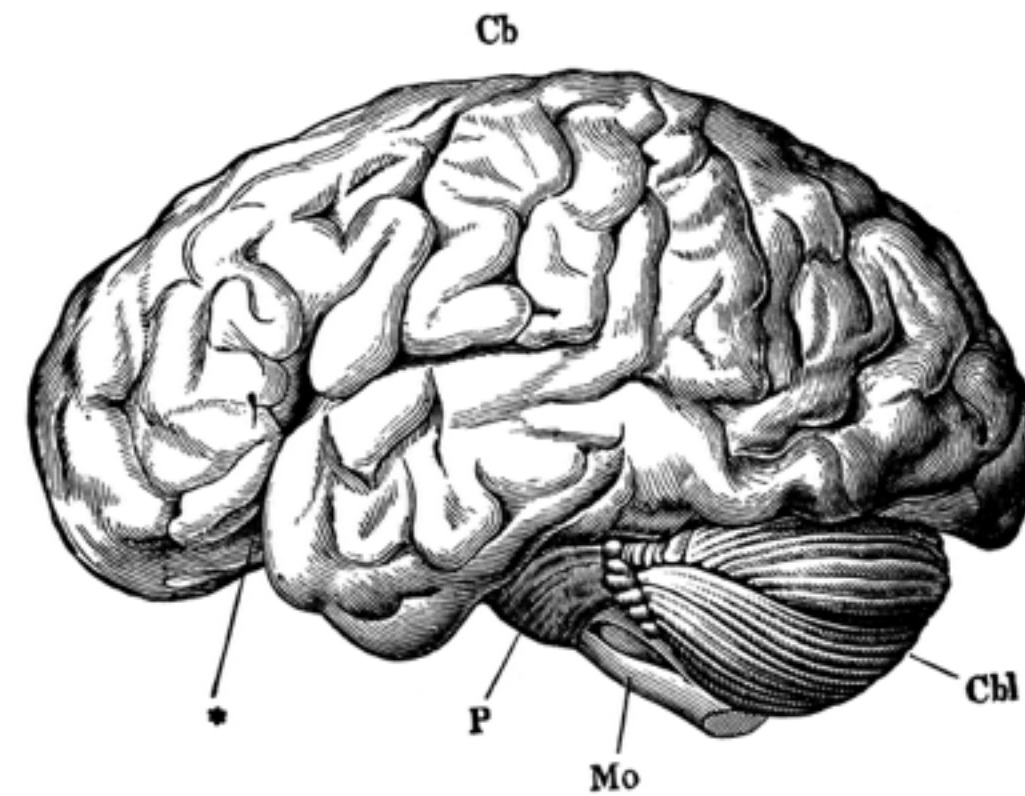


Data Science

Suggestions for statisticians



Garrett Grolemond

Data Scientist, Educator

March 2016



1. For Statisticians

2. For Universities

1. For Statisticians

2. For Universities



SCIENCE CENTER



garrettgrolemond — R — 81×29

```
> lm(weight ~ height, data = ?
```

Tips for Statisticians

Learn to:

- 1.** Use Databases (SQL)
- 2.** Wrangle Data
- 3.** Learn to program (R, Python, etc.)

