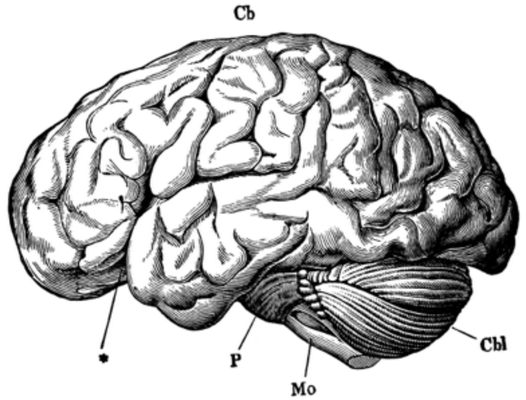
Data Science

Suggestions for statisticians



Garrett Grolemund

Data Scientist, Educator

March 2016


RStudio®

1. For Statisticians

2. For Universities

1. For Statisticians

2. For Universities


SCIENCE CENTER


garrettgrolemund — 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

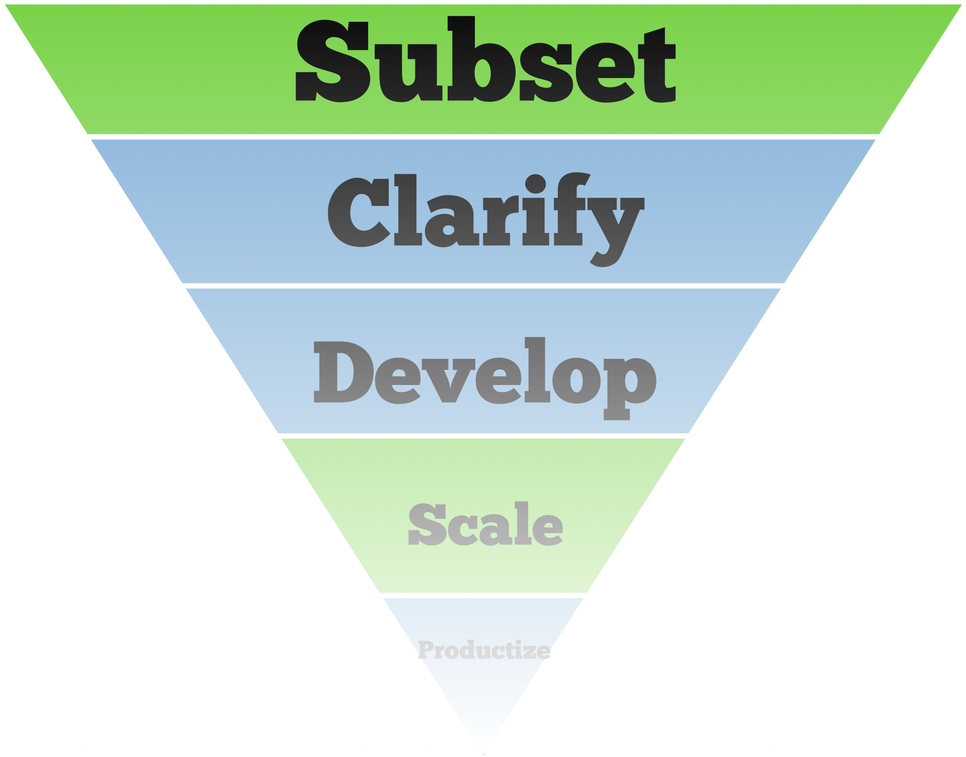
Automate and integrate

Publish

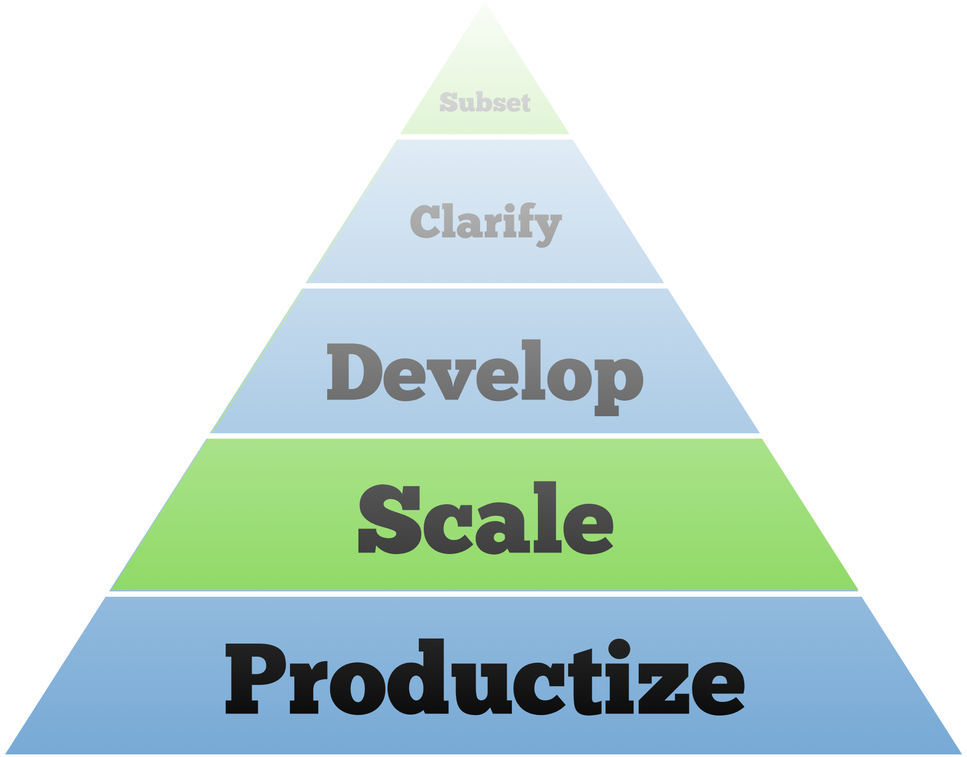
