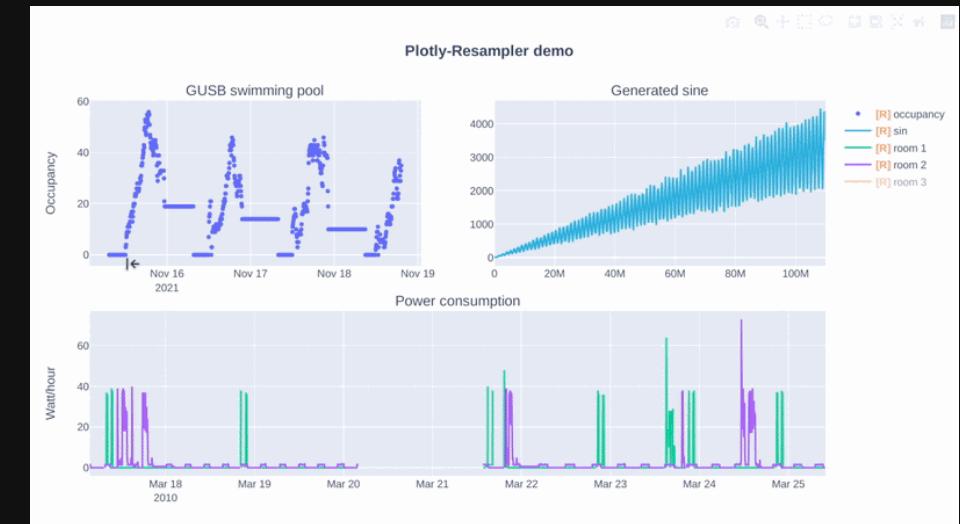


# 数据可视化：常见工具

- 基于编程语言
  - RSutdio, Matplotlib, Seaborn, Bokeh, Plotly, Streamlit, Gradio, ...
- 不基于编程语言
  - Plotly, Tableau, Icharts, QlikView, FineBI, Power BI, Infogram, RAW Graphs, ...



- 本课程是基于 Python 编程语言，讲解开源免费的数据可视化工具。
  - **Matplotlib**: 满足基本的需求（用的好可以满足所有的需求，就是用起来太麻烦）
  - **Seaborn**: 满足颜控的需求（非常漂亮！非常容易！）
  - **Bokeh**: 满足交互呈现的需求
  - **Plotly**: 强大的在线交互可视化框架
  - **Streamlit**: 专注于机器学习和数据科学团队的用户交互可视化 app

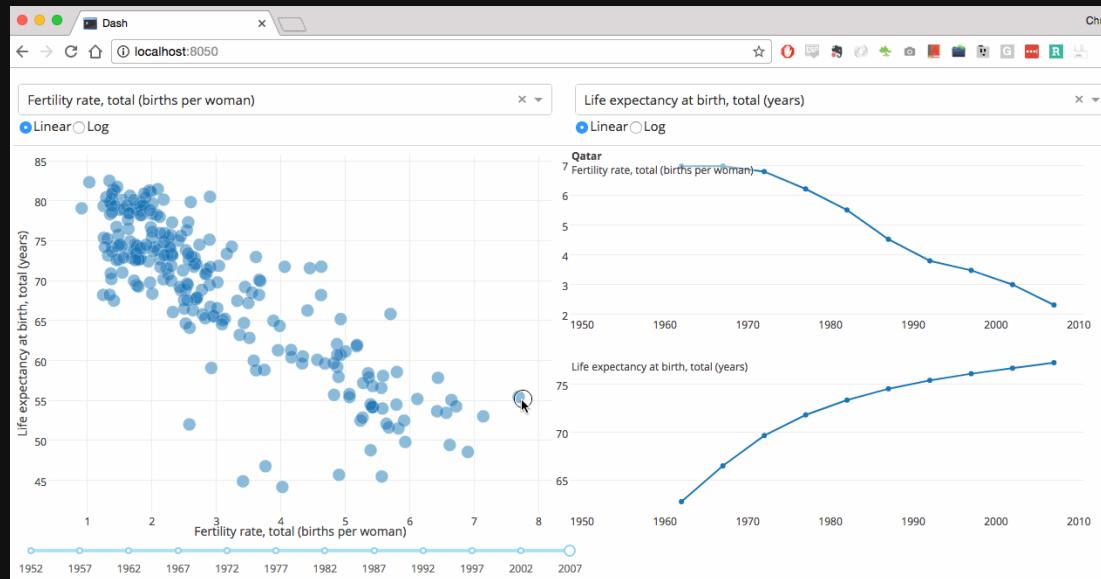




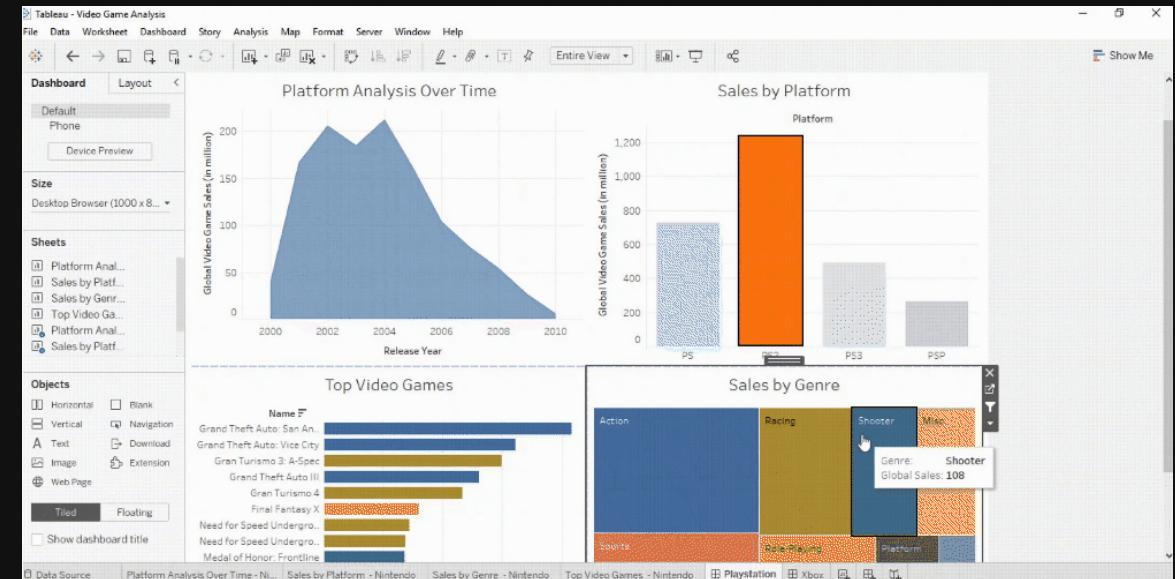
# 数据可视化：常见工具

- 基于编程语言
  - RSutdio, Matplotlib, Seaborn, Bokeh, Plotly, Streamlit, Gradio, ...
- 不基于编程语言
  - Plotly, Tableau, Icharts, QlikView, FineBI, Power BI, Infogram, RAW Graphs, ...

- 本课程是基于 Python 编程语言，讲解开源免费的数据可视化工具。
  - **Matplotlib**: 满足基本的需求（用的好可以满足所有的需求，就是用起来太麻烦）
  - **Seaborn**: 满足颜控的需求（非常漂亮！非常容易！）
  - **Bokeh**: 满足交互呈现的需求
  - **Plotly**: 强大的在线交互可视化框架
  - **Streamlit**: 专注于机器学习和数据科学团队的用户交互可视化 app

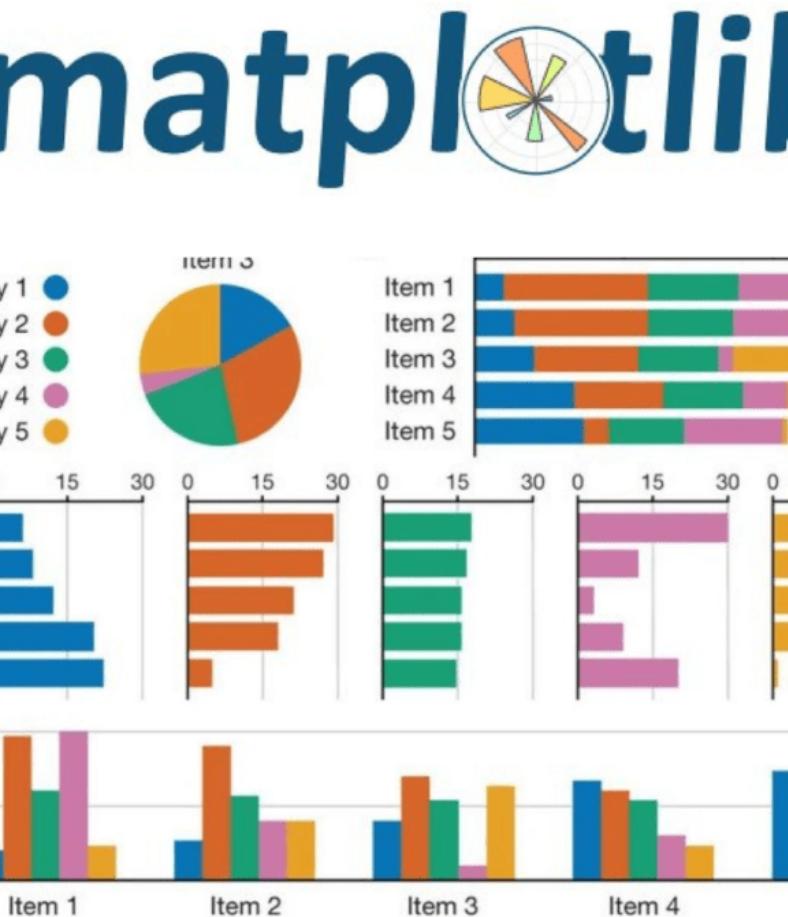


Plotly



Tableau

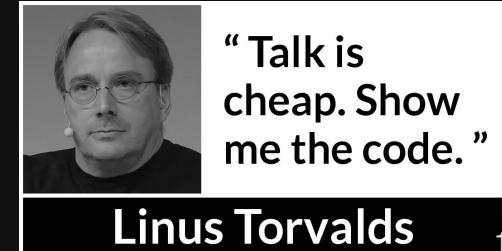
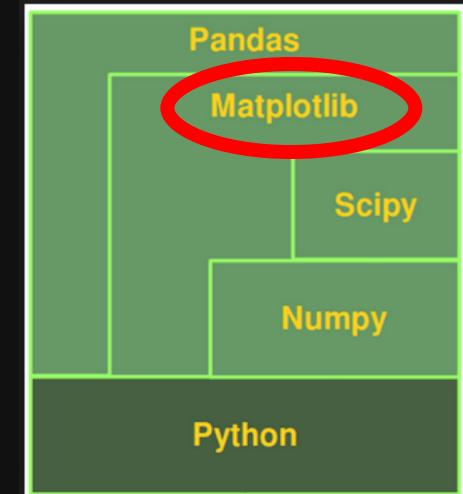
# Data Jisualization in Python using