Lifecycle of an Analysis Project

Clarify

Become familiar with the data,
template a solution

Develop

Create a working model

Productize

Automate and integrate

Publish

Socialize

Lifecycle of an Analysis Project

Subset

Extract data to explore, work with

Clarify

Become familiar with the data,
template a solution

Develop

Create a working model

Scale Up*

Generalize to entire data set

Productize

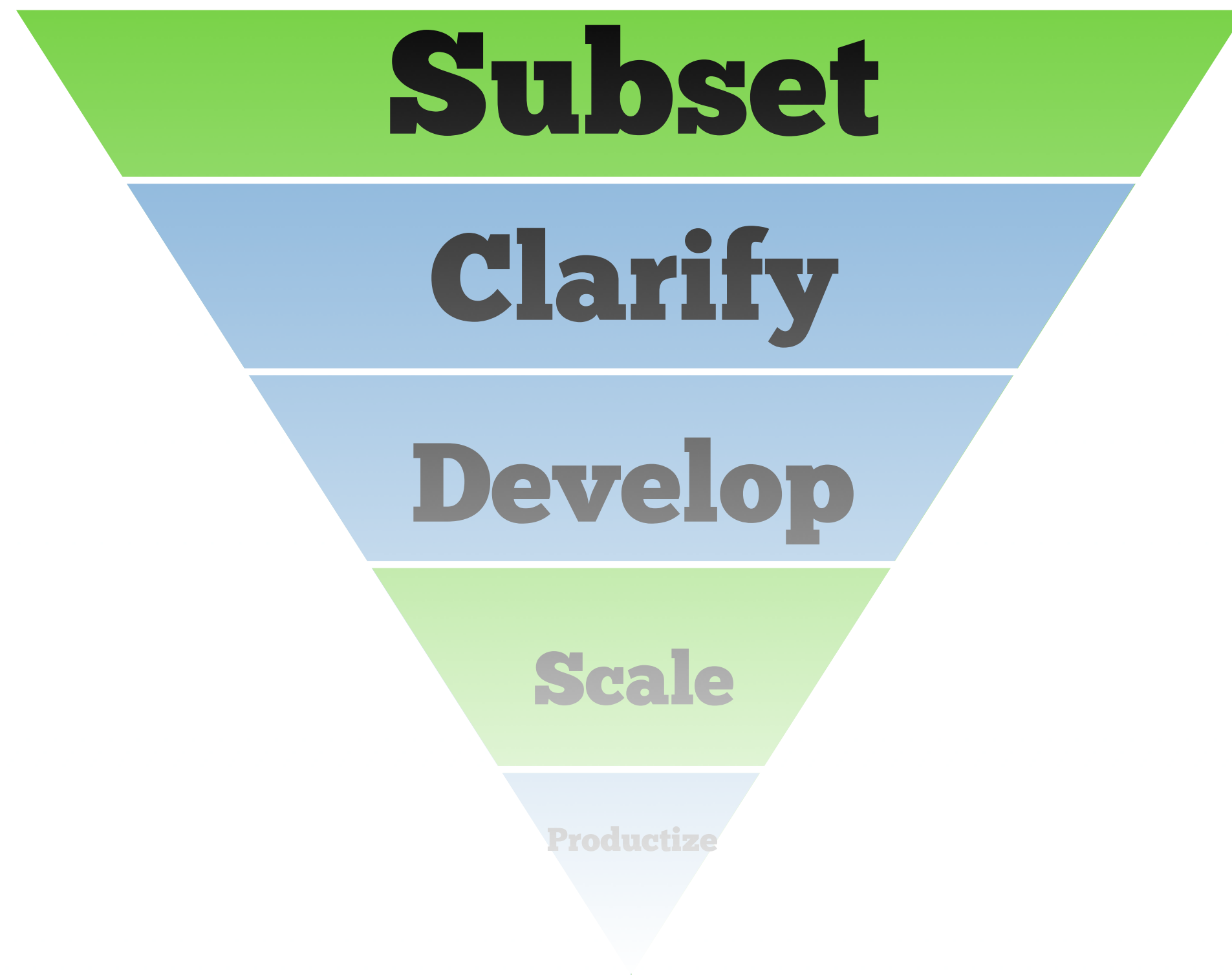
Automate and integrate

Publish

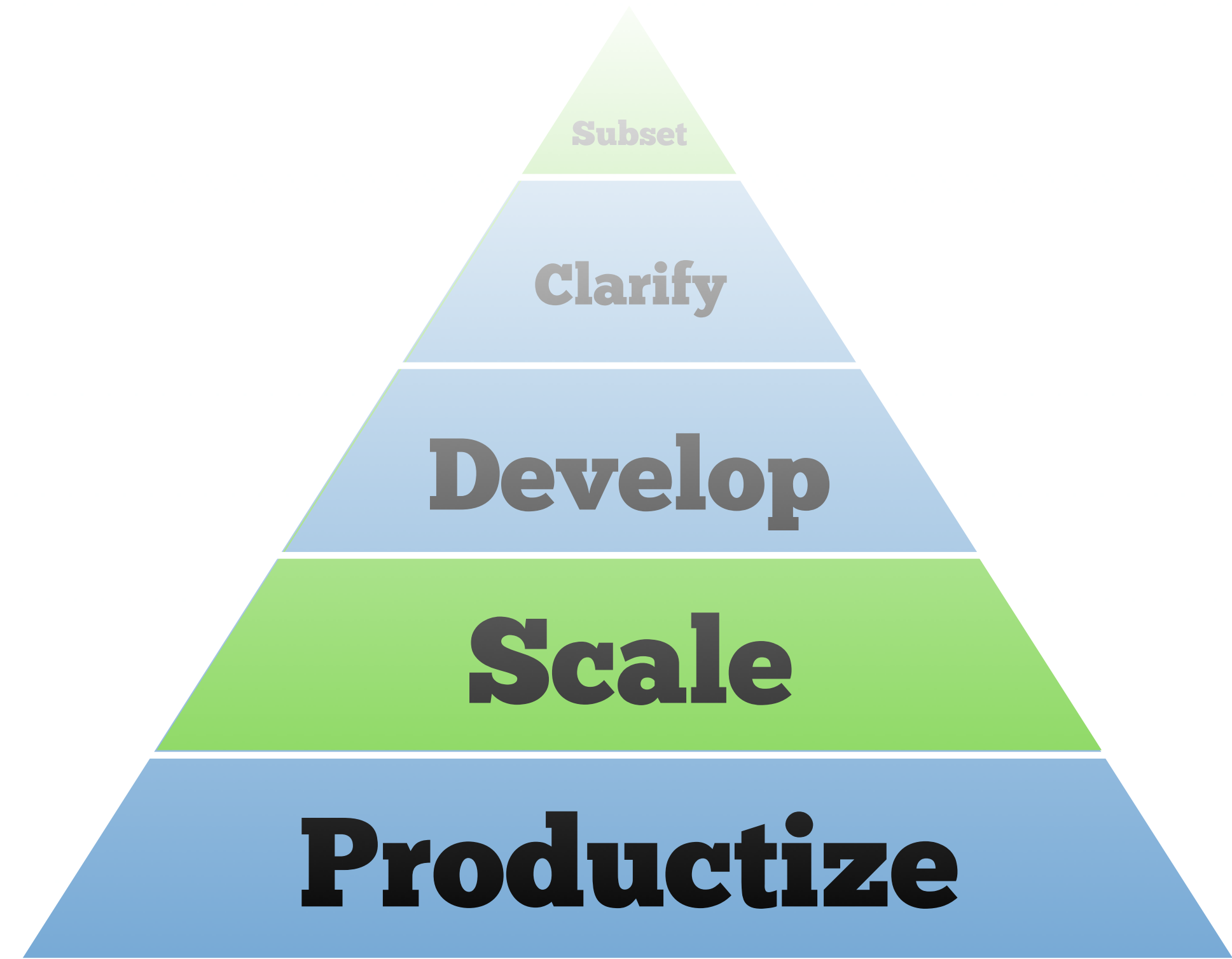
Socialize

* **sometimes**

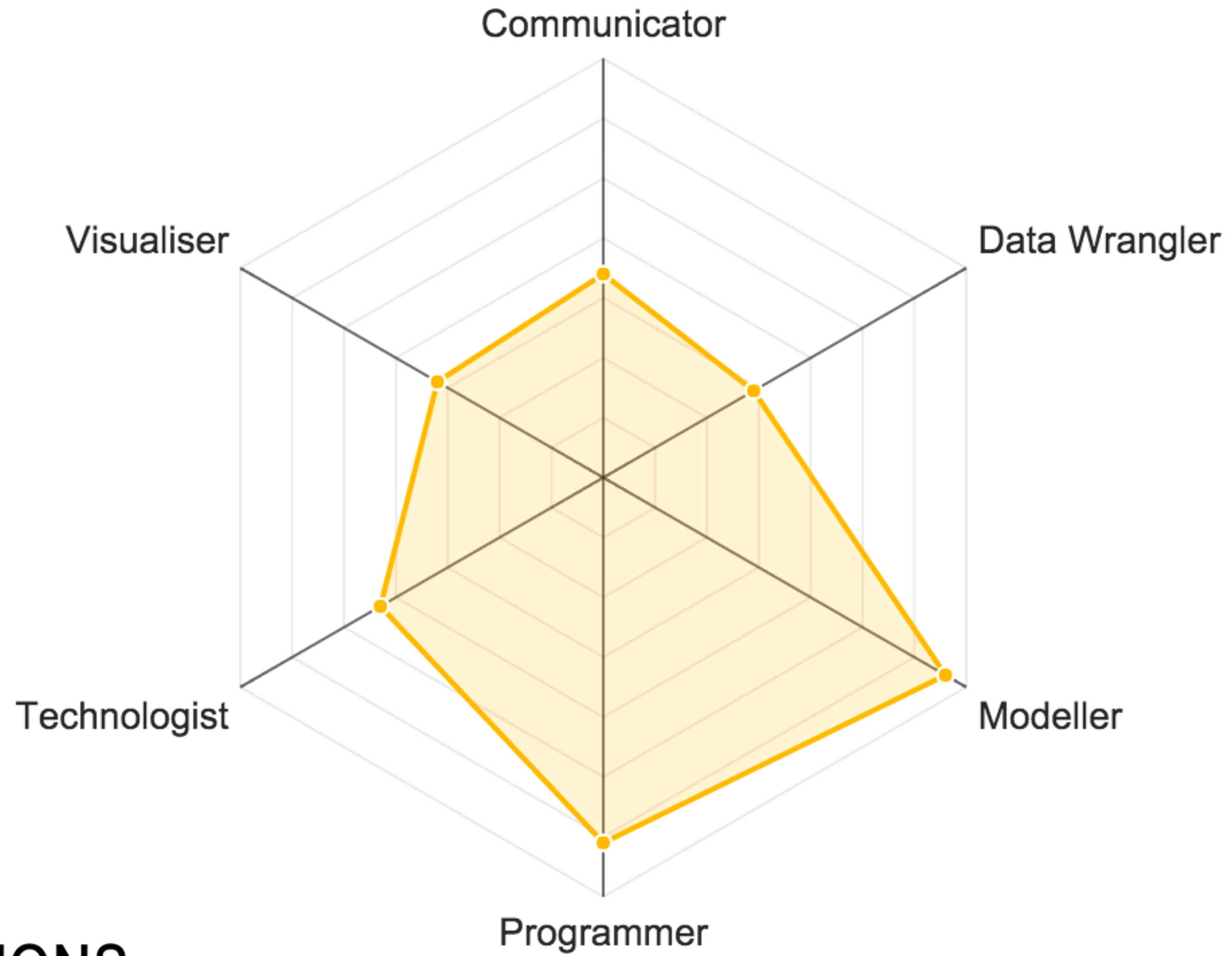
Data Roles

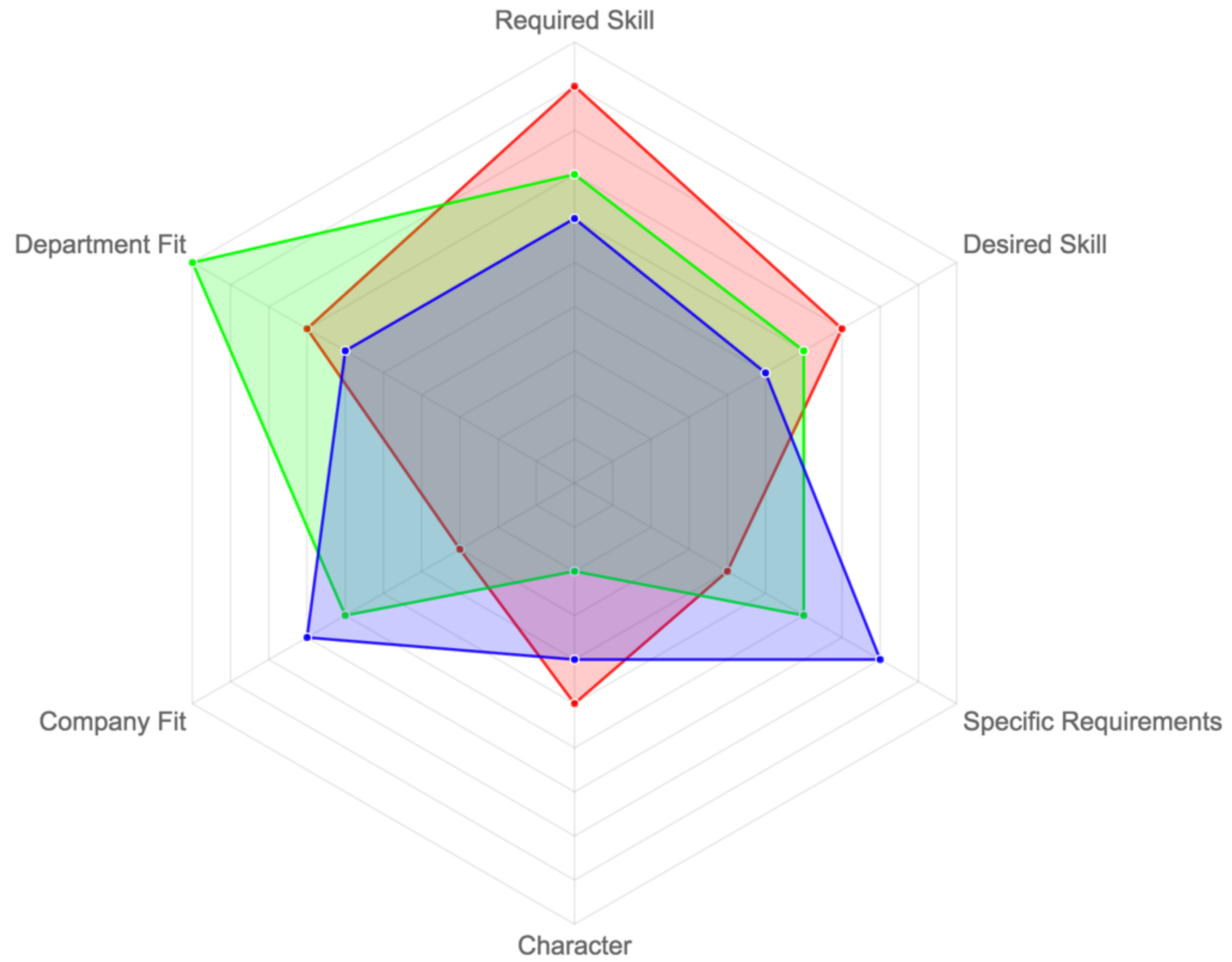


Analyst



IT/Manager/Engineer





1. For Statisticians

2. For Universities

1. Embrace Exploratory Data Analysis (EDA)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
data = pd.read_csv('data.csv')