Socialize

* sometimes

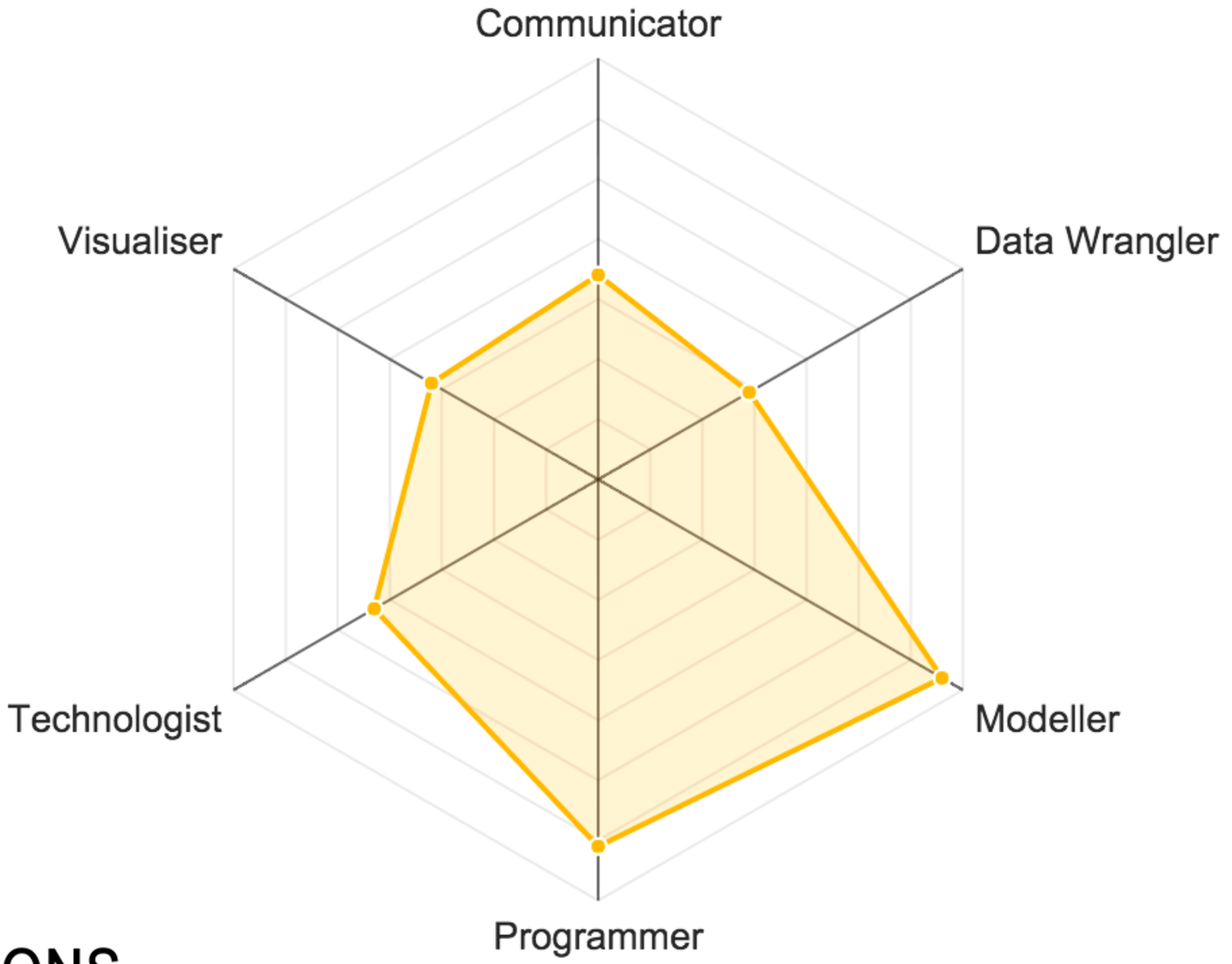
Data Roles



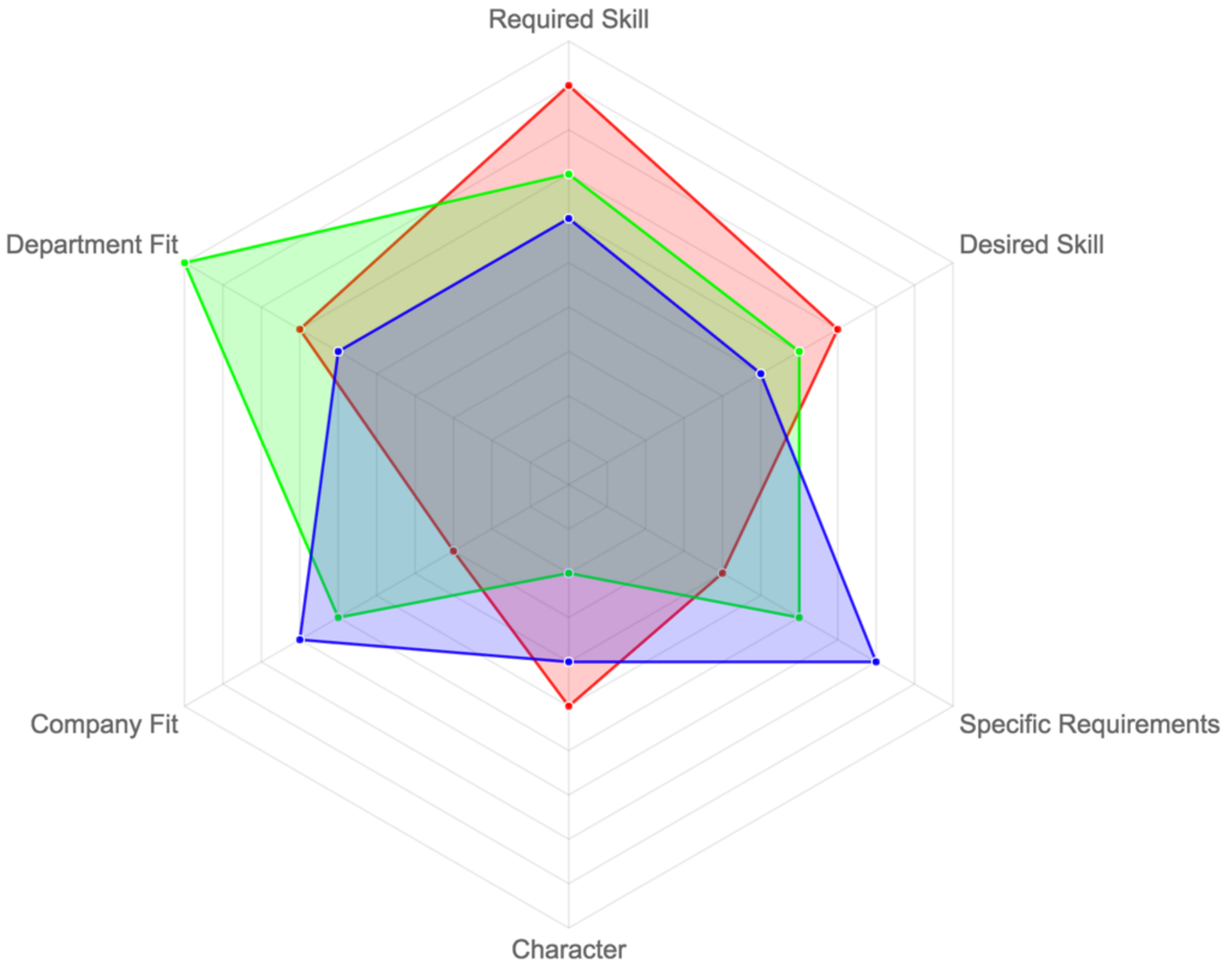
Analyst



IT/Manager/Engineer



MANGOSOLUTIONS
data analysis that delivers

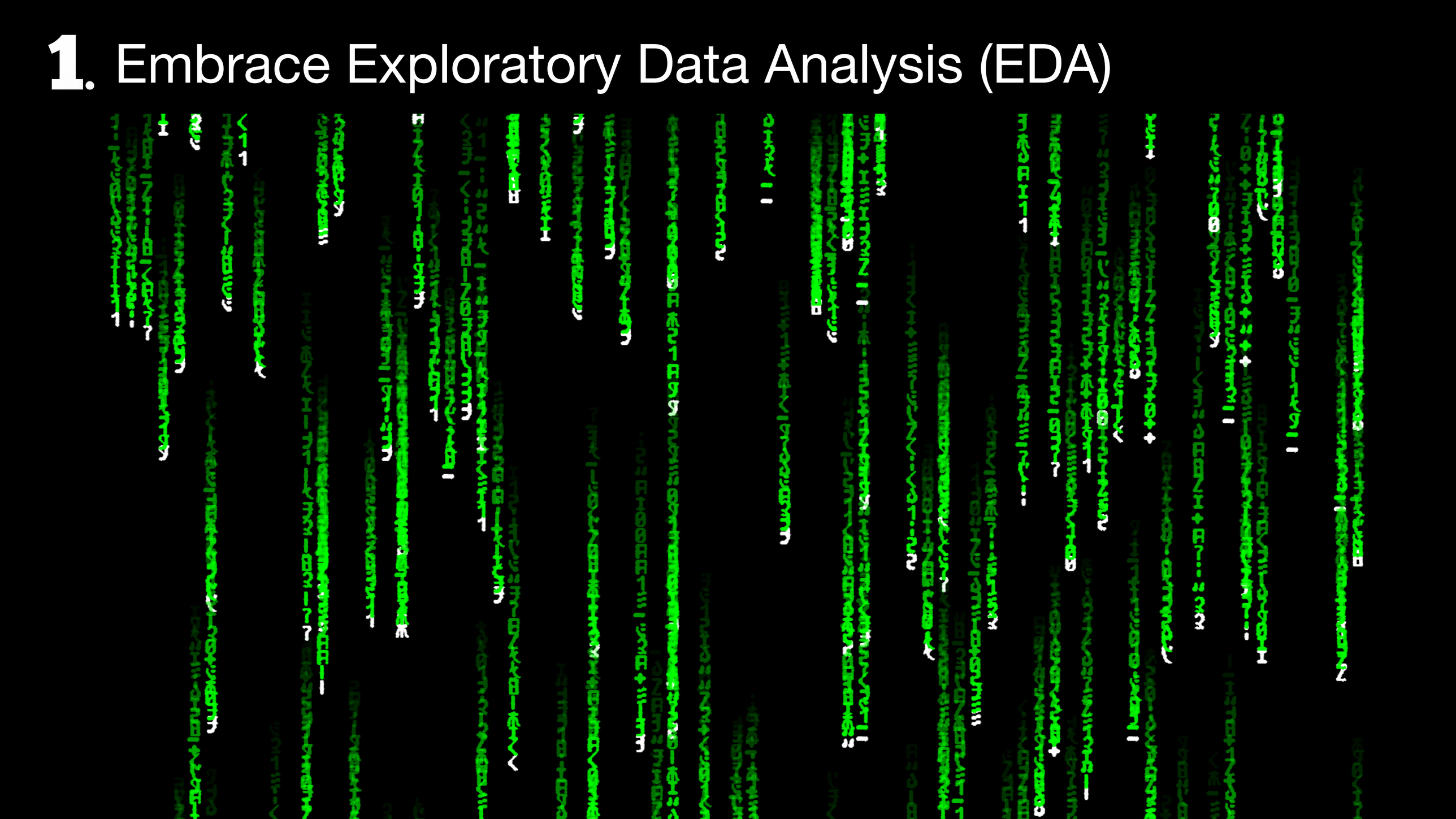