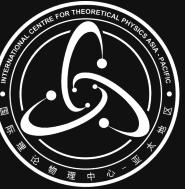


matplotlib

## 数据分析可视化之 Matplotlib

- 数据可视化：基本处理流程
  - 数据准备
  - 确定图表
  - 分析迭代
  - 输出结论

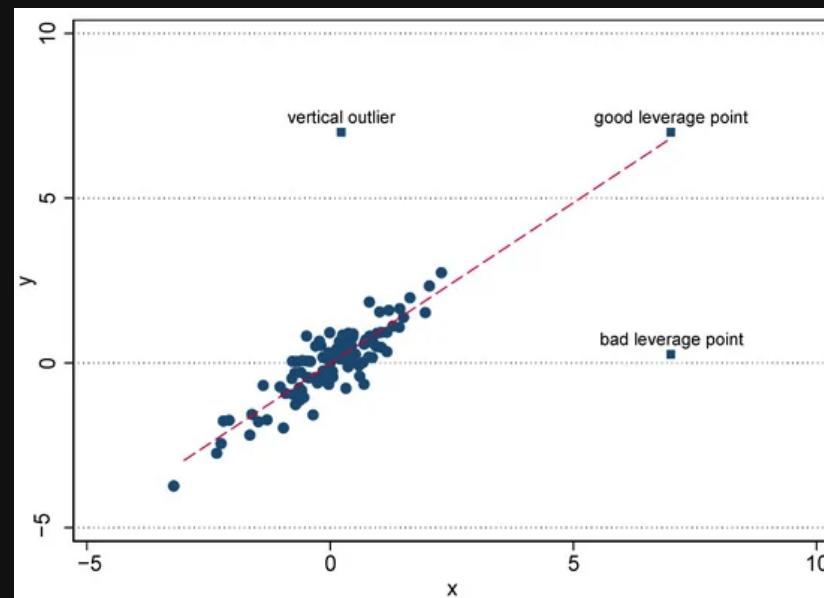




# 数据可视化：基本处理流程

## 1. 数据准备

- 数据规模：数据分组、数据采样（处理大数据时尤为需要）
- 数据类型：数值数据、分类数据（一定要对数据结构特别清楚：连续？离散？有序吗？）
- 数据异常：取值异常、数据缺失（通常应对的是数据爬虫等采集后的原始数据）



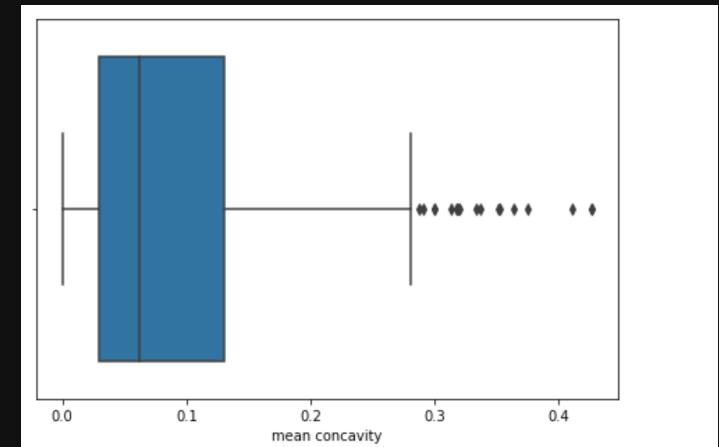
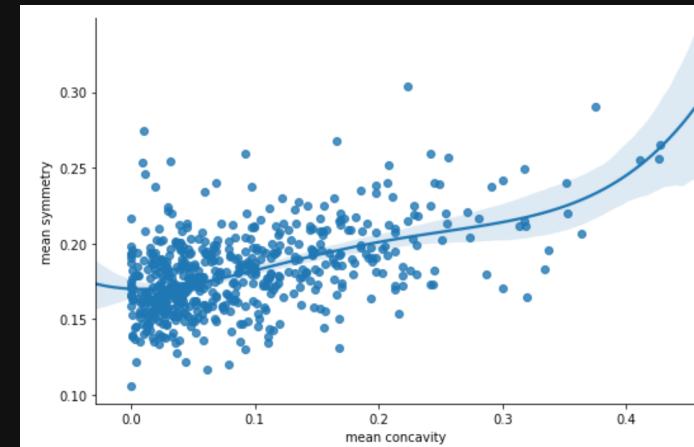
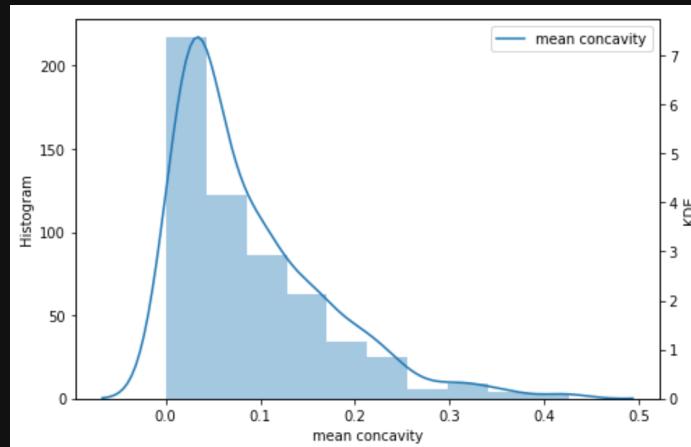
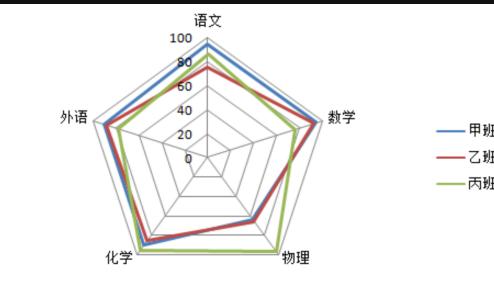
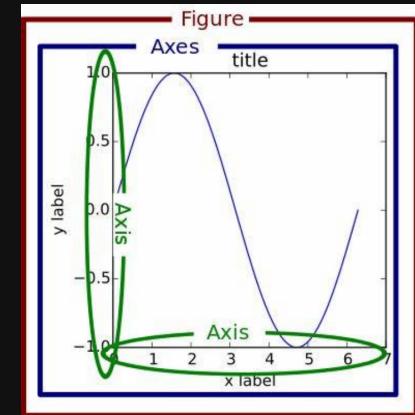


# 数据可视化：基本处理流程

## 1. 数据准备

## 2. 确定图表。数据可视化里通常面临的三类问题：

- **关联分析、定量数值比较**: 散点图, 曲线图 (scatter, plot)
- **分布分析 (定量数据: 粗粒度 / 细粒度)**: 灰度图, 密度图 (hist, gaussian\_kde, plot)
- **分类分析 (关于定序 / 定类数据)**: 柱状图, 箱式图 (bar, boxplot)

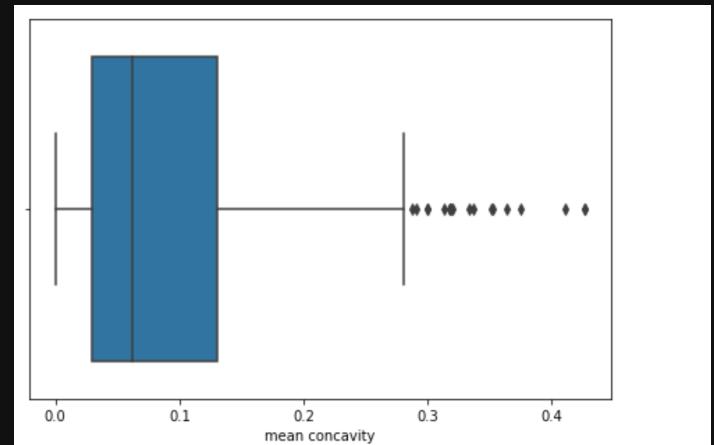
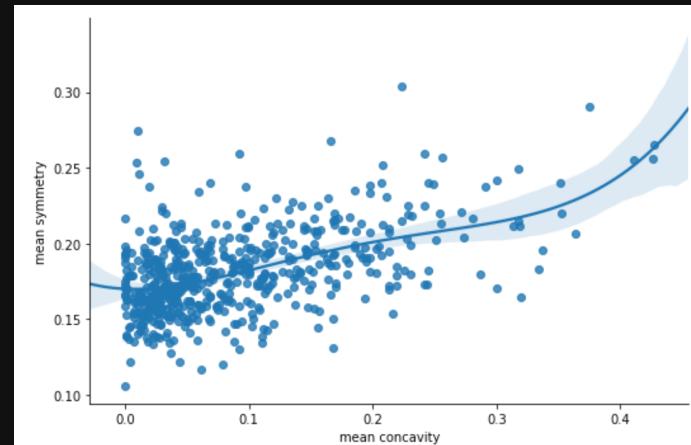
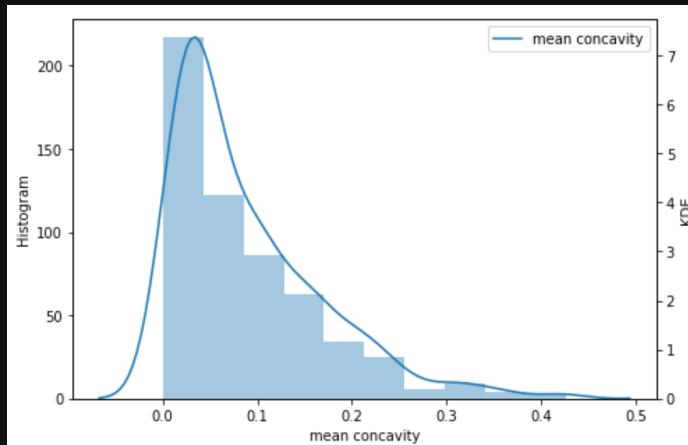