# Check the shape of the data
print(data.shape)

# Check the data types
print(data.dtypes)

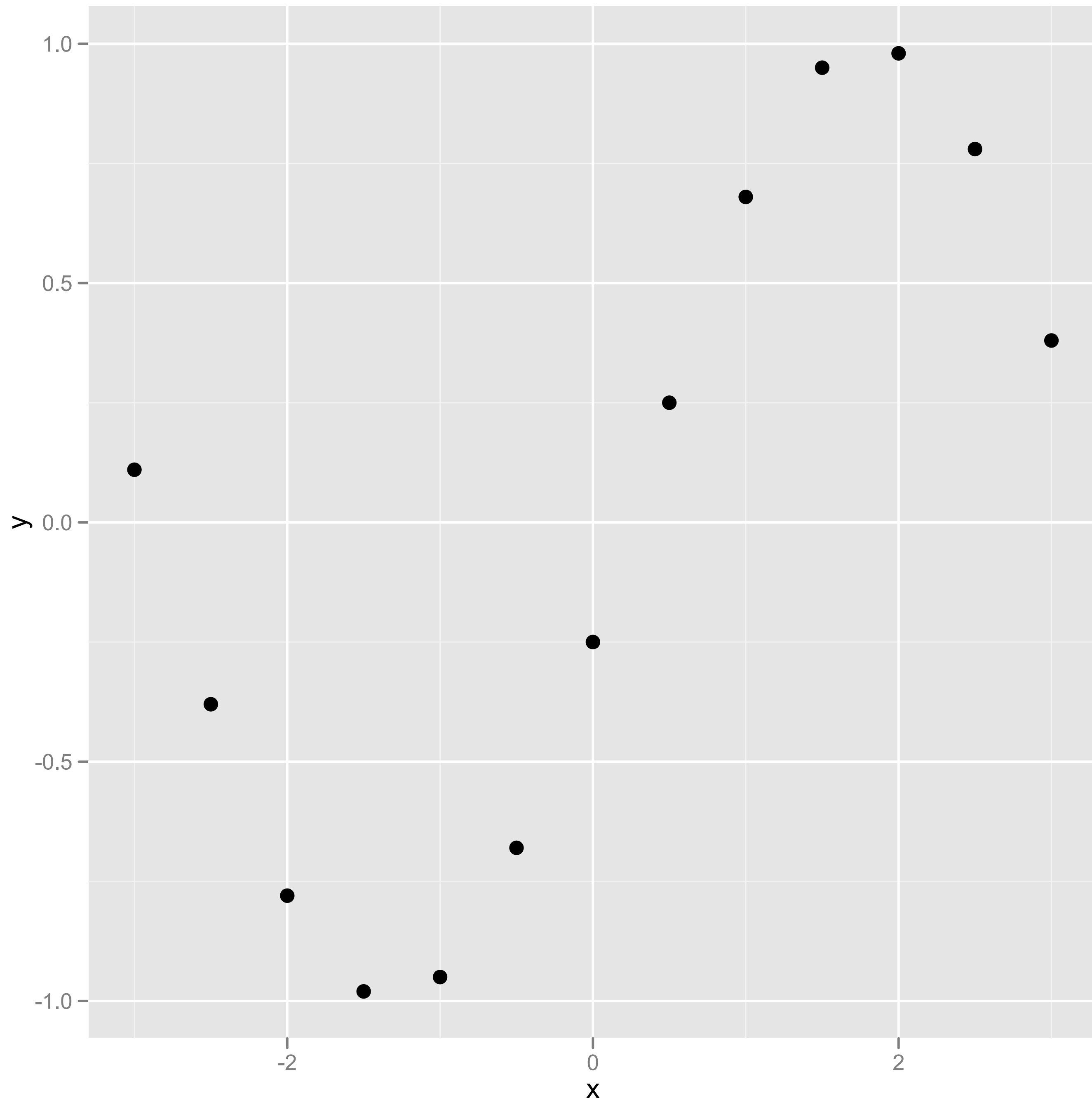
# Check for missing values
print(data.isnull().sum())

# Check the distribution of the data
print(data.describe())

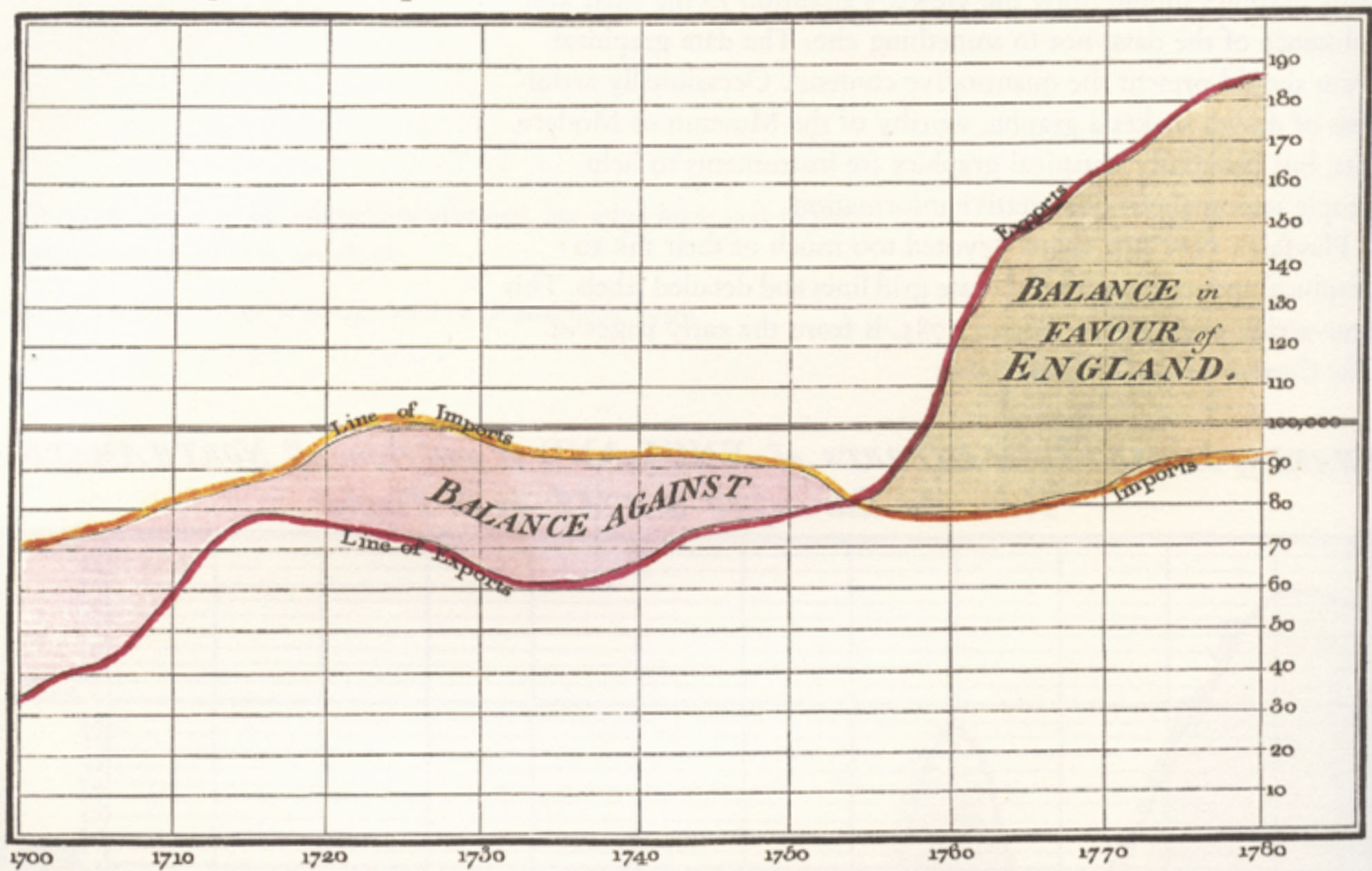
# Visualize the data
sns.pairplot(data)