MANGOSOLUTIONS
data analysis that delivers

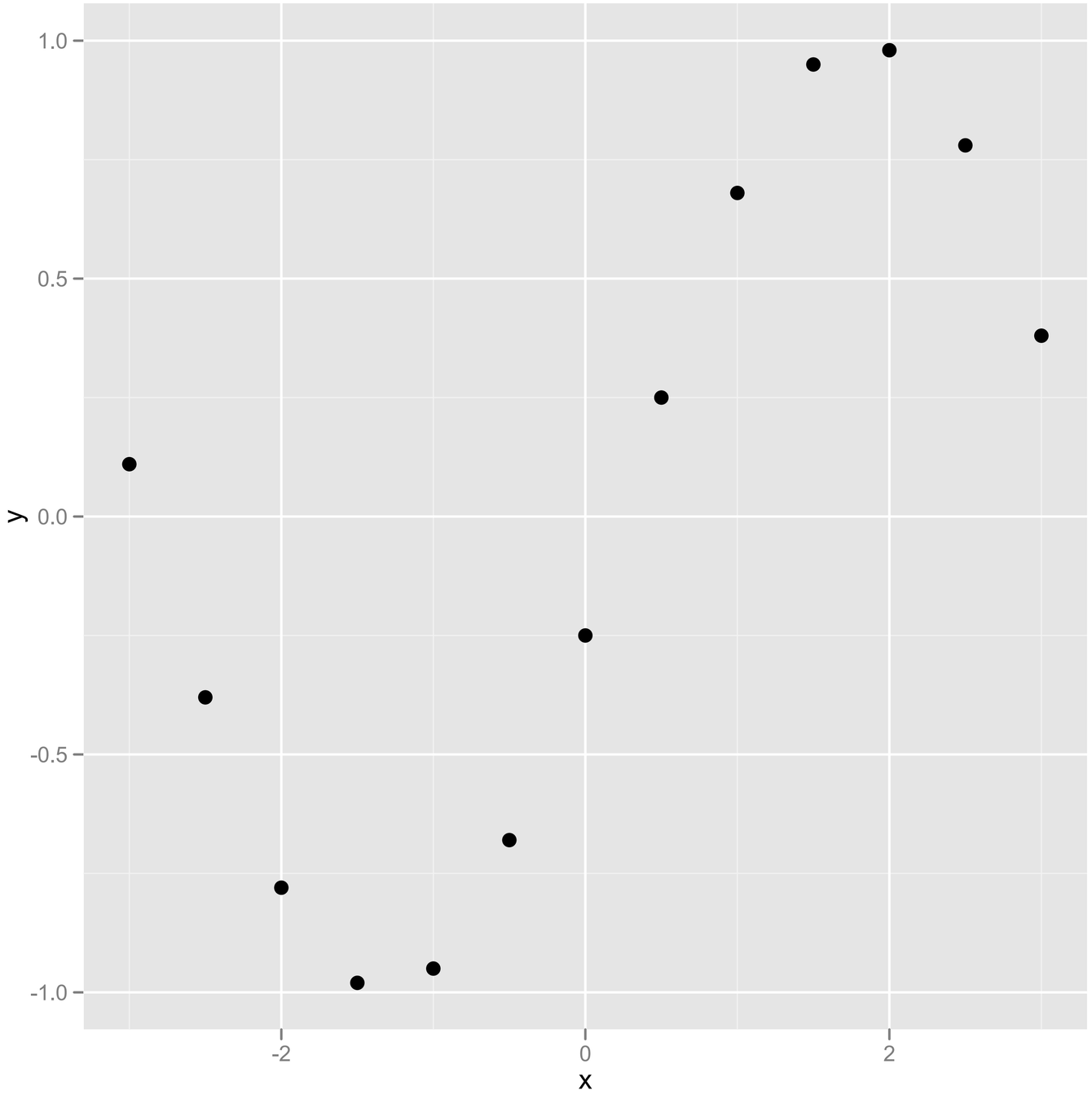
1. For Statisticians

2. For Universities

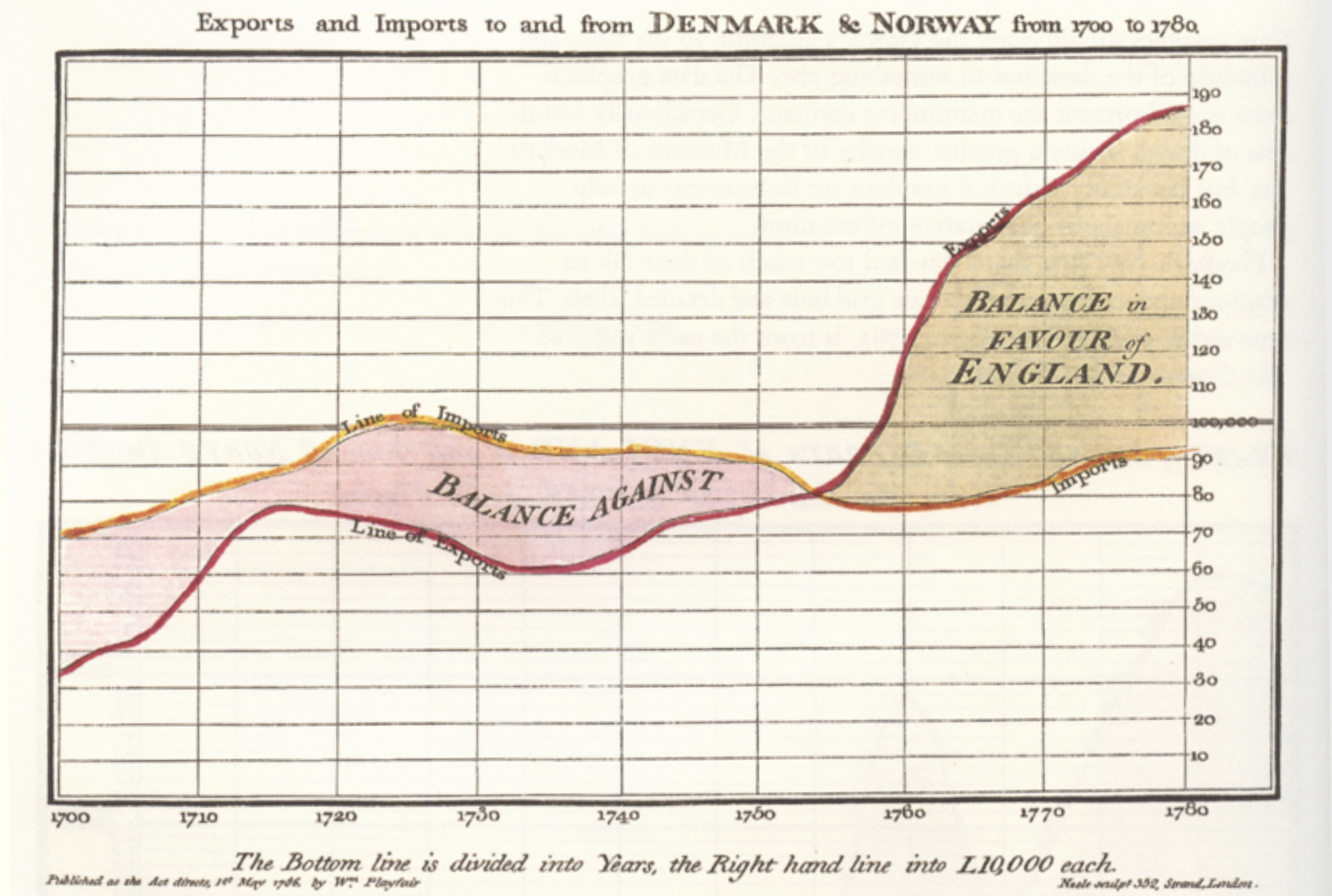
1. Embrace Exploratory Data Analysis (EDA)