# 数据可视化：基本处理流程

1. 数据准备
2. 确定图表
3. 分析迭代

- 确定拟合模型: OLS, fit OLS = 最小二乘; fit = 拟合
- 分析拟合性能: summary\_table 统计学汇总
- 确定数据分布: hist
- 确定重点区间: quartile 分布的上下四分位数, 以及各分位数之间的区间

<b>count</b>	569
<b>mean</b>	0.088799
<b>std</b>	0.079720
<b>min</b>	0.000000
<b>25%</b>	0.029560
<b>50%</b>	0.061540
<b>75%</b>	0.130700
<b>max</b>	0.426800

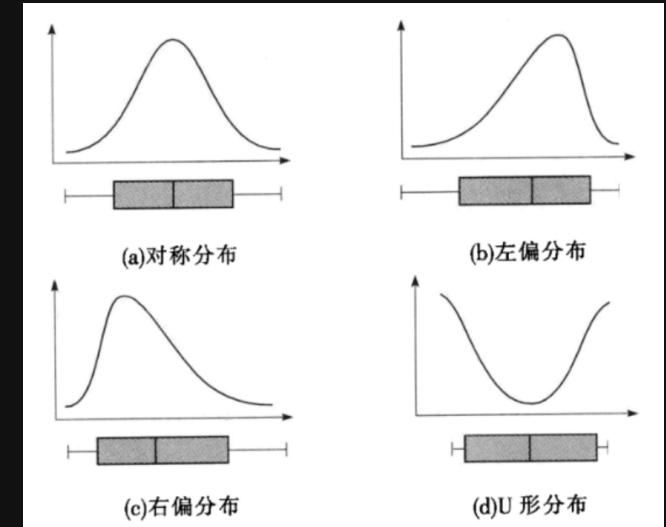
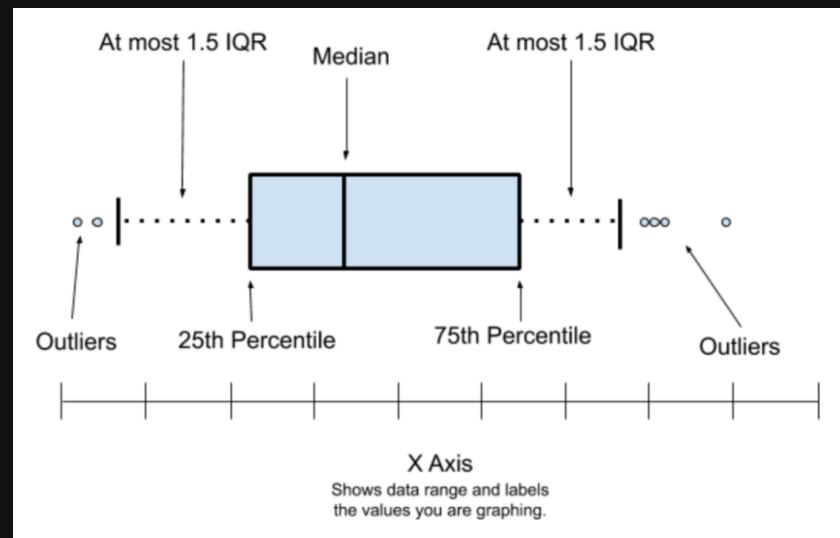


分析迭代的要素，不仅依赖于数据本身，也依赖于人的分析角度



# 数据可视化：基本处理流程

1. 数据准备
2. 确定图表
3. 分析迭代
  - 箱型图 Boxplot: 观察分布的对称性和偏性

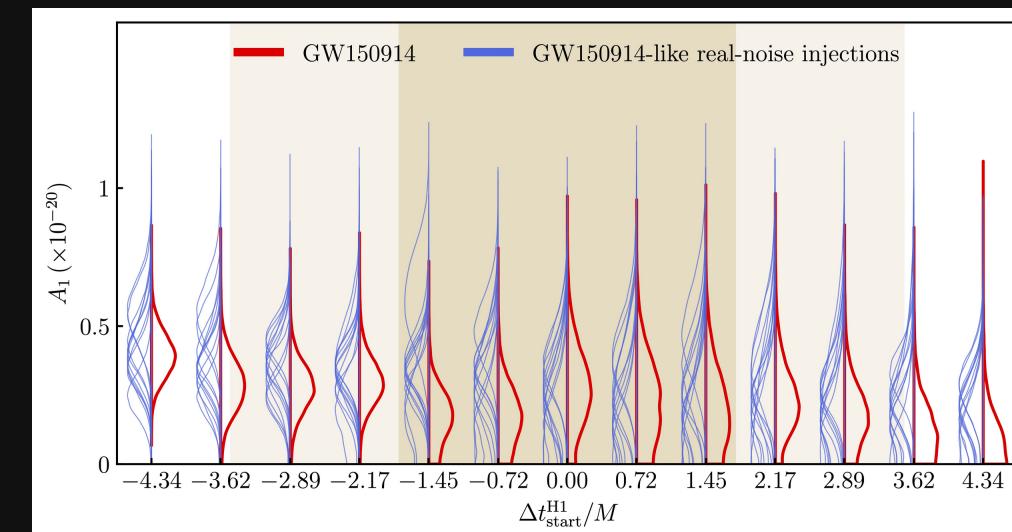
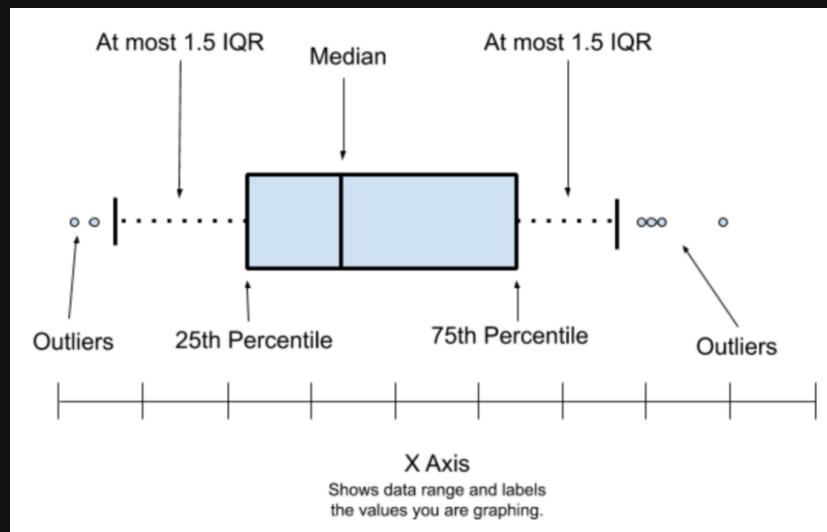
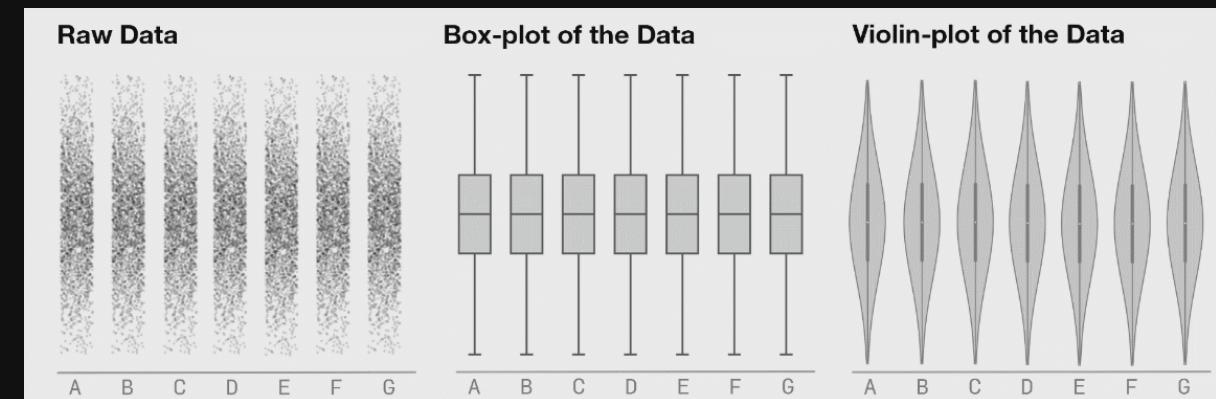




# 数据可视化：基本处理流程

1. 数据准备
2. 确定图表
3. 分析迭代

- 箱型图 Boxplot: 观察分布的对称性和偏性
- 箱型图的局限性: 压缩信息的代价。。。
  - 提琴图



Red violin plots: reconstructed amplitude of the first overtones of GW150914 for different estimates of the waveform peak time. Blue violin plots: amplitudes inferred by injecting a GW150914-like signal in different noise realizations at those same starting times. (Phys. Rev. Lett. 129, 111102)



