

Teaching Data Science

Garrett Grolemund

Master Instructor, RStudio

May 2018



- 1. What to teach**
- 2. How to teach it**

A Hands-on Guide to Transforming, Visualizing & Modeling Data



Data Analysis with R

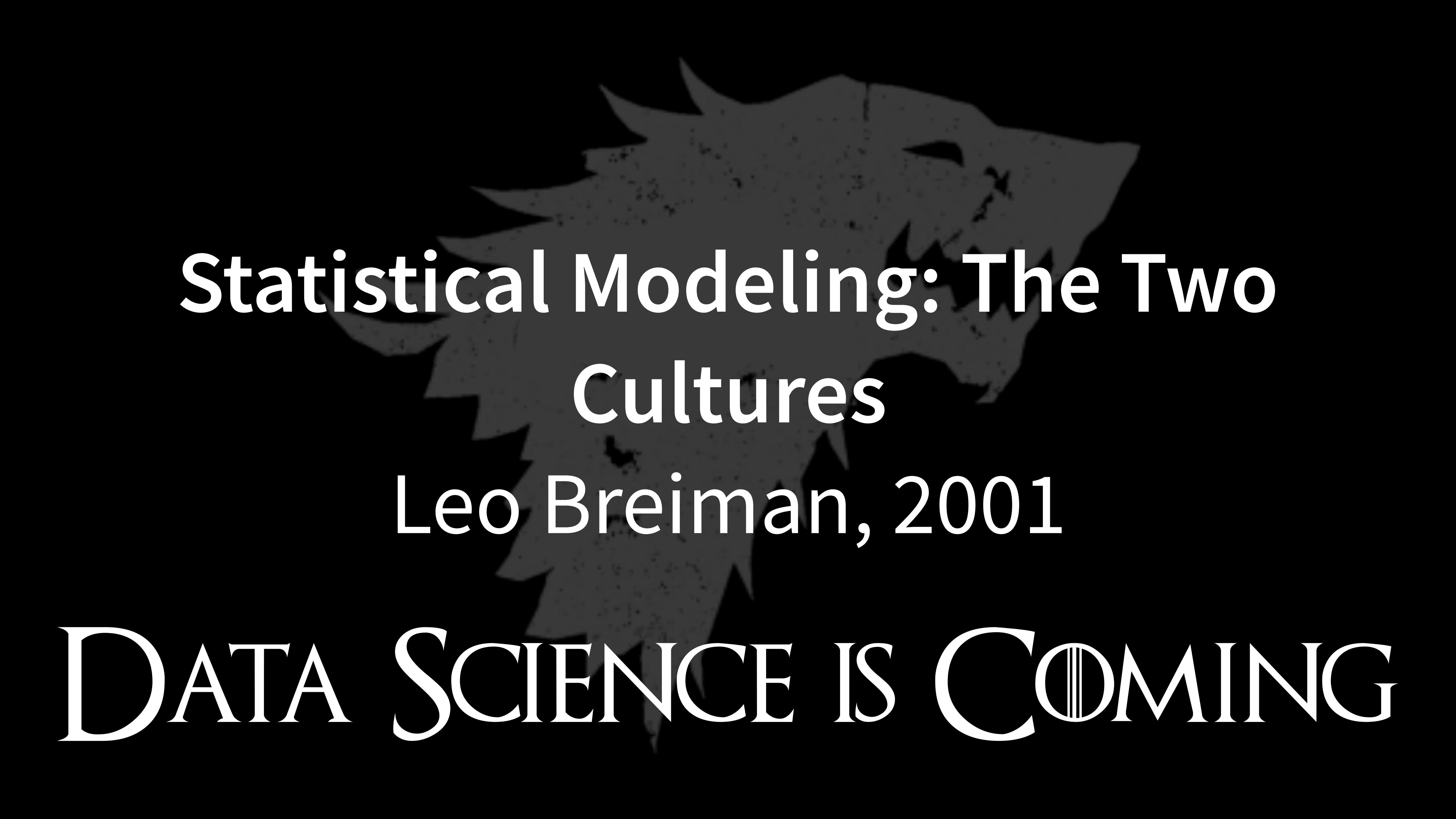