# Summary statistics
print(data.mean())
print(data.std())
```

x	y
0	-0.25
-0.5	-0.68
-1	-0.95
-1.5	-0.98
-2.5	-0.38
0.5	0.25
2	0.98
1.5	0.95
3	0.38
1	0.68
2.5	0.78
-3	0.11
-2	-0.78



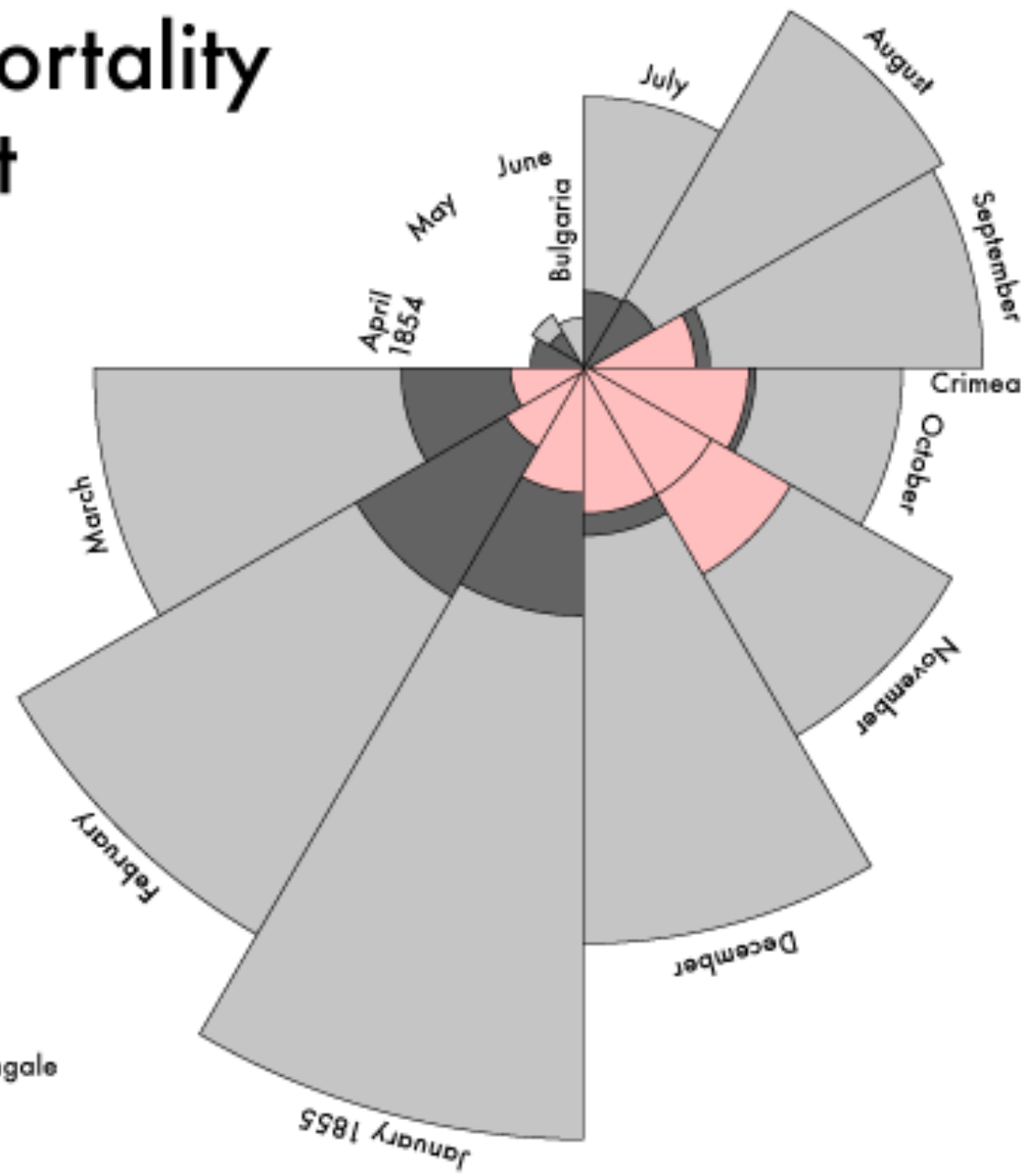
x	y
0	-0.25
-0.5	-0.68
-1	-0.95
-1.5	-0.98
-2.5	-0.38
0.5	0.25
2	0.98
1.5	0.95
3	0.38
1	0.68
2.5	0.78
-3	0.11
-2	-0.78



The Bottom line is divided into Years, the Right hand line into £10,000 each.
 Published as the Act directs, 1st May 1786, by W. Flaxman, No. 52, Strand, London.

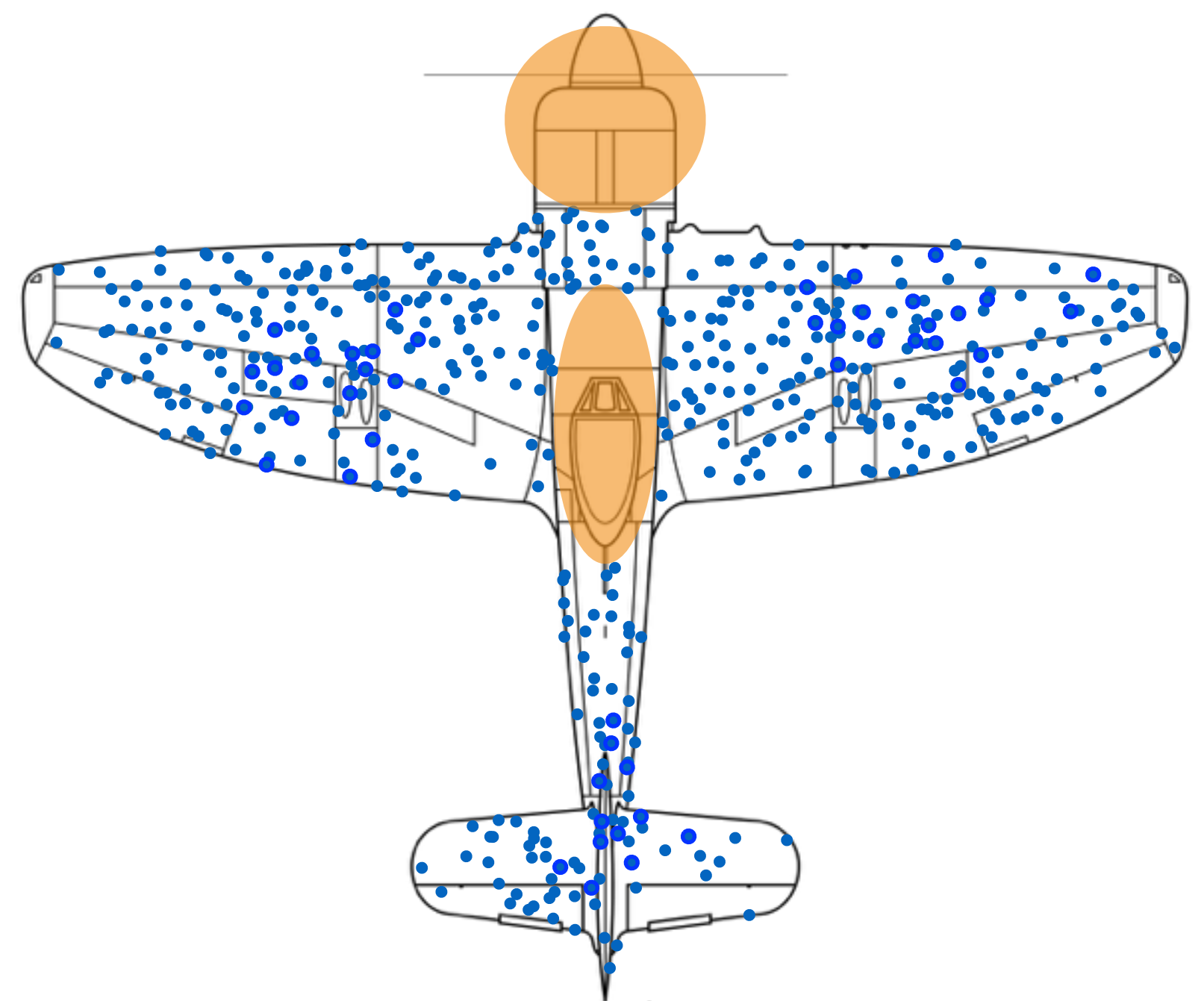
Diagram of the Causes of Mortality in the Army in the East

- Preventable or Mitigable Zymotic disease
- Wounds
- All other causes



The black line across November 1854 marks the boundary of the deaths from all other causes during that month. In October 1854, the black coincides with the red.

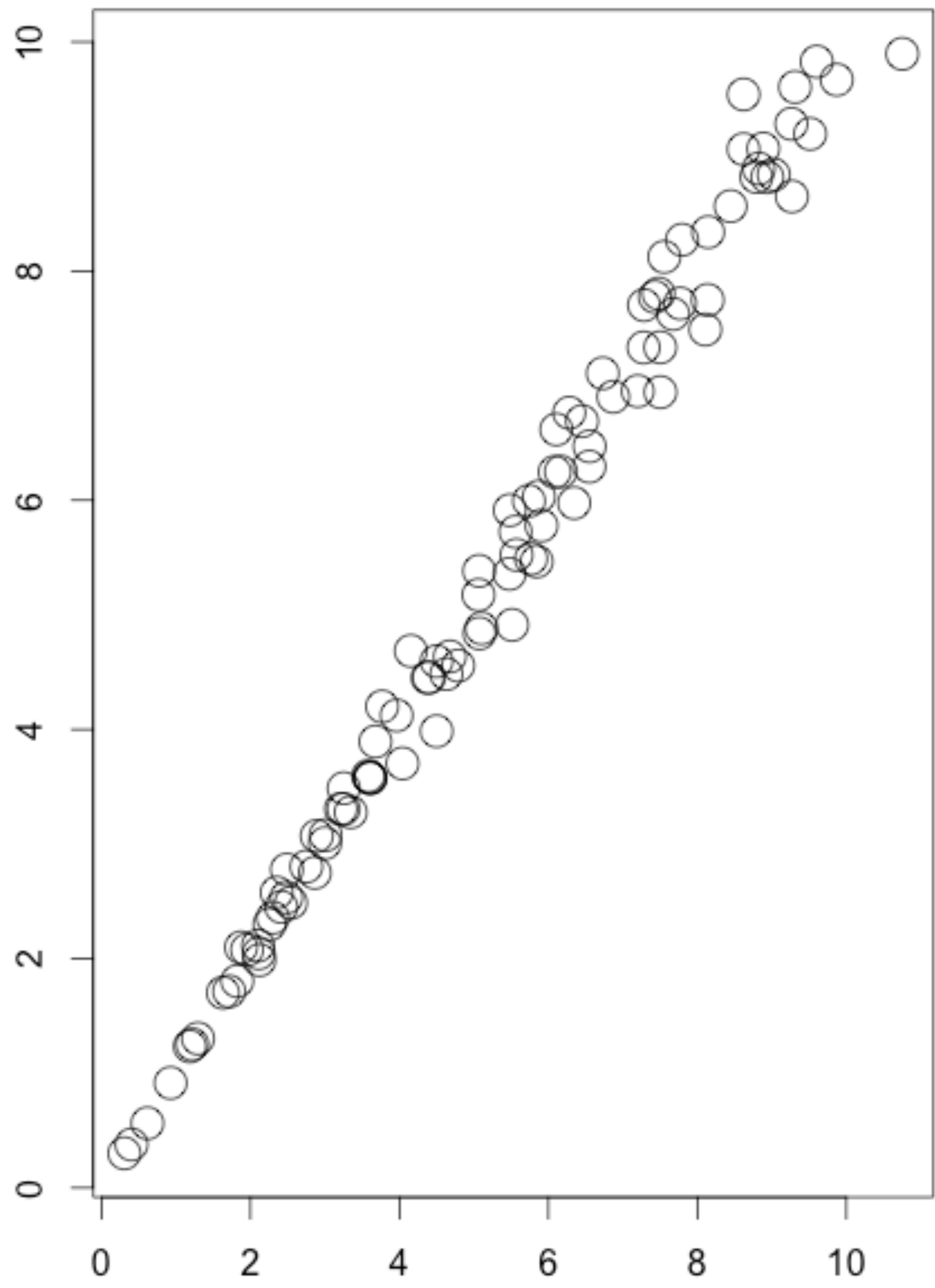
Florence Nightingale 1856