# 数据可视化：基本处理流程

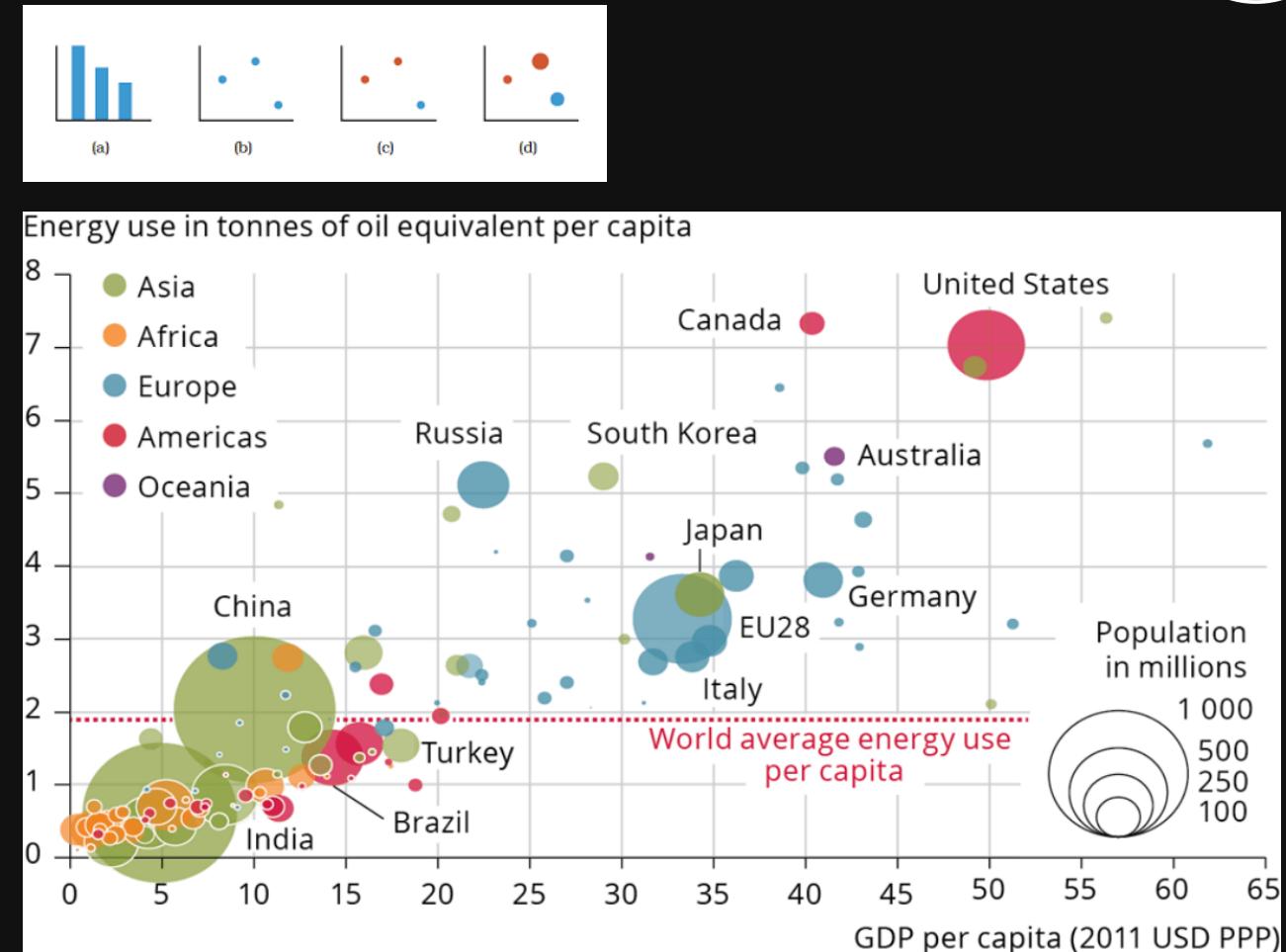
1. 数据准备
2. 确定图表
3. 分析迭代
4. 输出结论

- 养成看图说话的好习惯
- 提出一个好问题，画出一个好图像，给出一个好结论

Intensified global competition for resources?



Global demand vs GDP per person

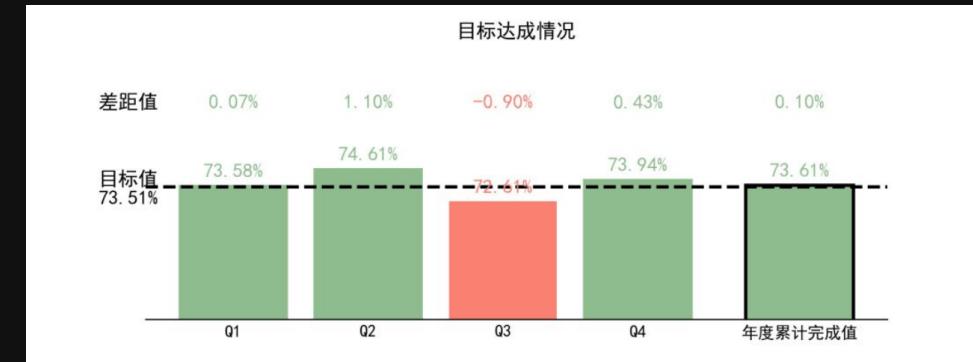


Data source: World Bank World development indicators Note: The graph shows the correlation of national per capita energy consumption and per capita GDP. The size of the bubbles denotes total population per country. All values refer to the year 2011.



# 数据可视化：思维方式

季度	实际值	目标值	差距值 (实际值-目标值)
<b>Q1</b>	73.58%	73.51%	<b>0.07%</b>
<b>Q2</b>	74.61%	73.51%	<b>1.10%</b>
<b>Q3</b>	72.61%	73.51%	<b>-0.90%</b>
<b>Q4</b>	73.94%	73.51%	<b>0.43%</b>
<b>年累计</b>	<b>73.61%</b>	<b>73.51%</b>	<b>0.10%</b>



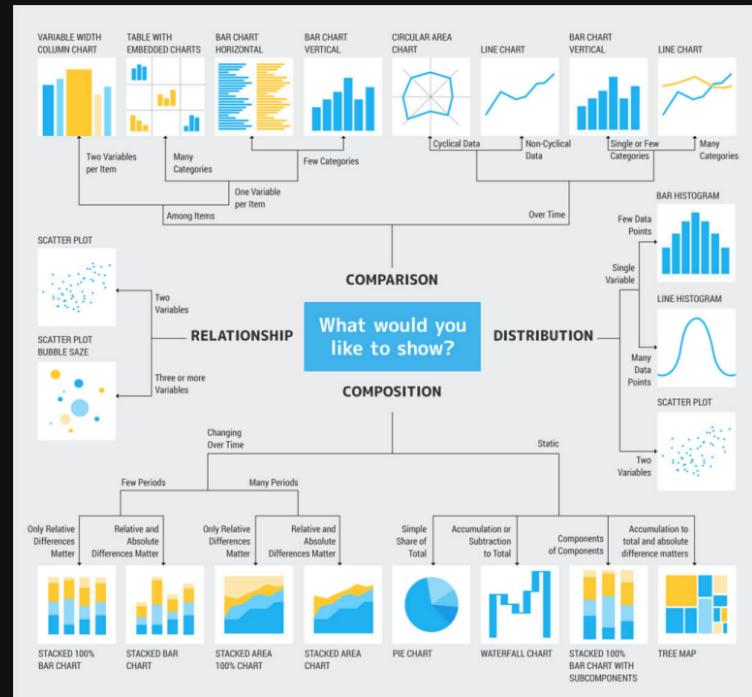
图片来源：CrossHands - AhongPlus

## 问题导向

- 了解数据来源的背景或数据应用场景，以问题为导向的探索性分析思路，以得到鲜明的观点和分析结果为目标。

## 方法导向

- 熟悉可视化工具，在方法论层面上充分积累，更多的是注重工具本身，以实现经过精心设计可视化呈现为目标。



图片来源：<https://vizard.co/tableau-interview-questions/>



# 数据可视化：思维方式

马世权老师在「乐见数据：商业数据可视化思维」里提出，一个成功的商业数据可视化要满足两要素：

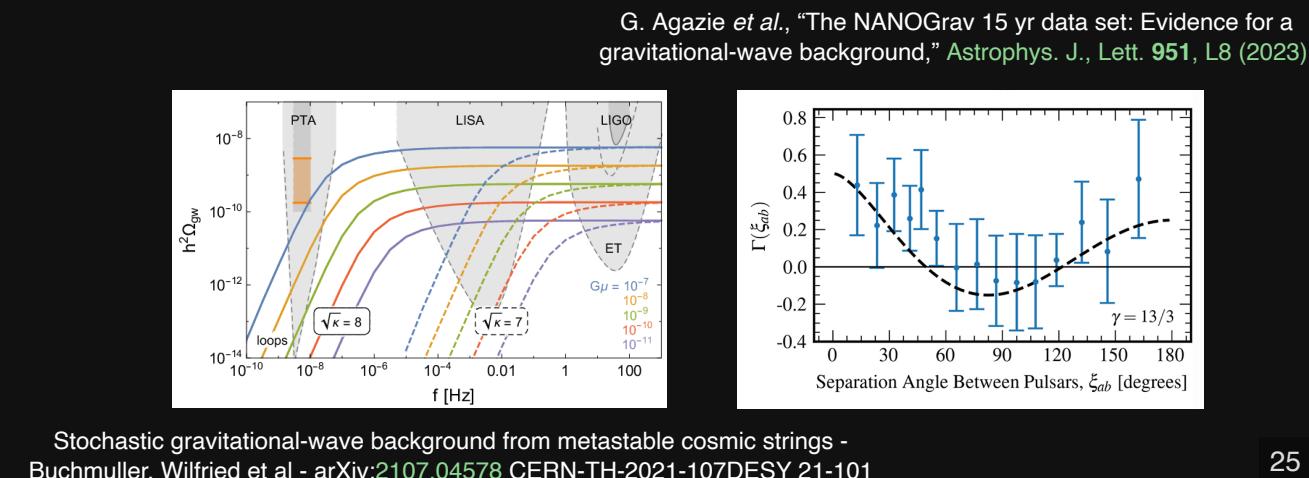
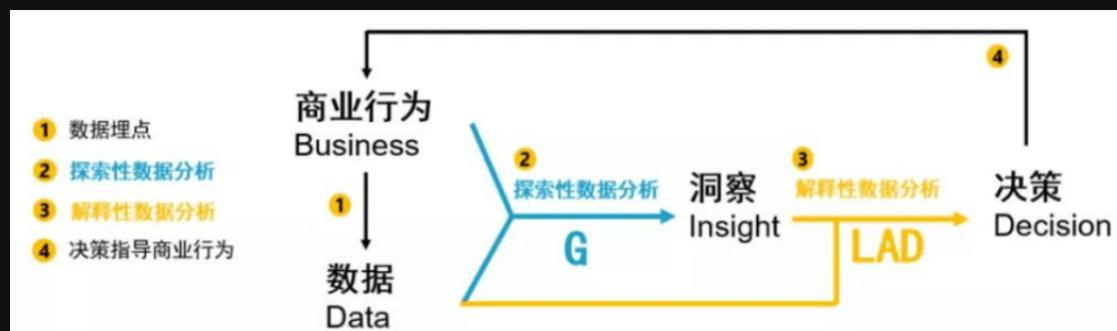
- 提供足够的商业价值      恰如其分地表达科学观点
- 帮助读者快速理解信息      帮助读者快速理解观点