O'REILLY®

Garrett Grolemund





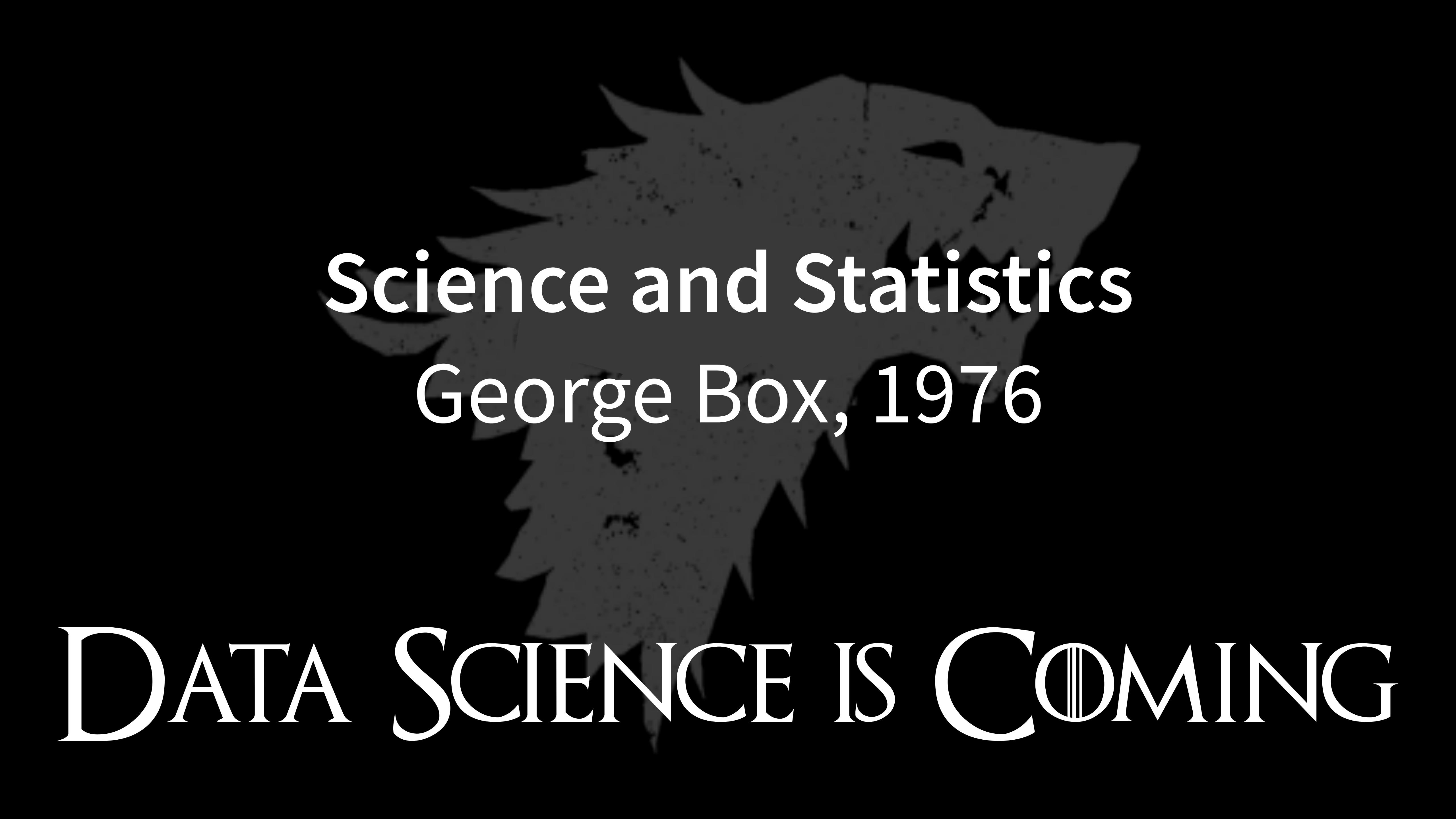
DATA SCIENCE IS COMING



Statistical Modeling: The Two Cultures

Leo Breiman, 2001

DATA SCIENCE IS COMING



Science and Statistics
George Box, 1976

DATA SCIENCE IS COMING