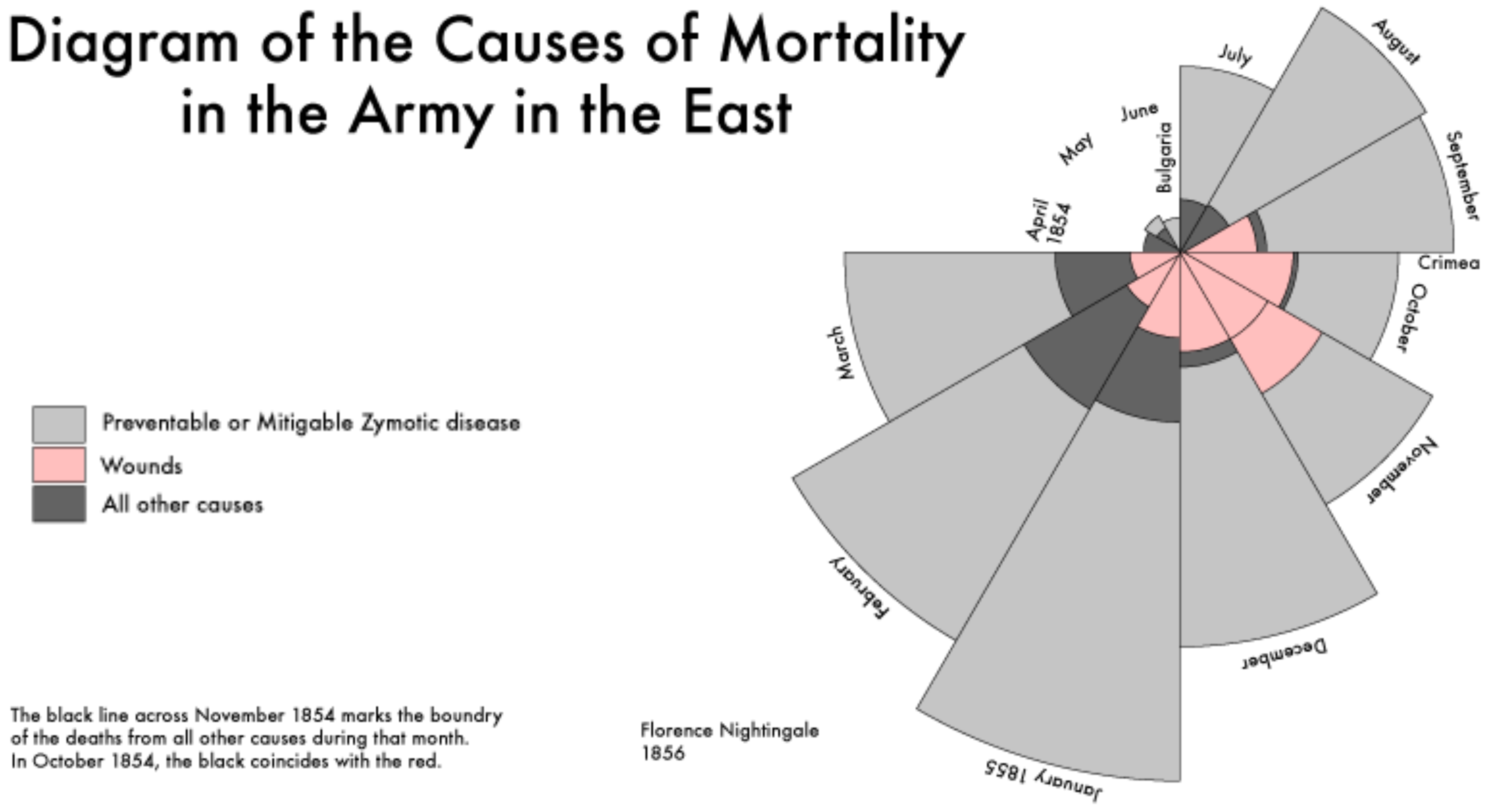


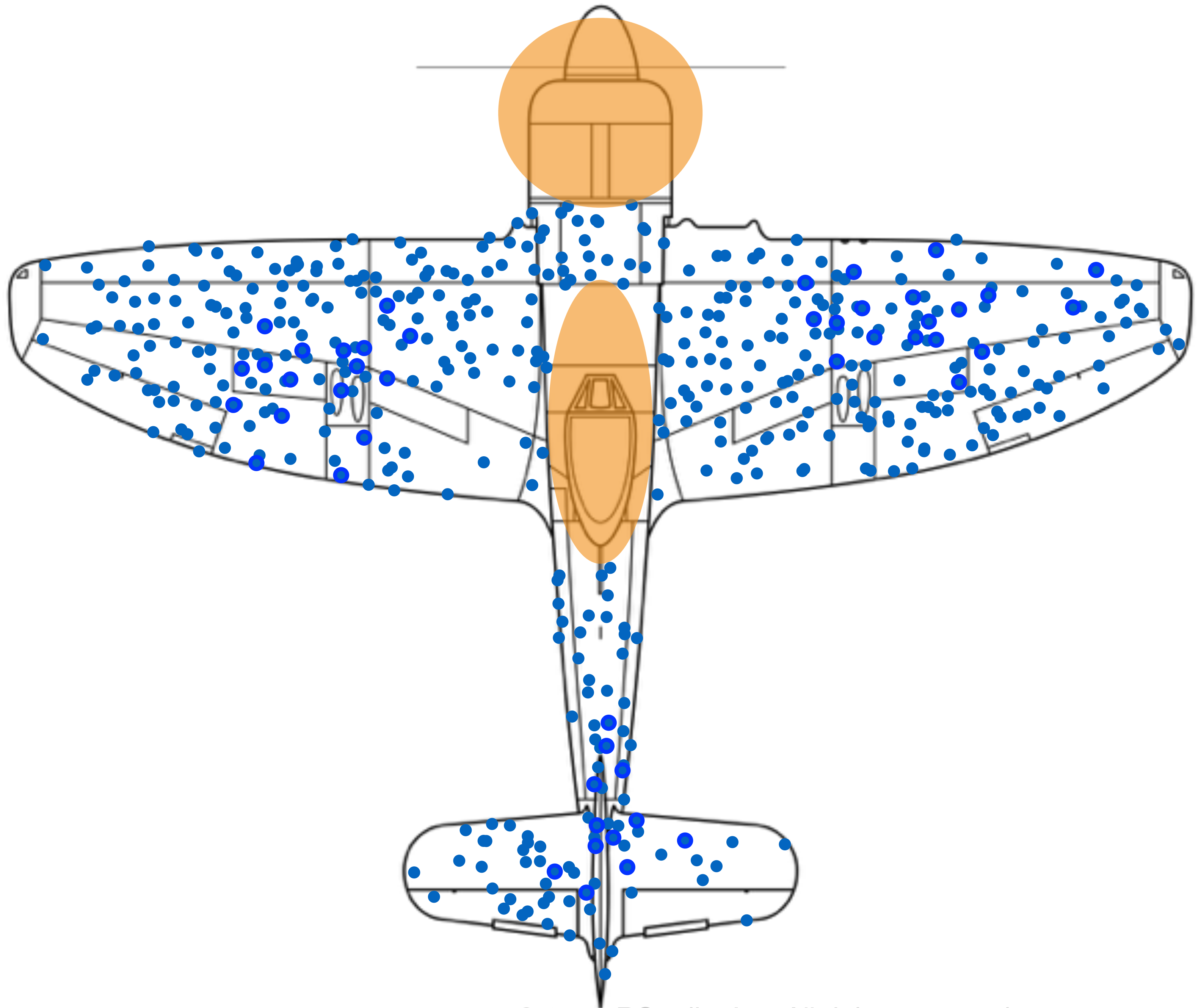
x	y
0	-0.25
-0.5	-0.68
-1	-0.95
-1.5	-0.98
-2.5	-0.38
0.5	0.25
2	0.98
1.5	0.95
3	0.38
1	0.68
2.5	0.78