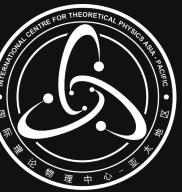
那什么是好的商业数据可视化图表？

- 答案：符合 **GLAD** 原则的图表



图片来源于网络





# 数据可视化：思维方式

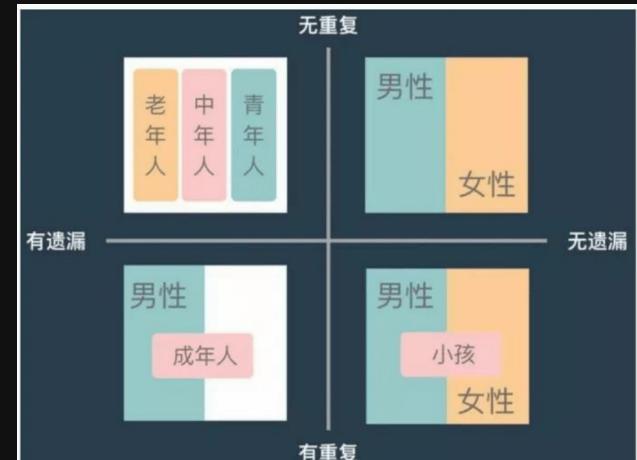
- MECE: Mutually Exclusive, Collectively Exhaustive. 即,相互独立,完全穷尽,不重叠,不遗漏,是麦肯锡在金字塔原理中提出的思维原则。
- 如果能在解决问题、沟通问题、撰写报告时熟练运用该原则,能够让自己的逻辑更清晰、考虑问题更全面,也更容易让人记忆和理解。

**GLAD** 原则:

	中文含义	探索性数据分析	思考问题
G	图表的灵魂: 发现好数据和好洞察	数据是否恰当 洞察能否恰当	技术
L	降噪: 简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	
A	精准表达: 提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	
D	画龙点睛: 突出洞察信息的标识	是否有突出洞察的标识	

如类别和度量使用是否恰当

- 类别的不恰当使用:
  - 类别不符合 MECE 原则: 有重复、有遗漏
  - 分类不均匀
- 度量指标的不恰当使用:
  - 绝对值指标与相对值指标混淆
  - 时间段指标与时间点指标混淆



图片来源于网络



# 数据可视化：思维方式

- 下图这里使用绝对值——采购量指标来统计，回答的问题是整体采购量的变化如何。但是无法直接回答采购工作质量如何，是变好还是变差呢？使用相对值百分比统计指标来监控问题的占比更恰当。

**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪里	技术
L	降噪：简约至上	特效/颜色 辅助信息	
A	精准表达：提升数据表达的准确度	图形元素 数据密度 数据显示	
D	画龙点睛：突出洞察信息的标识	是否有突	

图片来源：  
<https://help.fanruan.com/dv/doc-view-81.html>

如类别和度量使用是否恰当

- 类别的不恰当使用：
  - 类别不符合 MECE 原则：有重复、有遗漏
  - 分类不均匀
- 度量指标的不恰当使用：
  - 绝对值指标与相对值指标混淆
  - 时间段指标与时间点指标混淆





# 数据可视化：思维方式

**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	<u>数据是否恰当</u> 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色 辅助信息	
A	精准表达：提升数据表达的准确度	图形元 数据密度 数据显	
D	画龙点睛：突出洞察信息的标识	是否有	

如类别和度量使用是否恰当

- 类别的不恰当使用：
  - 类别不符合 MECE 原则：有重复、有遗漏
  - 分类不均匀
- 度量指标的不恰当使用：
  - 绝对值指标与相对值指标混淆
  - **时间段指标与时间点指标混淆**
- 在职人数为时间点指标，离职人数为时间段指标，时间段指标与时间点指标混淆不清。这时可以将时间度量统一，都修改为时间段指标。

图片来源：<https://help.fanruan.com/dvg/doc-view-81.html>



# 数据可视化：思维方式

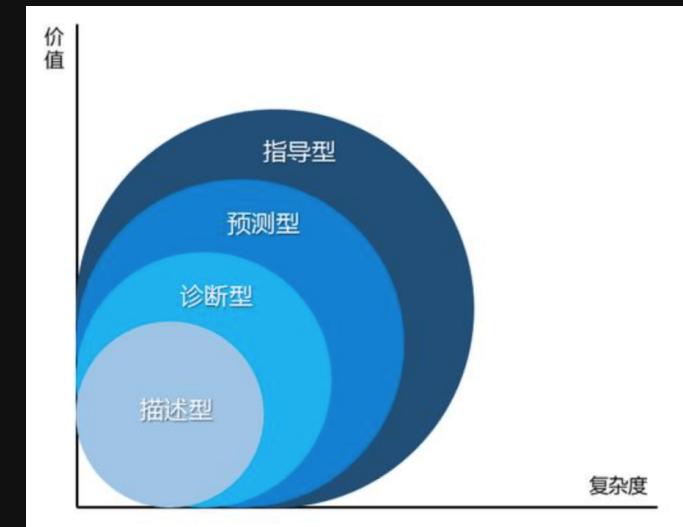
**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	
D	画龙点睛：突出洞察信息的标识	是否有突出洞察的标识	

- 在探索性数据分析领域，一般将数据分析划分为四类：描述型、诊断型、预测型和指导型。
- 这就像医生的工作内容一样：描述型分析是体检，先客观地检查身体健康指标情况，判断是否偏离正常值范围并陈述观点；诊断型分析即通过进一步的询问和信息挖掘诊断出问题是什么，病根在哪里；预测型分析即结合对病人的了解分析病情目前处于哪个阶段，预测病情会怎样发展；指导型分析即最后开出针对性的药方和治疗建议。

描述性分析 < 诊断/预测性分析 < 指导性分析

1. 描述型：观察当前数据发生了什么？
2. 诊断型：理解为什么会发生？
3. 预测型：预测未来会发生什么？
4. 指导型：怎样达到更好地商业决策 (例子)



图片来源：<https://help.fanruan.com/dvg/doc-view-81.html>



# 数据可视化：思维方式

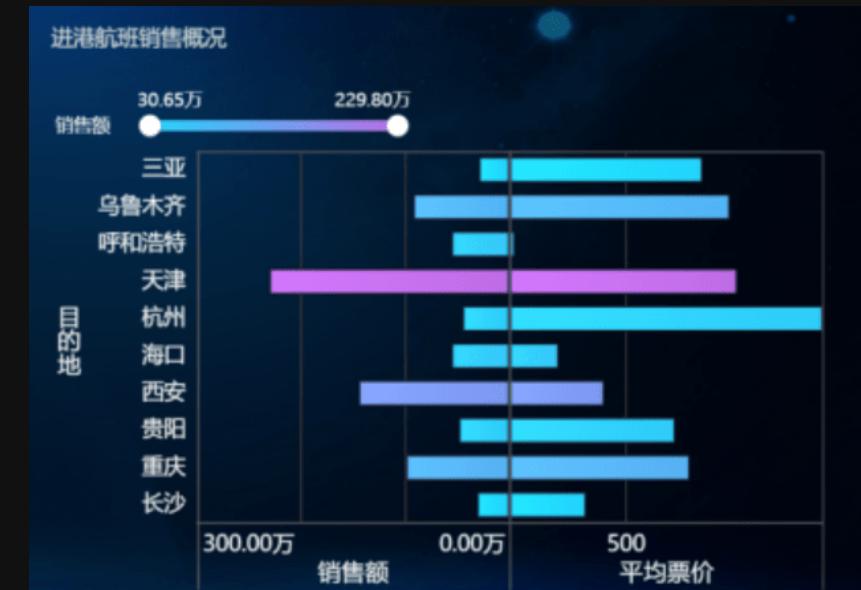
**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	
D	画龙点睛：突出洞察信息的标识	是否有突出洞察的标识	

- 在探索性数据分析领域，一般将数据分析划分为四类：描述型、诊断型、预测型和指导型。
- 这就像医生的工作内容一样：描述型分析是体检，先客观地检查身体健康指标情况，判断是否偏离正常值范围并陈述观点；诊断型分析即通过进一步的询问和信息挖掘诊断出问题是什么，病根在哪里；预测型分析即结合对病人的了解分析病情目前处于哪个阶段，预测病情会怎样发展；指导型分析即最后开出针对性的药方和治疗建议。

描述性分析 < 诊断/预测性分析 < 指导性分析

1. **描述型**: 观察当前数据发生了什么?
2. 诊断型: 理解为什么会发生?
3. 预测型: 预测未来会发生什么?
4. 指导型: 怎样达到更好地商业决策