A long-exposure photograph of a comet streaking across a starry night sky. The comet's tail is a bright, horizontal streak of light, transitioning from a white-yellow core on the left to a blue-purple glow on the right. The background is a dense field of stars of various colors, including white, yellow, and red, set against a dark, deep blue-purple sky.

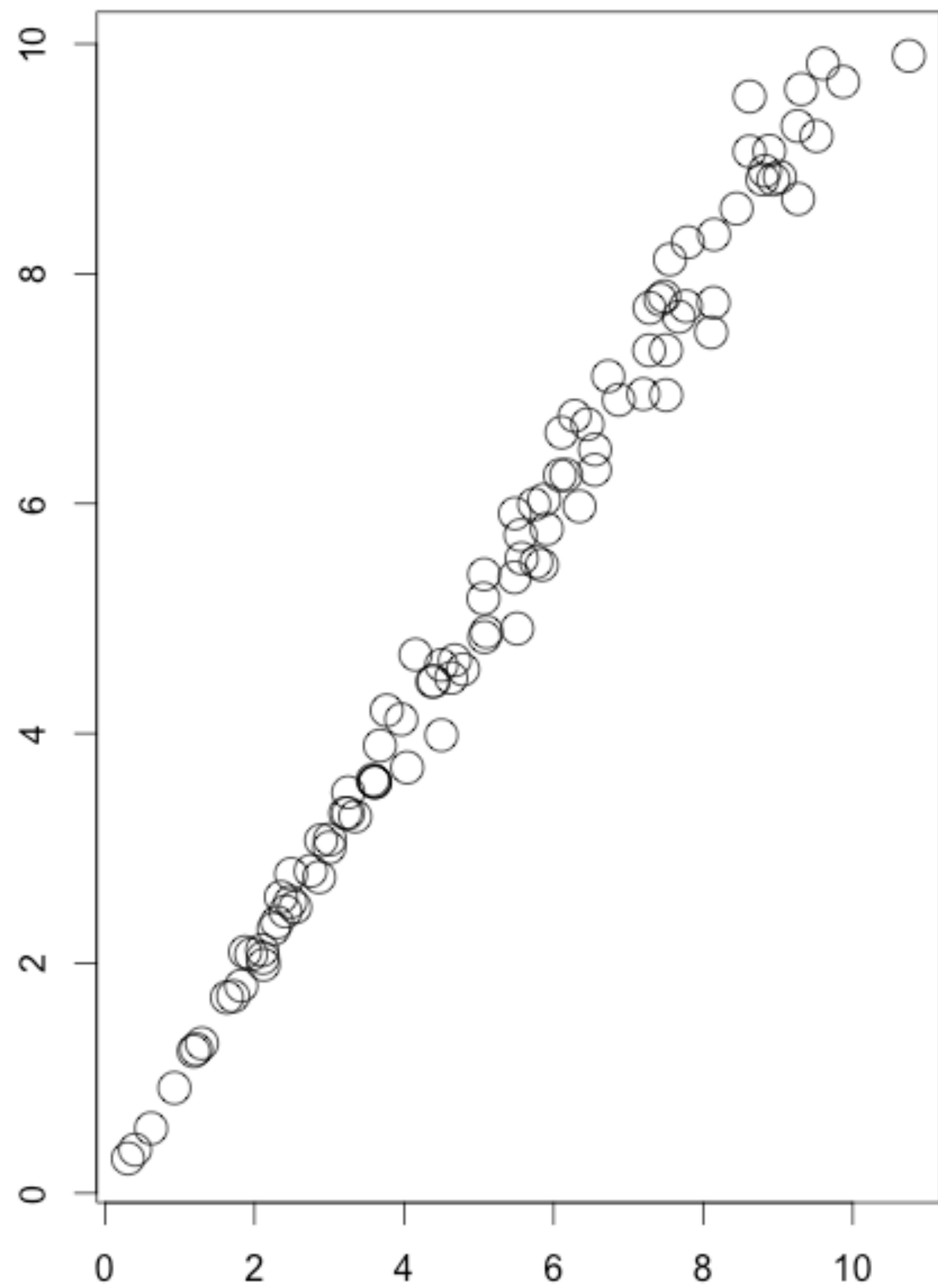
2. Embrace Determinism

A



F

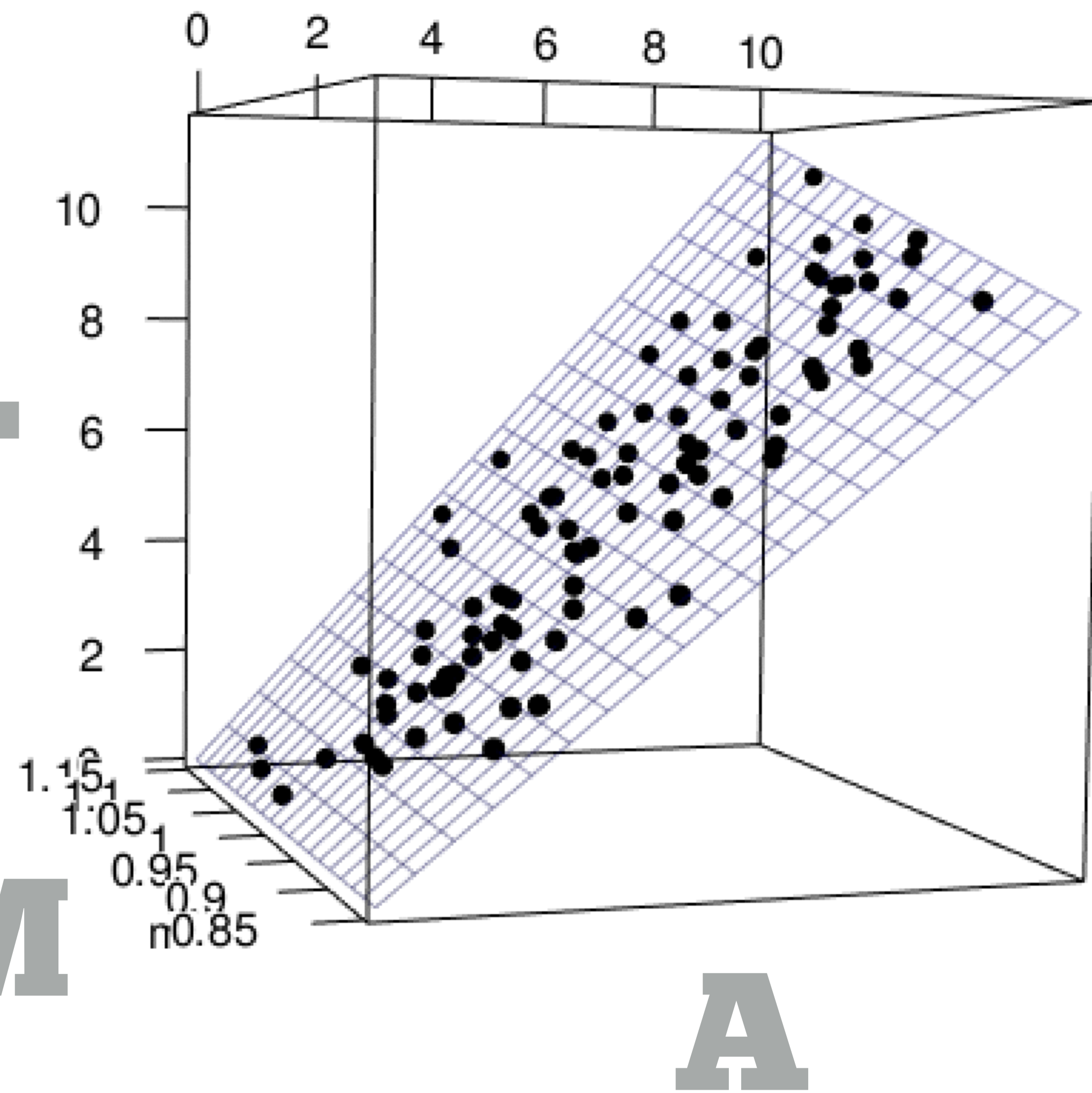
A



F

F

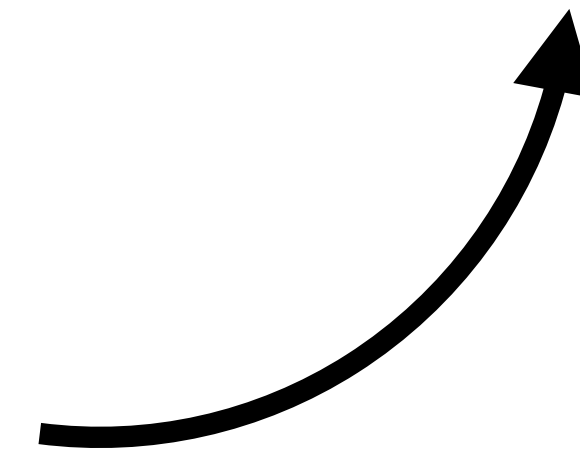
M



A

$$\hat{Y} = \alpha + \beta X + \epsilon$$

~~Random errors~~



Effects of unmeasured variables

3. Embrace Philosophy of Science



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions,

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05. Research is not most appropriately represented and summarized by p -values, but, unfortunately, there is a widespread

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability

Quantifying Global Warming from the Retreat of Glaciers

Johannes Oerlemans

Safety of Newly Approved Drugs Implications for Prescribing

Robert J. Temple, MD

