第4章 数据可视化

1.本章定位与内容简介

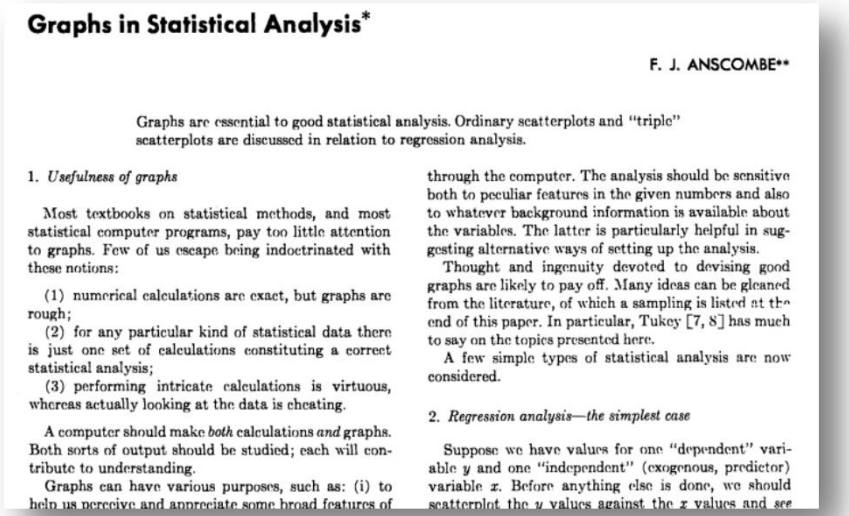
2.本章学习提示及要求

了解	理解	掌握	熟练掌握
• 数据可视化与数据科学的区别与联系 • 大数据环境下数据可视化的发展趋势	• 数据可视化的基本原则 • 可视分析学及其核心模型	• 数据类型的划分方法及视觉编码方法的选择 • 常用统计图表的绘制方法	• 基于Python的数据可视化

3. 数据可视化在数据科学中的重要地位

Anscombe的四组数据（Anscombe’s Quartet）

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	25
4.0	4.26	4.0	3.10	4.0	39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
0	68	0	4.74	0	73	8.0	6.89



Graphs in Statistical Analysis*

F. J. ANSCOMBE**

Graphs are essential to good statistical analysis. Ordinary scatterplots and "triple" scatterplots are discussed in relation to regression analysis.

1. *Usefulness of graphs*

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

(1) numerical calculations are exact, but graphs are rough;
(2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
(3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

Graphs can have various purposes, such as: (i) to help us perceive and appreciate some broad features of

through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

Thought and ingenuity devoted to devising good graphs are likely to pay off. Many ideas can be gleaned from the literature, of which a sampling is listed at the end of this paper. In particular, Tukey [7, 8] has much to say on the topics presented here.