图片来源：<https://bbs.fanruan.com/thread-135963-1-1.html>

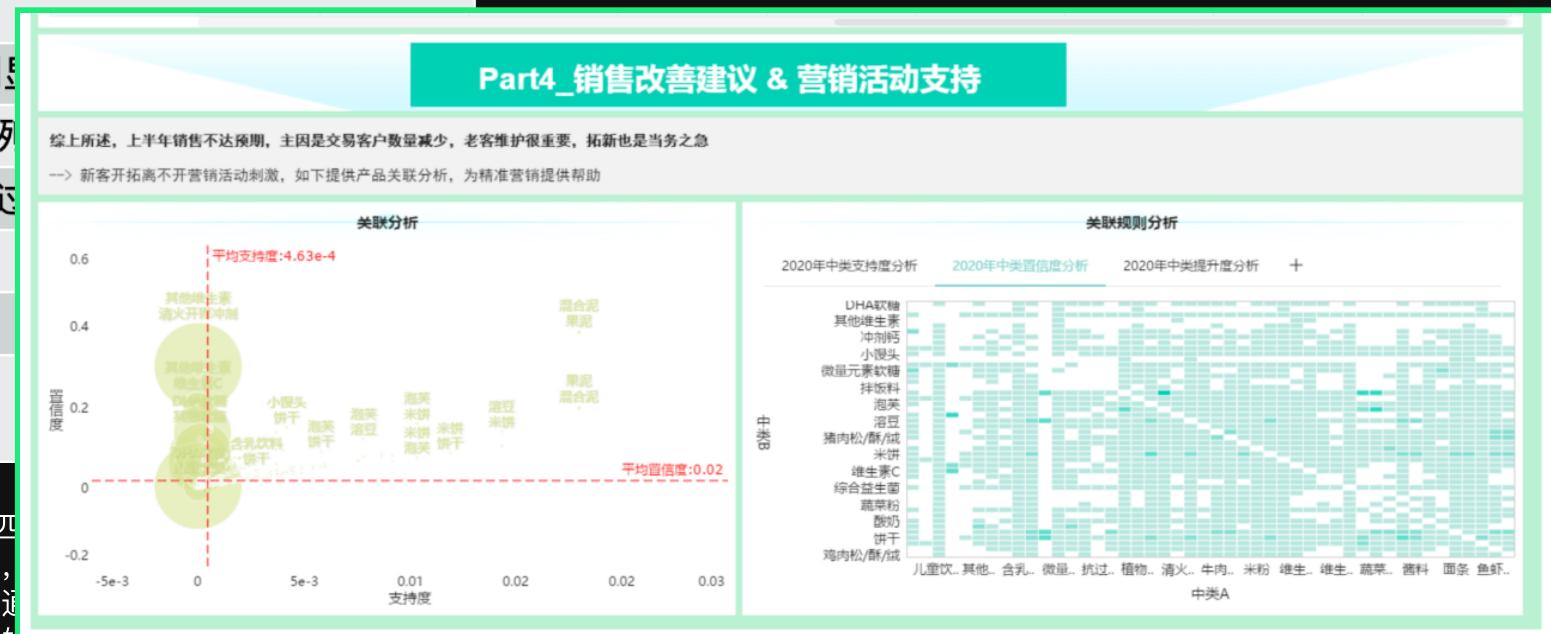


# 数据可视化：思维方式

## GLAD 原则：

中文含义		探索性数据分析	思考问题	技术
G	图表的灵魂：发现好数据和好洞察		数据是否恰当 <u>洞察在哪个层次</u>	
L	降噪：简约至上		特效/颜色字体是否有明显 辅助信息(文字/标签/图标)	
A	精准表达：提升数据表达的准确度		图形元素的精确度是否过 数据密度是否合适 数据显示效果是否准确	<p>综上所述，上半年销售不达预期，主因是交易客户数量减少， → 新客开拓离不开营销活动刺激，如下提供产品关联分析，</p> <p>关联分析</p>
D	画龙点睛：突出洞察信息的标识		是否有突出洞察的标识	

- 在探索性数据分析领域，一般将数据分析划分为四类
  - 这就像医生的工作内容一样： 描述型分析是体检，是否偏离正常值范围并陈述观点； 诊断型分析即通过什么，病根在哪里； 预测型分析即结合对病人的病情会怎样发展； 指导型分析即最后开出针对性的治疗方案。



图片来源：<https://bbs.fanruan.com/thread-135963-1-1.html>

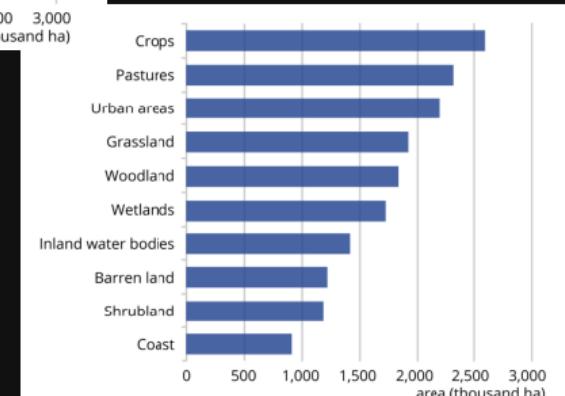
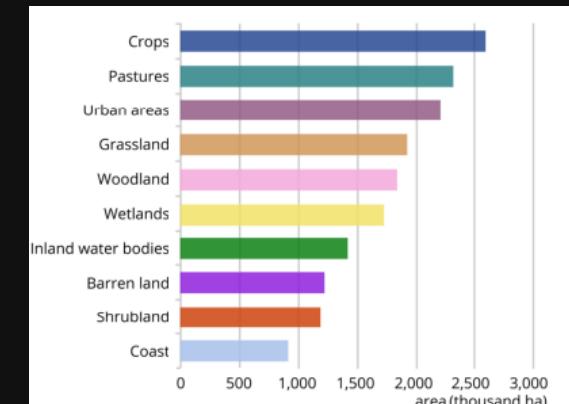


# 数据可视化：思维方式

**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	<u>特效/颜色字体是否有明显「噪声」</u> <u>辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」</u>	艺术
A	精准表达：提升数据表达的准确度	<b>解释性数据分析</b> 是否过低 数据密度是否合适 数据显示效果是否准确	
D	画龙点睛：突出洞察信息的标识	是否有突出洞察的标识	

确保图表中的颜色用于传递特定的信息，如果不是或有其他方式能够更有效地传递该信息，那就避免使用颜色。



- 任意或无意义地使用颜色，极大程度上会对用户造成噪音干扰。如下图所示：加入太多的颜色，其实并没有额外的意义，所以保持一种颜色即可，



# 数据可视化：思维方式

**GLAD** 原则：

中文含义		探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	艺术
	精准表达：提升数 字+文本+图标+色块	解释性数据分析 是否过低	



当对象与信息标注分隔较远的话，用户需要花较多的时间让视线来回切换，不利于信息快速获取。

- 例如，图例和数据序列相距较远的话，用户在解读数据时需要辛苦地在图例和数据之间来回切换，如下图：解决方案之一就是直接标记各种数据，利用格式塔临近原则，减少受众在图例和数据之间的来回切换。

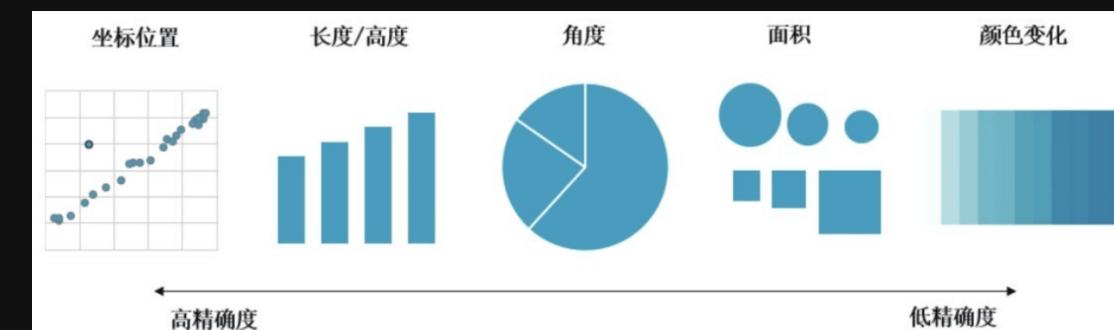


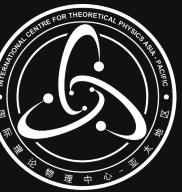
# 数据可视化：思维方式

**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	艺术
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	技术
D	画龙点睛：突出洞察信息的标识	解释性数据分析	标识

选择图形元素要准确。





# 数据可视化：思维方式

**GLAD 原则：**



数据的密度是指单位面积图表中所包含的数据量。图表所能承载的数据量是有限的：

- 过低会造成图表的丰富度不够，没有回答读者的问题。
- 过高会导致负载过重，读者无法理解图表要传达的信息。

一张图表的数据密度 = 类别的数量 + 度量指标的数量

L	辅助信息(文字/标签/图例/标尺等)是
A	精准表达：提升数据表达的准确度 <u>图形元素的精确度是否过低</u> <u>数据密度是否合适</u> 数据显示效果是否准确
D	画龙点睛：突出洞察信息的标识 <b>解释性数据分析</b>

- 很多人为了压缩展示的空间和精简图表，会使用组合图把很多信息拼到一起，但是这样的话数据过于集中，会导致读者无法一下子得到信息要点。如上图：
  - 类别：流动量、离职人数、离职率、流动率
  - 度量：人数、百分比
- 这样这张图表的数据密度 =  $4 + 2 = 6$  这个时候，比如我们看离职率或者流失率就看不出趋势。对于这个图表我们可以做降维处理，拆分为两个组合图，每个组合图的数据密度降低到 3，如右侧两个图。





# 数据可视化：思维方式

**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	艺术
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	技术
D	画龙点睛：突出洞察信息的标识	解释性数据分析	标识

选择合适的图表类型和把控数据密度属于粗调，在最后的展示前还需要对显示效果的细节做精调，否则也可能导致数据与事实的偏离。

- 调节可视化准确度的工作就像拧螺丝一样，选用合适口径的螺丝，把螺丝放到孔中，先用手快速旋转到一定紧度，再使用螺丝刀一步步拧紧固定，粗调和精调并用。
- 选择合适的图表类型和把控数据密度属于粗调，在最后的展示前还需要对显示效果的细节做精调，否则也可能导致数据与事实的偏离。
- 这里折线图很难看出趋势的变化，因为离职率最大的月份只有 3.9%，而纵坐标轴的最大刻度为 50%，这样趋势就被压平了，建议将纵坐标轴的最大刻度调整小一些。





# 数据可视化：思维方式

**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	艺术
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	技术
D	画龙点睛：突出洞察信息的标识	<u>是否有突出洞察的标识</u>	艺术

**解释性数据分析**

- 这些「前注意属性」主要包括以下几种：我们需要把通过数据比较得到的差异部分、体现洞察信息的内容利用明显不同的颜色、形状、文字标注等手段进行区别，让读者的视线聚焦到那里去。