-3	0.11
-2	-0.78



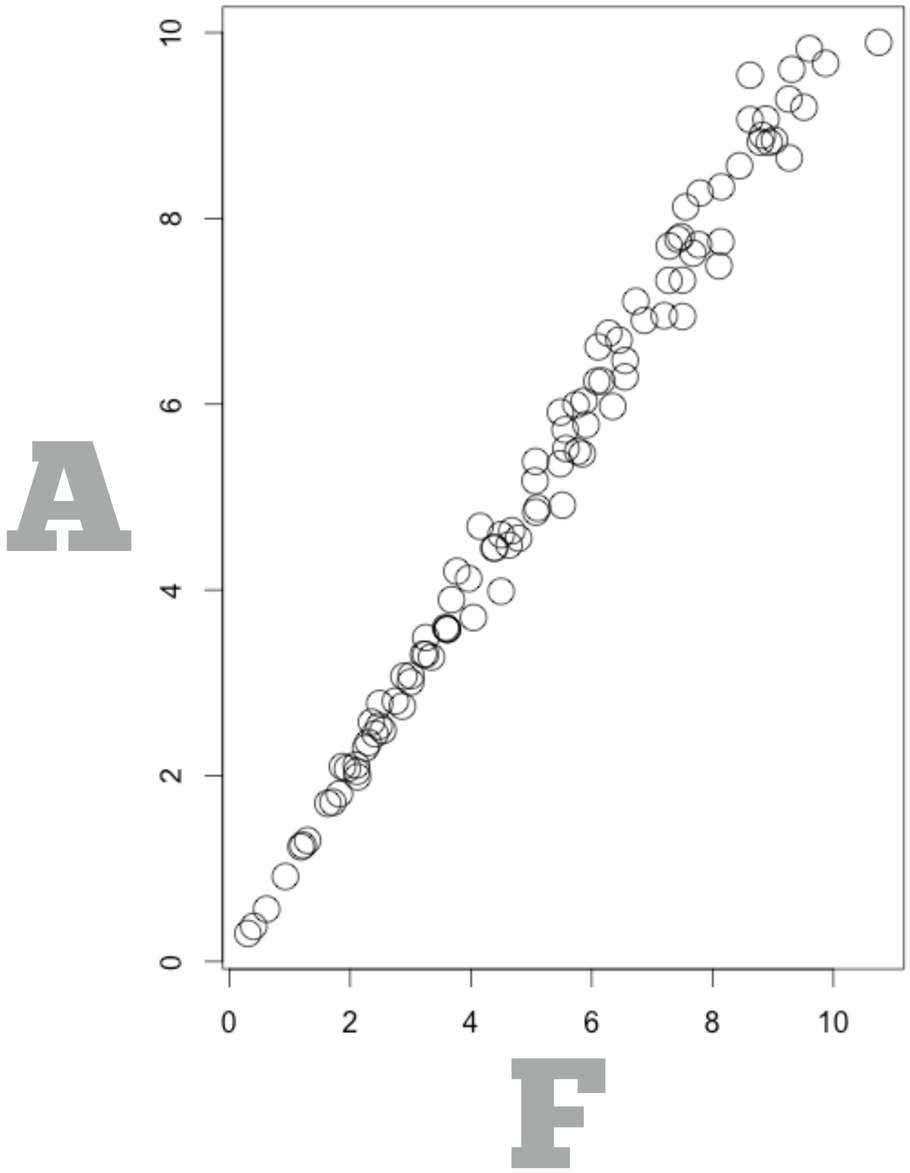
x	y
0	-0.25
-0.5	-0.68
-1	-0.95
-1.5	-0.98
-2.5	-0.38
0.5	0.25
2	0.98
1.5	0.95
3	0.38
1	0.68
2.5	0.78
-3	0.11
-2	-0.78