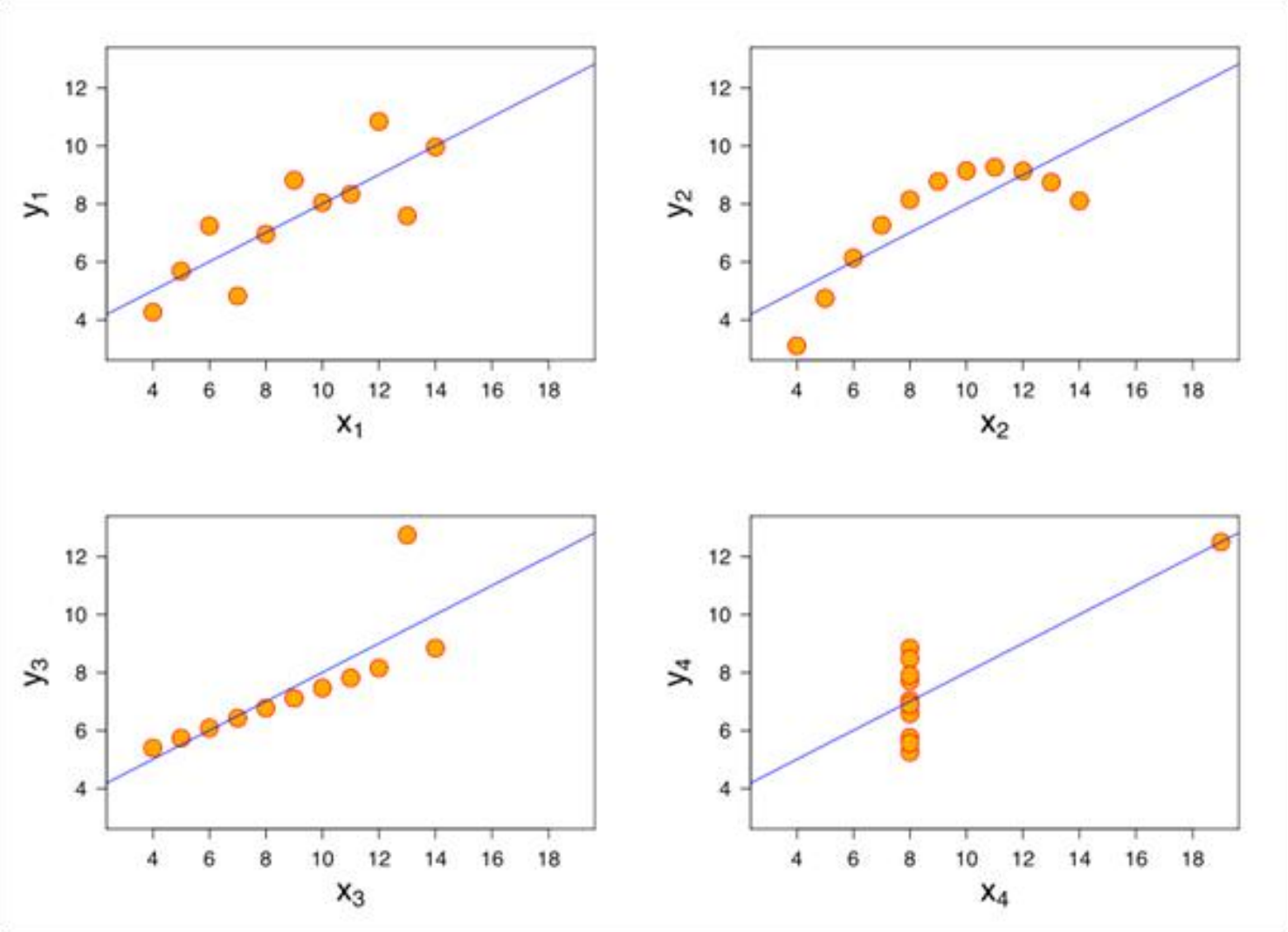
A few simple types of statistical analysis are now considered.