现代心理学把颜色、形状等能快速引起心理反应的信号统称为「**前注意属性**」，它们在我们的潜意识中活动，只需要0.25秒就可以作出识别。

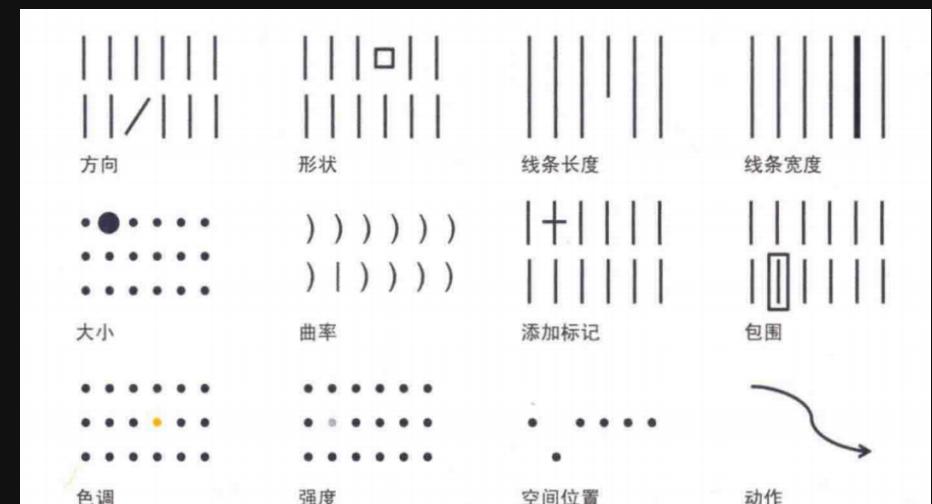


图 4-4 前注意属性

来源：引自 2004 年 Stephen Few 的 *Show Me the Numbers*。



# 数据可视化：思维方式

**GLAD** 原则：

		中文含义	探索性数据分析	思考问题																																																																									
G	图表的灵魂：发现好数据和好洞察	数据是否恰当			<p>打造视觉反差：我们可以利用颜色、形状、线的粗细用来打造视觉差异。</p> <p>举几个例子（你第一眼被什么吸引了？）：</p>																																																																								
		洞察在哪个层次																																																																											
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」			<table border="1"> <caption>2021年销售目标与毛利目标完成情况</caption> <thead> <tr> <th>序号</th> <th>品类</th> <th>年度销售目标</th> <th>累计销售</th> <th>销售达成率(%)</th> <th>年度毛利目标</th> <th>累计毛利</th> <th>毛利达成率(%)</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>海鲜</td> <td>50,000,000</td> <td>32,500,000</td> <td>65.00%</td> <td>40,000,000</td> <td>24,375,000</td> <td>60.94%</td> </tr> <tr> <td>2</td> <td>肉类</td> <td>50,000,000</td> <td>27,000,000</td> <td>54.00%</td> <td>37,500,000</td> <td>20,250,000</td> <td>54.00%</td> </tr> <tr> <td>3</td> <td>饮品</td> <td>50,000,000</td> <td>24,000,000</td> <td>48.00%</td> <td>32,500,000</td> <td>18,000,000</td> <td>55.38%</td> </tr> <tr> <td>4</td> <td>蔬菜</td> <td>2,000,000</td> <td>780,000</td> <td>39.00%</td> <td>1,300,000</td> <td>585,000</td> <td>45.00%</td> </tr> <tr> <td>5</td> <td>日用品</td> <td>5,000,000</td> <td>1,650,000</td> <td>33.00%</td> <td>3,000,000</td> <td>1,237,500</td> <td>41.25%</td> </tr> <tr> <td>6</td> <td>办公用品</td> <td>1,000,000</td> <td>320,000</td> <td>32.00%</td> <td>500,000</td> <td>240,000</td> <td>48.00%</td> </tr> <tr> <td>7</td> <td>食品</td> <td>2,500,000</td> <td>525,000</td> <td>21.00%</td> <td>15,000,000</td> <td>393,750</td> <td>2.63%</td> </tr> <tr> <td colspan="2" rowspan="2">合计</td><td>160,500,000</td><td>86,775,000</td><td>41.71%</td><td>129,800,000</td><td>65,081,250</td><td>43.89%</td></tr> </tbody> </table>	序号	品类	年度销售目标	累计销售	销售达成率(%)	年度毛利目标	累计毛利	毛利达成率(%)	1	海鲜	50,000,000	32,500,000	65.00%	40,000,000	24,375,000	60.94%	2	肉类	50,000,000	27,000,000	54.00%	37,500,000	20,250,000	54.00%	3	饮品	50,000,000	24,000,000	48.00%	32,500,000	18,000,000	55.38%	4	蔬菜	2,000,000	780,000	39.00%	1,300,000	585,000	45.00%	5	日用品	5,000,000	1,650,000	33.00%	3,000,000	1,237,500	41.25%	6	办公用品	1,000,000	320,000	32.00%	500,000	240,000	48.00%	7	食品	2,500,000	525,000	21.00%	15,000,000	393,750	2.63%	合计		160,500,000	86,775,000	41.71%	129,800,000	65,081,250	43.89%
序号	品类	年度销售目标	累计销售	销售达成率(%)	年度毛利目标	累计毛利	毛利达成率(%)																																																																						
1	海鲜	50,000,000	32,500,000	65.00%	40,000,000	24,375,000	60.94%																																																																						
2	肉类	50,000,000	27,000,000	54.00%	37,500,000	20,250,000	54.00%																																																																						
3	饮品	50,000,000	24,000,000	48.00%	32,500,000	18,000,000	55.38%																																																																						
4	蔬菜	2,000,000	780,000	39.00%	1,300,000	585,000	45.00%																																																																						
5	日用品	5,000,000	1,650,000	33.00%	3,000,000	1,237,500	41.25%																																																																						
6	办公用品	1,000,000	320,000	32.00%	500,000	240,000	48.00%																																																																						
7	食品	2,500,000	525,000	21.00%	15,000,000	393,750	2.63%																																																																						
合计		160,500,000	86,775,000	41.71%	129,800,000	65,081,250	43.89%																																																																						
		辅助信息(文字/标签/图例/标尺等)是否有																																																																											
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低			<table border="1"> <caption>车订单各层级城市明细下钻和环比分析-点击日期和城市分级下钻</caption> <thead> <tr> <th rowspan="2">城市分级</th> <th colspan="2">一线城市</th> <th colspan="2">新一线城市</th> <th colspan="2">二线城市</th> <th colspan="2">三线城市</th> </tr> <tr> <th>日期</th> <th>成交订单量</th> <th>成交订单量环比增长率(%)</th> <th>成交订单量</th> <th>成交订单量环比增长率(%)</th> <th>成交订单量</th> <th>成交订单量环比增长率(%)</th> <th>成交订单量</th> <th>成交订单量环比增长率(%)</th> </tr> </thead> <tbody> <tr> <td>2015</td> <td>506</td> <td></td> <td>548</td> <td></td> <td>369</td> <td></td> <td>192</td> <td></td> </tr> <tr> <td>2016</td> <td>853</td> <td>68.58% ↑</td> <td>1,124</td> <td>105.11% ↑</td> <td>650</td> <td>76.15% ↑</td> <td>366</td> <td>90.63% ↑</td> </tr> <tr> <td>2017</td> <td>1,593</td> <td>86.75% ↑</td> <td>2,060</td> <td>83.27% ↑</td> <td>1,229</td> <td>89.08% ↑</td> <td>558</td> <td>52.46% ↑</td> </tr> <tr> <td>2018</td> <td>1,845</td> <td>15.82% ↑</td> <td>2,334</td> <td>13.30% ↑</td> <td>1,512</td> <td>23.03% ↑</td> <td>869</td> <td>55.73% ↑</td> </tr> <tr> <td>2019</td> <td>1,097</td> <td>-40.54% ↓</td> <td>1,553</td> <td>-33.46% ↓</td> <td>1,086</td> <td>-28.17% ↓</td> <td>599</td> <td>-31.07% ↓</td> </tr> <tr> <td colspan="2" rowspan="3">合计</td><td>5,894</td><td>7,619</td><td>4,846</td><td>2,584</td><td></td><td></td><td></td></tr> </tbody> </table>	城市分级	一线城市		新一线城市		二线城市		三线城市		日期	成交订单量	成交订单量环比增长率(%)	成交订单量	成交订单量环比增长率(%)	成交订单量	成交订单量环比增长率(%)	成交订单量	成交订单量环比增长率(%)	2015	506		548		369		192		2016	853	68.58% ↑	1,124	105.11% ↑	650	76.15% ↑	366	90.63% ↑	2017	1,593	86.75% ↑	2,060	83.27% ↑	1,229	89.08% ↑	558	52.46% ↑	2018	1,845	15.82% ↑	2,334	13.30% ↑	1,512	23.03% ↑	869	55.73% ↑	2019	1,097	-40.54% ↓	1,553	-33.46% ↓	1,086	-28.17% ↓	599	-31.07% ↓	合计		5,894	7,619	4,846	2,584			
城市分级	一线城市		新一线城市		二线城市		三线城市																																																																						
	日期	成交订单量	成交订单量环比增长率(%)	成交订单量	成交订单量环比增长率(%)	成交订单量	成交订单量环比增长率(%)	成交订单量	成交订单量环比增长率(%)																																																																				
2015	506		548		369		192																																																																						
2016	853	68.58% ↑	1,124	105.11% ↑	650	76.15% ↑	366	90.63% ↑																																																																					
2017	1,593	86.75% ↑	2,060	83.27% ↑	1,229	89.08% ↑	558	52.46% ↑																																																																					
2018	1,845	15.82% ↑	2,334	13.30% ↑	1,512	23.03% ↑	869	55.73% ↑																																																																					
2019	1,097	-40.54% ↓	1,553	-33.46% ↓	1,086	-28.17% ↓	599	-31.07% ↓																																																																					
合计		5,894	7,619	4,846	2,584																																																																								
		数据密度是否合适																																																																											
		数据显示效果是否准确																																																																											
D	画龙点睛：突出洞察信息的标识	是否有突出洞察的标识																																																																											

## 解释性数据分析

- 这些「前注意属性」主要包括以下几种：我们需要把通过数据比较得到的差异部分、体现洞察信息的内容利用明显不同的颜色、形状、文字标注等手段进行区别，让读者的视线聚焦到那里去。