Martin H. Himmel, MD, MPH

drawal or black box warning, able, as the earlier detection mean fewer late discoveries.

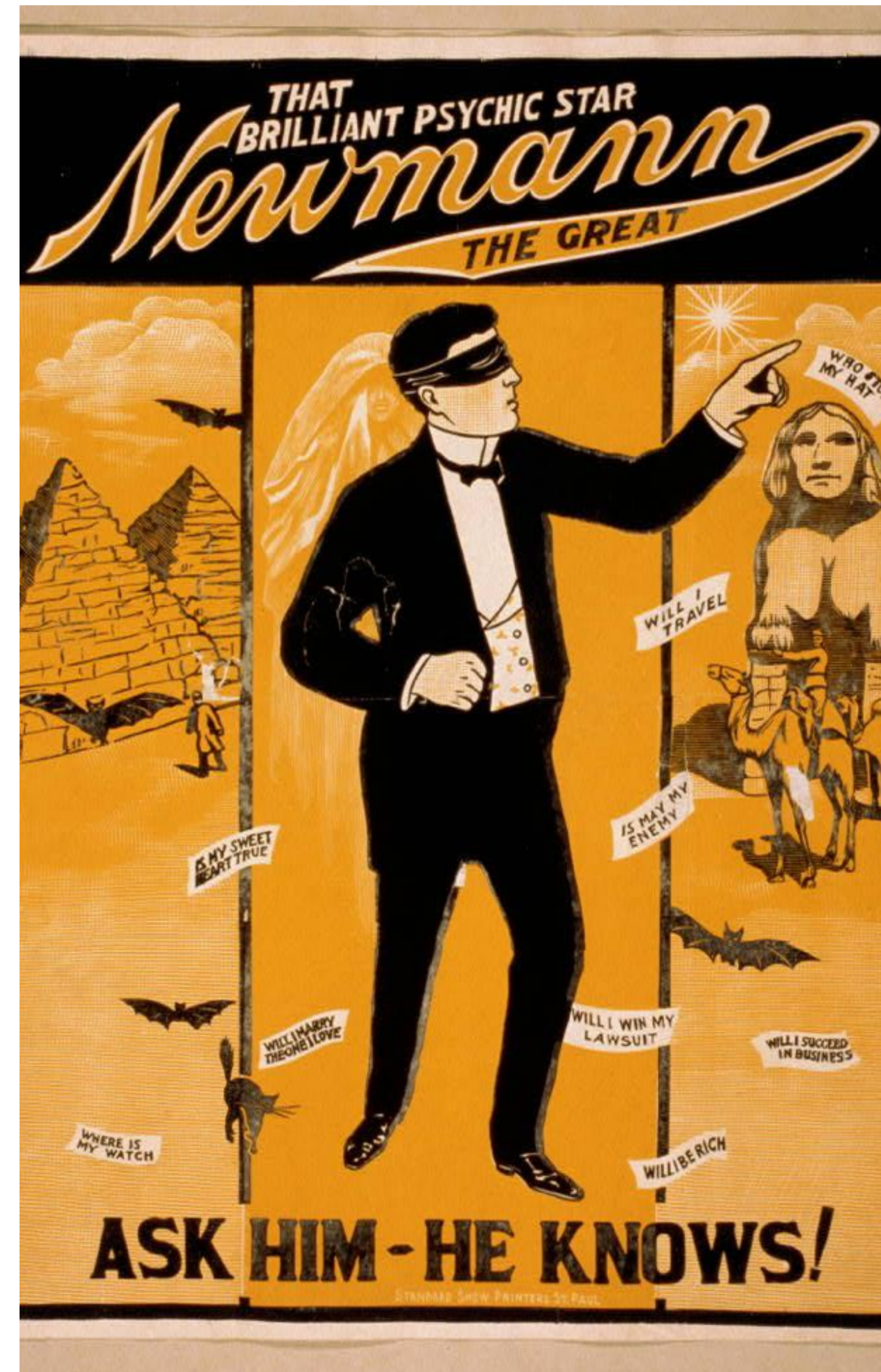
Closing in on a Breast Cancer Gene

J. M. Hall,* L. Friedman,* C. Guenther,* M. K. Lee,* and M.-C. King*

*School of Public Health and Department of Molecular and Cell Biology, University of Wisconsin-Madison, Marshfield, WI; and ‡Imperial Cancer Research Fund, London

Housing valuations: no bubble apparent

Kathleen Stephansen and Maxine Koster



A close-up, artistic photograph of an acoustic guitar. The focus is on the soundhole, which is surrounded by a decorative, multi-layered rosette. The guitar's body is a warm, reddish-brown color, and the neck is dark wood with light-colored frets. The strings are visible, extending from the soundhole towards the top right. The lighting is dramatic, highlighting the textures of the wood and the intricate details of the rosette.

4. Be Picky about Statistics

The Introductory Statistics Course: A Ptolemaic Curriculum

George W. Cobb
Mount Holyoke College

Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up

George COBB

The last half-dozen years have seen *The American Statistician* publish well-argued and provocative calls to change our thinking about statistics and how we teach it, among them Brown and Kass, Nolan and Temple-Lang, and Legler et al. Within this past year, the ASA has issued a new and comprehensive set

rethink our curriculum from the ground up, starting necessarily with alternatives to the former consensus introductory course, but with a more ambitious goal to rebuild the entire undergraduate statistics curriculum. In my 2005 address at USCOTS (U. S. Conference on Teaching Statistics), I argued that the standard introductory course, which puts the normal distribution at its

The found
old of a f
and how
For two t
ago, the
applied
meant ari
eye that e

Keep

- 1.** Modeling and Machine Learning
- 2.** Bootstrap (for inference)
- 3.** Cross Validation (for model comparison)