2. *Regression analysis—the simplest case*

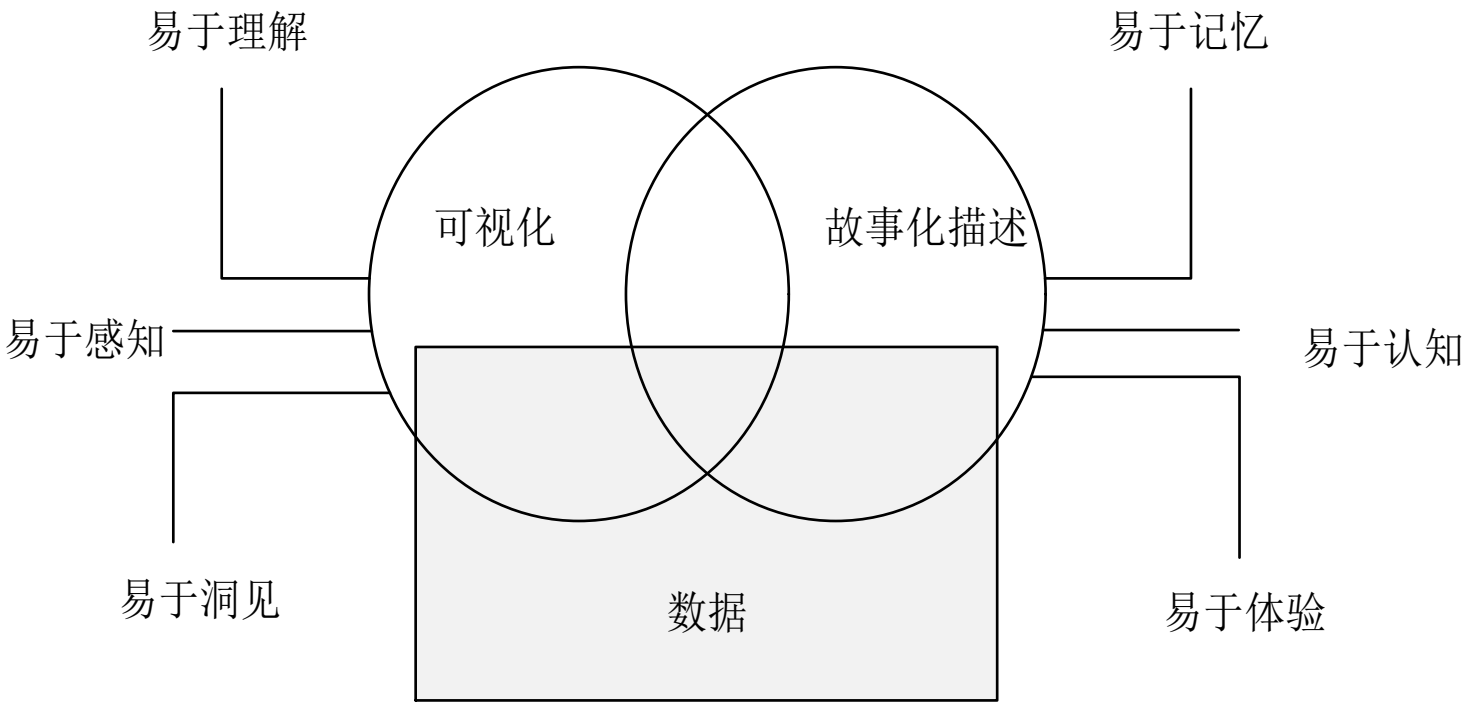
Suppose we have values for one "dependent" variable y and one "independent" (exogenous, predictor) variable x. Before anything else is done, we should scatterplot the y values against the x values and *see*

Anscombe F J. Graphs in statistical analysis[J]. The American Statistician, 1973, 27(1): 17-21.