# 数据可视化：思维方式

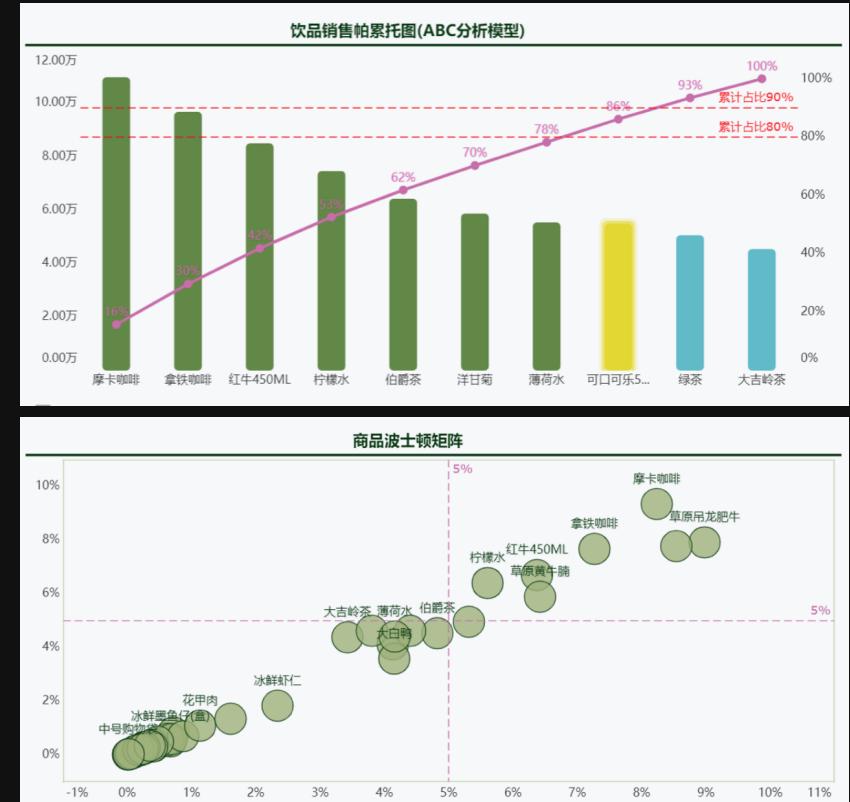
**GLAD** 原则：

	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	艺术
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	技术
D	画龙点睛：突出洞察信息的标识	<u>是否有突出洞察的标识</u>	艺术

**解释性数据分析**

- 这些「前注意属性」主要包括以下几种：我们需要把通过数据比较得到的差异部分、体现洞察信息的内容利用明显不同的颜色、形状、文字标注等手段进行区别，让读者的视线聚焦到那里去。

还可以绘制神奇的线：水平线、趋势线、划分区间...  
举几个例子（你第一眼被什么吸睛了？）：





# 数据可视化：思维方式

**GLAD** 原则：

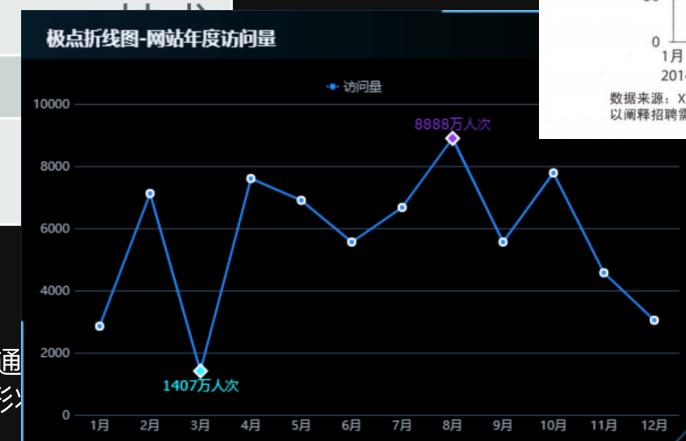
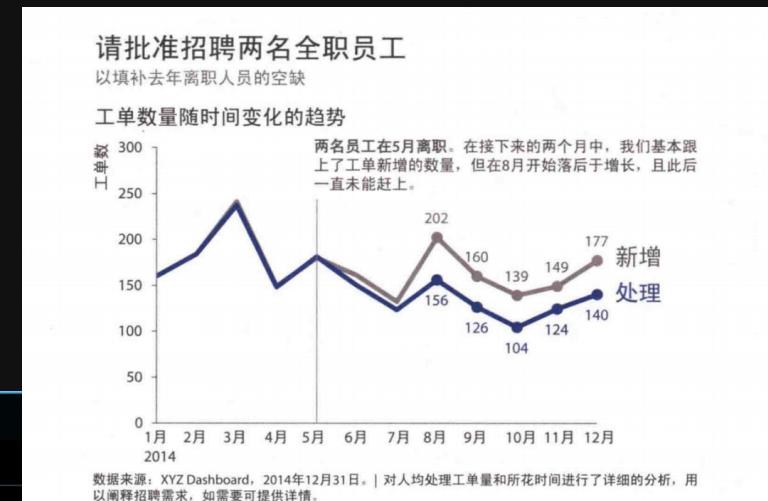
	中文含义	探索性数据分析	思考问题
G	图表的灵魂：发现好数据和好洞察	数据是否恰当 洞察在哪个层次	技术
L	降噪：简约至上	特效/颜色字体是否有明显「噪声」 辅助信息(文字/标签/图例/标尺等)是否有明显「噪声」	艺术
A	精准表达：提升数据表达的准确度	图形元素的精确度是否过低 数据密度是否合适 数据显示效果是否准确	
D	画龙点睛：突出洞察信息的标识	是否有突出洞察的标识	

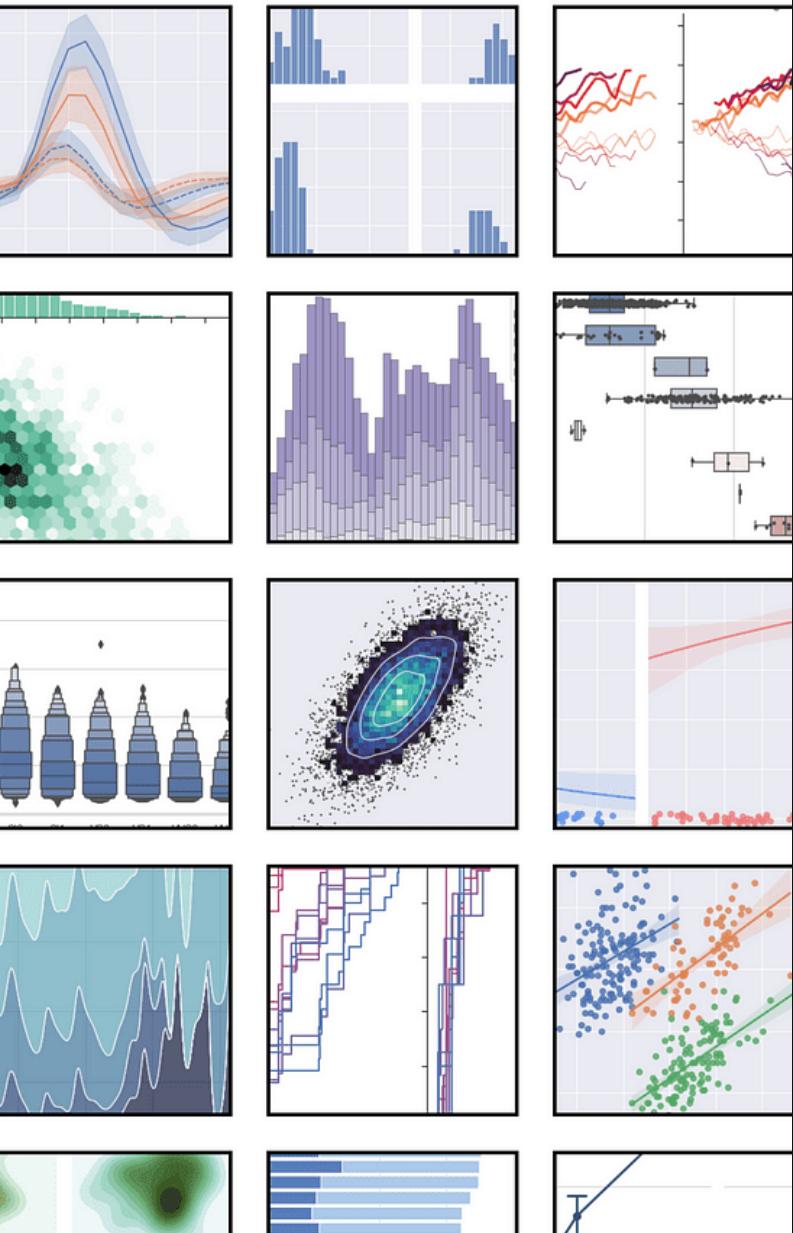
## 解释性数据分析

- 这些「**前注意属性**」主要包括以下几种：我们需要把通异部分、体现洞察信息的内容利用明显不同的颜色、形进行区别，让读者的视线聚焦到那里去。

添加文字：文字在数据沟通中能起到以下作用：标签、简介、解释、强调、突出、推荐和讲故事。

举几个例子（你第一眼被什么吸睛了？）：





# 数据分析可视化之 Seaborn

- 风格管理
- 颜色管理
- 皮尔森系数



seaborn



“Talk is cheap. Show me the code.”

Linus Torvalds

Repo of the course: <https://github.com/iophysresearch/GWData-Bootcamp>

# Homework

1. 航班乘客变化分析 (2个题)
2. 鸢尾花花型尺寸分析 (3个题)
3. 餐厅小费情况分析 (7个题)
4. 泰坦尼克号海难幸存状况分析 (8个题)

- 基础作业:
  - 数据可视化作业题目位于 [2023/python/homework-matplotlib\\_seaborn.ipynb](#), 用 matplotlib 或 seaborn 完成题目后, 把该 notebook 提交到学员自己的作业目录中, 最后 PR 即可。
- 拓展作业:
  - 分别用 matplotlib 和 seaborn 完成基础作业。