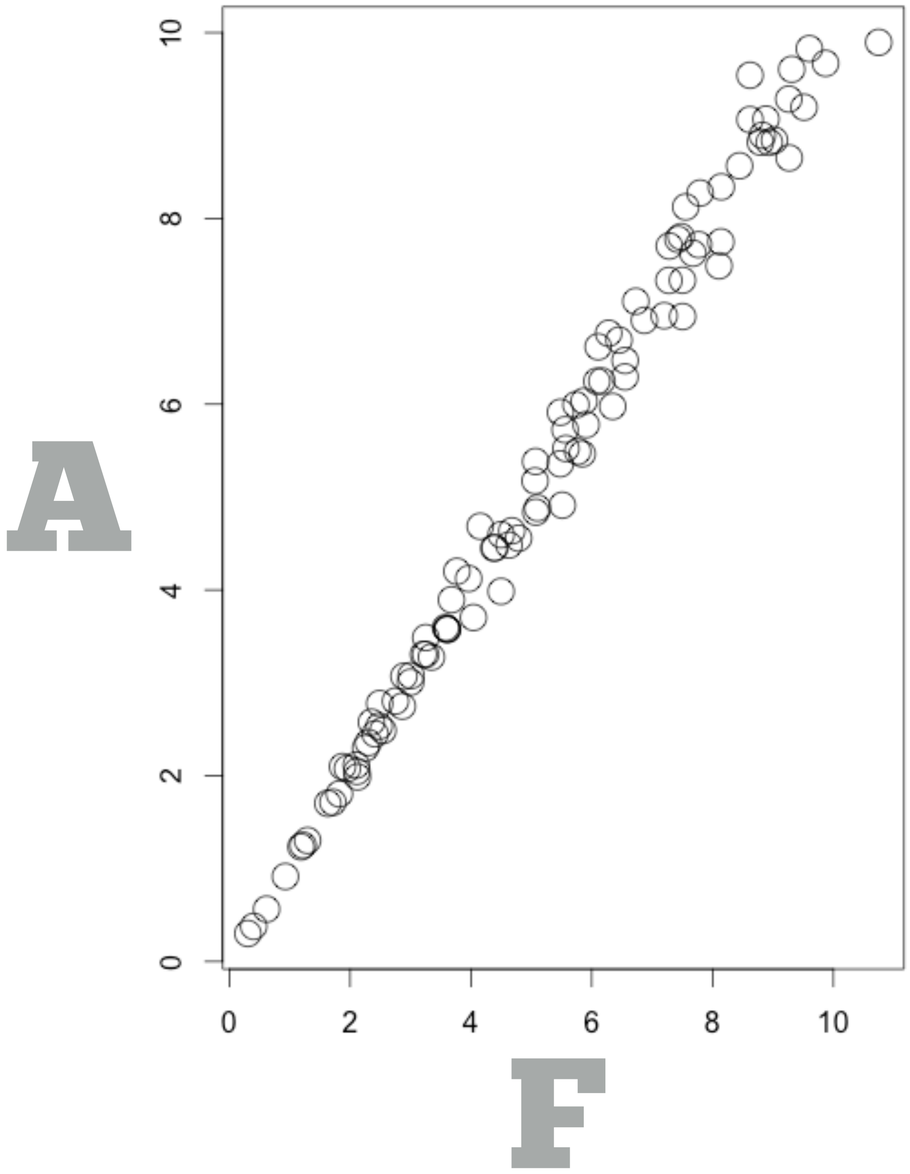


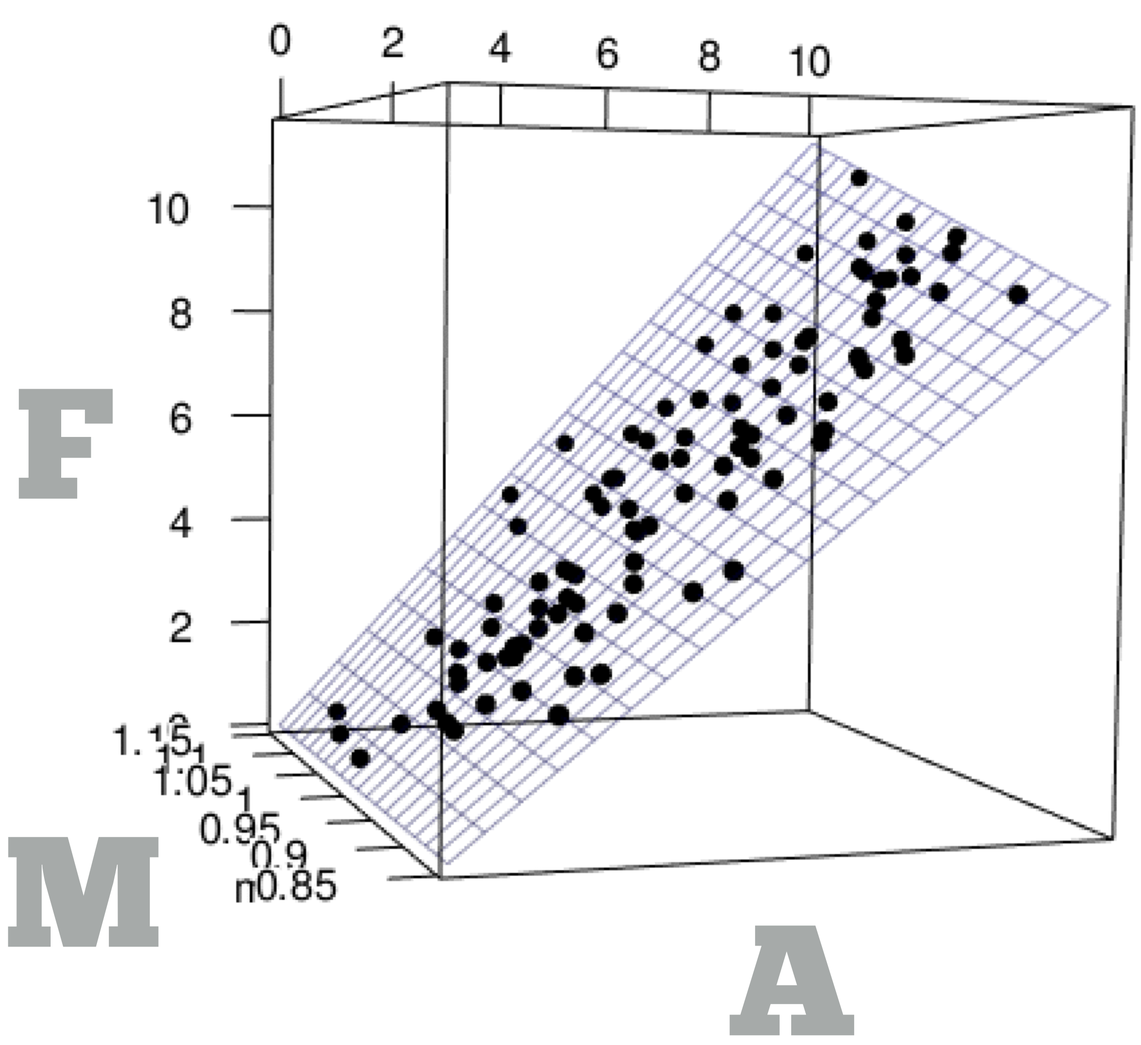




2. Embrace Determinism







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

~~Random errors~~

Effects of unmeasured variables

3. Embrace Philosophy of Science

Open access, freely available online

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.

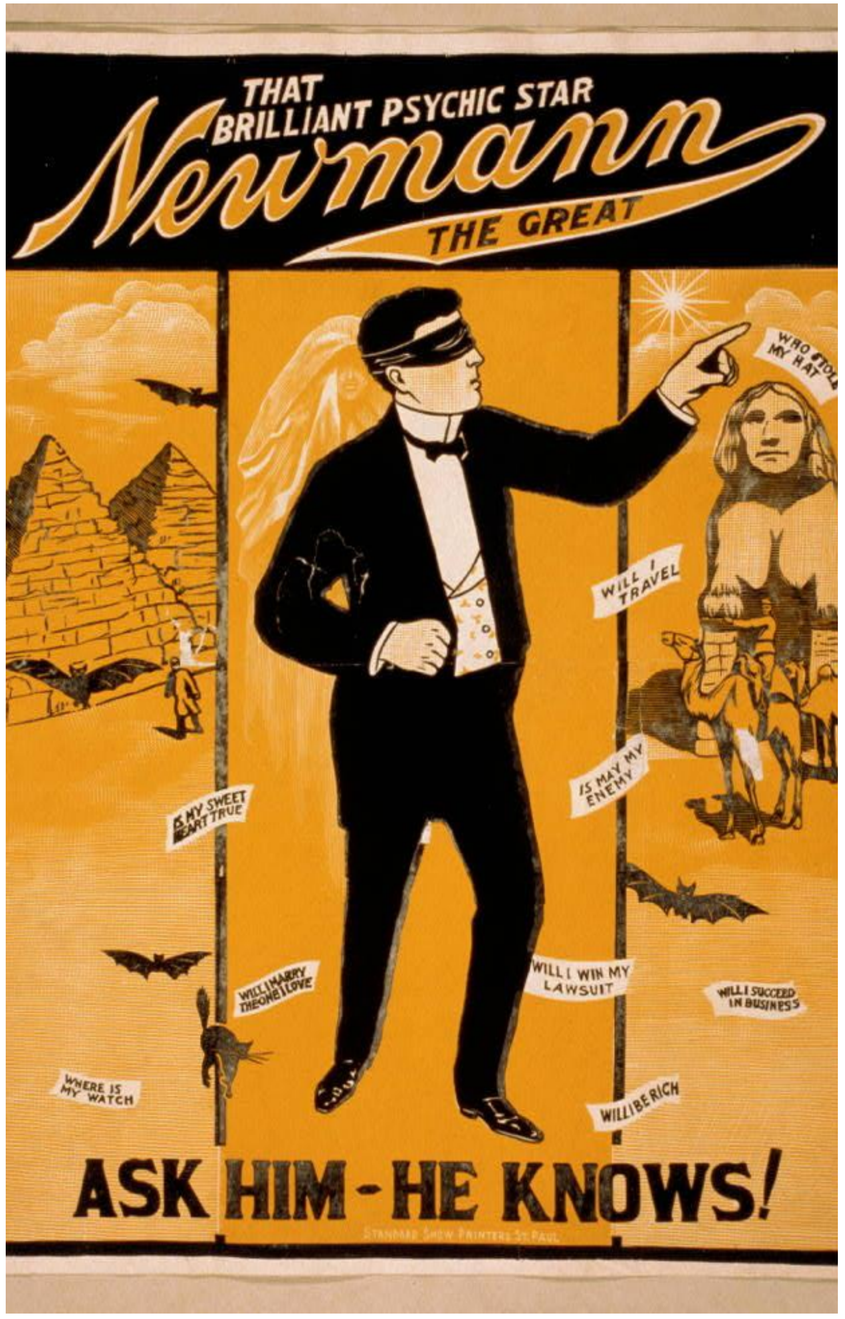
Closing in on a Breast Cancer Gene

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, Universit
Foundation, Marshfield, WI; and ‡Imperial Cancer Research Fund, London

Housing valuations: no bubble apparent

Kathleen Stephansen and Maxine Koster



4. Be Picky about Statistics

The Introductory Statistics Course: A Ptolemaic Curriculum

George W. Cobb
Mount Holyoke College

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

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

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

Keep

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