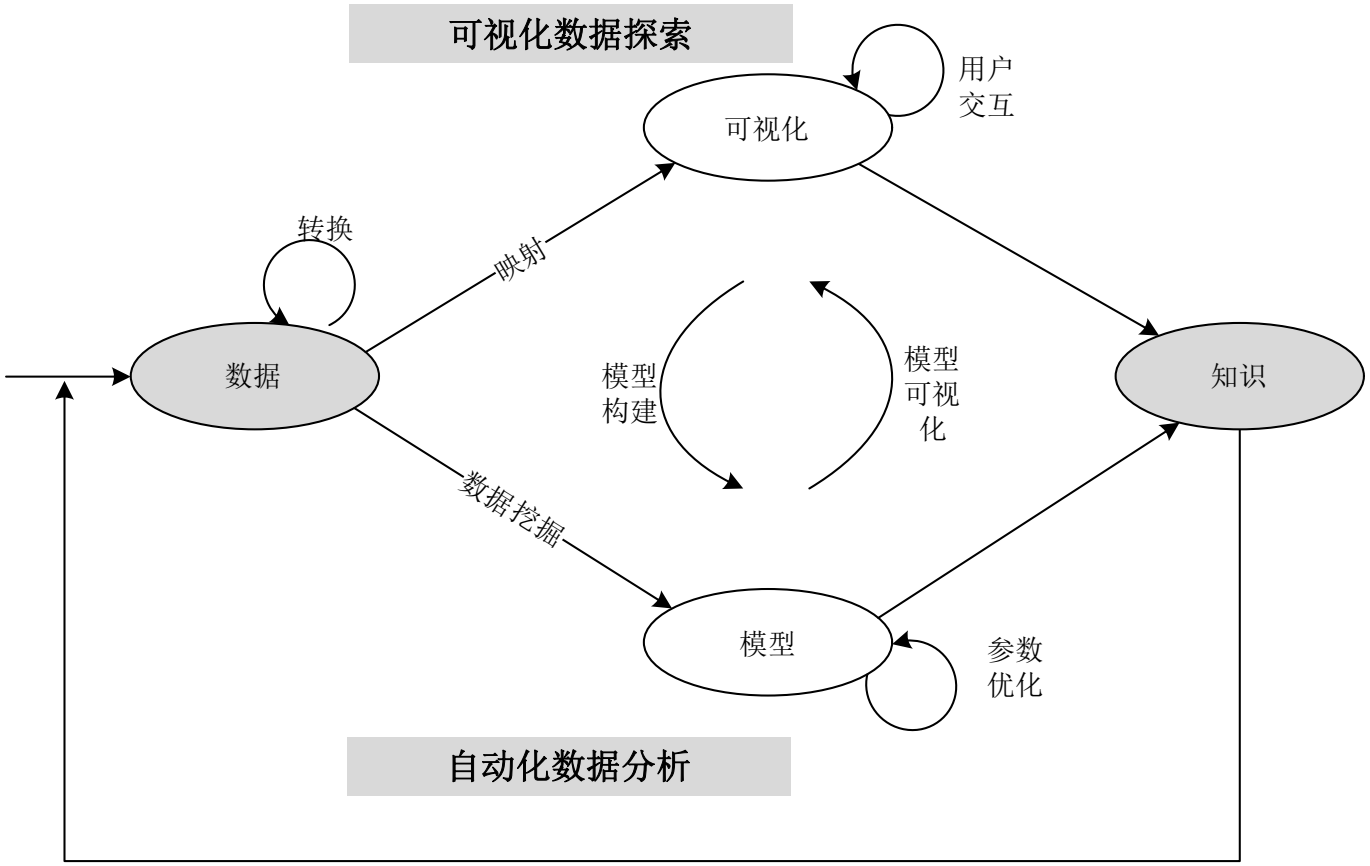
Anscombe's Quartet



数据的可视化与故事化



4.可视分析学



强调数据到知识的转换过程

强调可视化分析与自动化建模之间的相互作用

强调数据映射和数据挖掘的重要性

强调数据预处理工作的必要性

强调人机交互的重要性

5.视觉假象

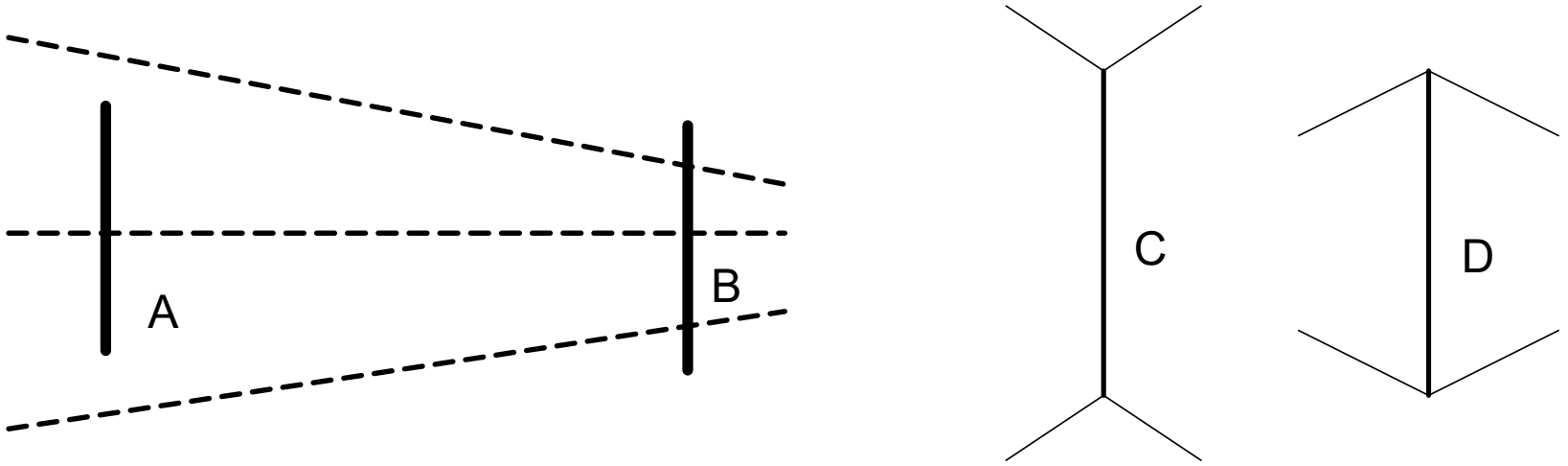
含义

- 是指给目标用户产生的错误或不准确的**视觉感知**，而这种感知与数据可视化者的意图或数据本身的真实情况不一致。

原因

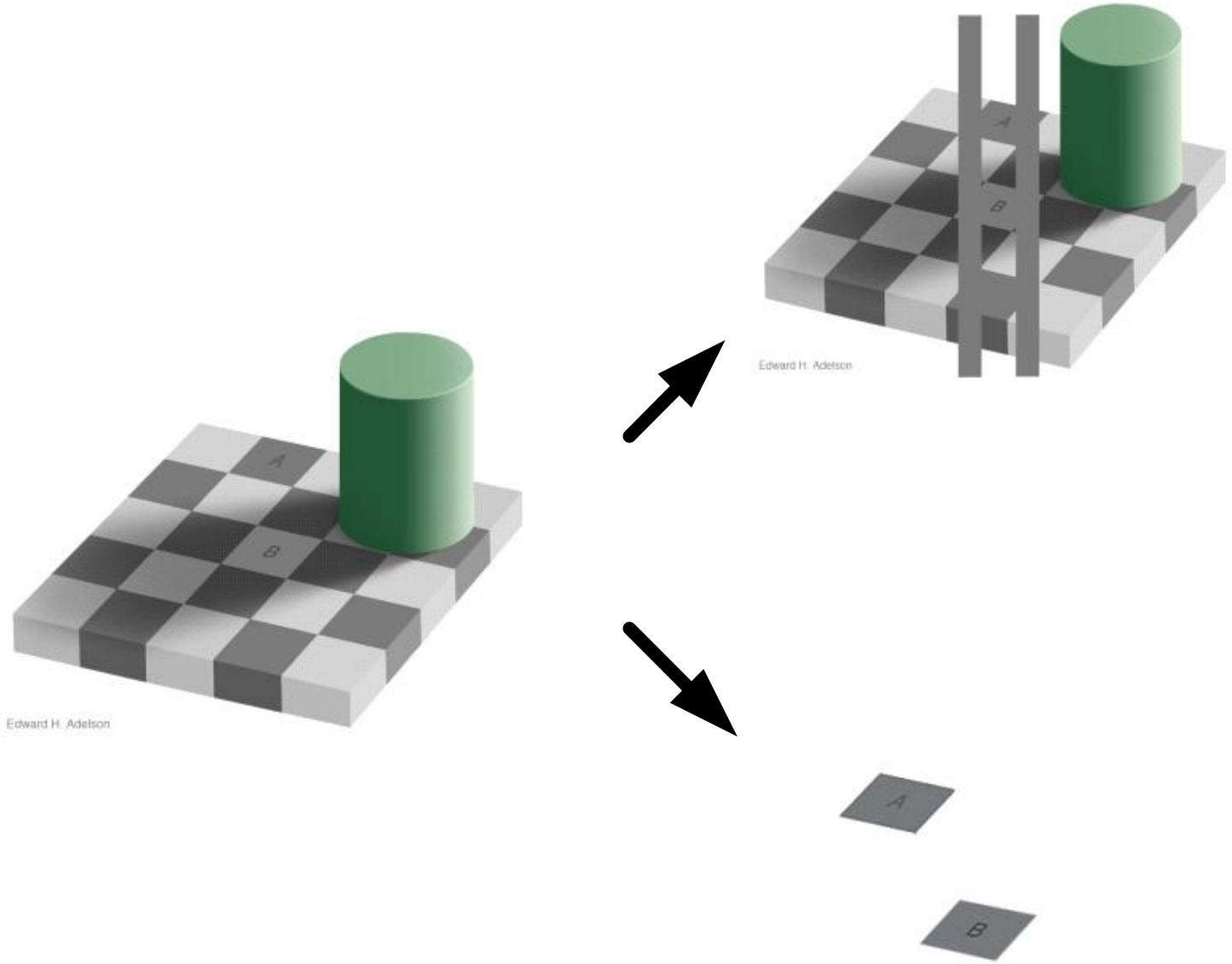
- 可视化视图所处的上下文（周边环境）可能导致视觉假象。
- 人们对亮度和颜色的相对判断容易造成视觉假象。
- 目标用户的经历与经验可能导致视觉假象。

（1）可视化视图所处的上下文（周边环境）可能导致视觉假象。



上下文可能导致视觉假象的示例

（2）人们对亮度和颜色的相对判断容易造成视觉假象。



色块A比色块B更亮？

（3）目标用户的经历与经验可能导致视觉假象

目标用户的经历与经验可能导致视觉假象

6.如何继续学习本章知识

面向特定专业领域的数据可视化方法

- 1931 年，机械制图员Henry Beck
- 借鉴电路图的制图方法设计出伦敦地铁线路图

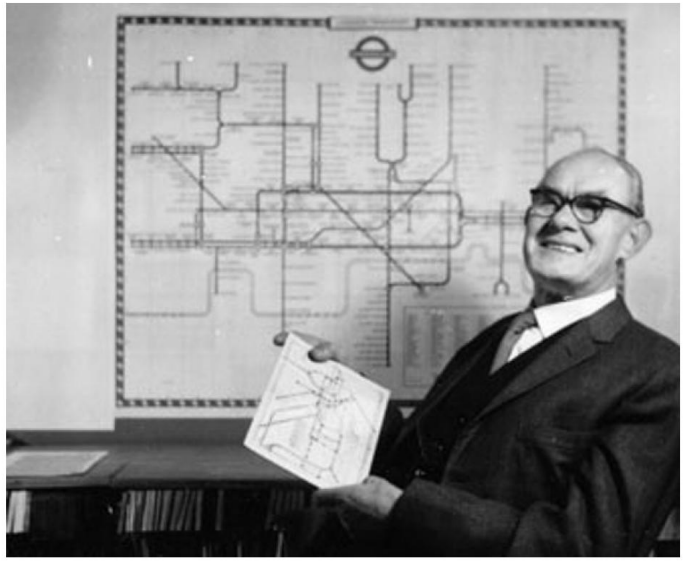


图 4-19　亨利·贝克（来源：伦敦交通博物馆）

数据可视化与其他数据呈现方式，尤其是数据故事化等有效结合

- 可理解性
- 可记忆性
- 可体验性

有效利用数据可视化方法

- 数据可视化也不是万能的
- 更不能“为了可视化而可视化”

重视数据可视化的动手实践

- Matplotlib、seaborn、Bokeh、Basemap、Plotly、NetworkX

小结

1.本章定位与内容简介

2.本章学习提示及要求

3.数据可视化在数据科学中的重要地位

4.可视分析学

5.视觉假象

6.如何继续学习本章知识