Embracing the "Wider View"
of Statistics
Chris Wild, 1994

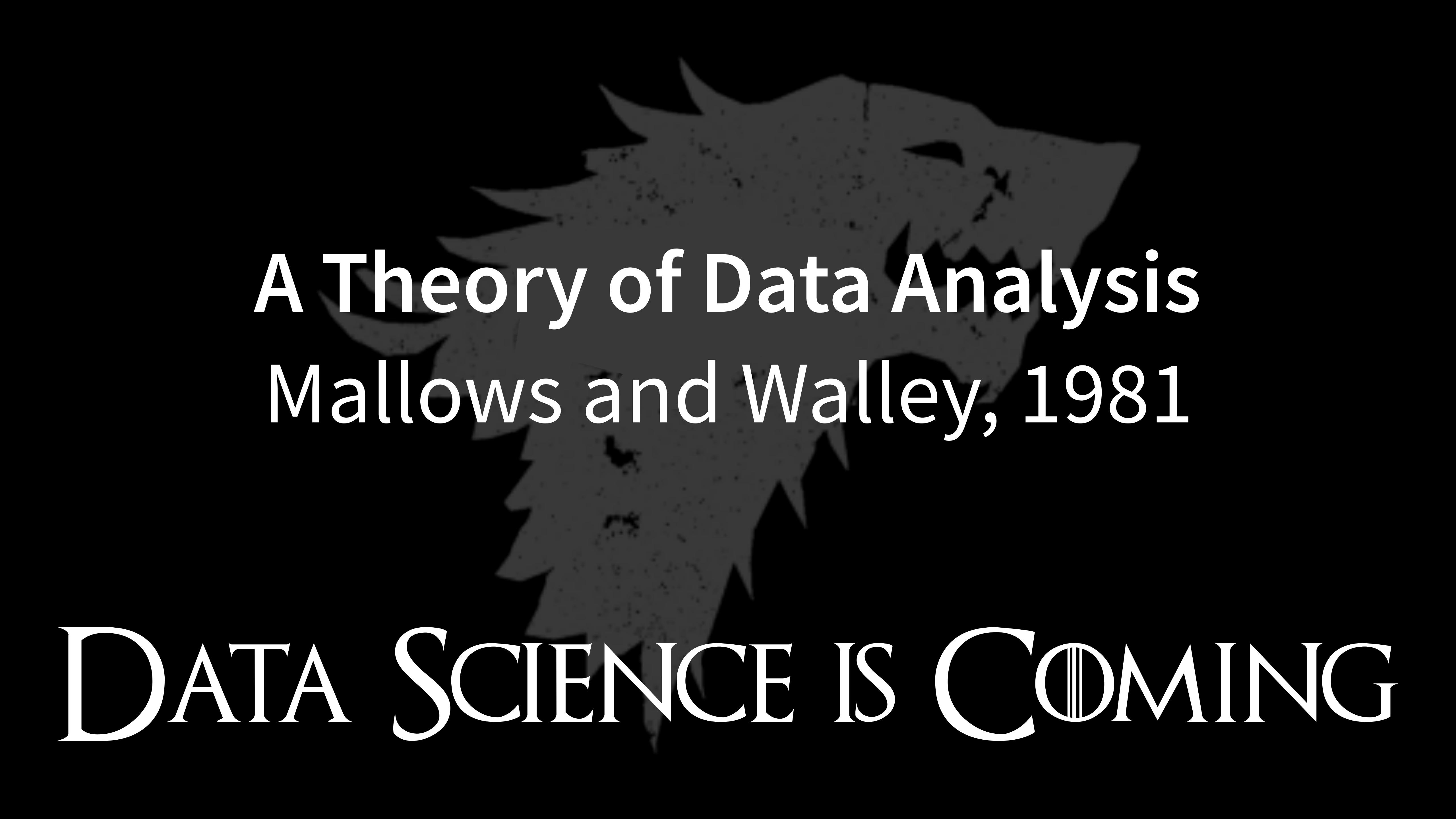
DATA SCIENCE IS COMING



Data Analysis: In Search of an Identity

Peter Huber, 1985

DATA SCIENCE IS COMING



A Theory of Data Analysis

Mallows and Walley, 1981

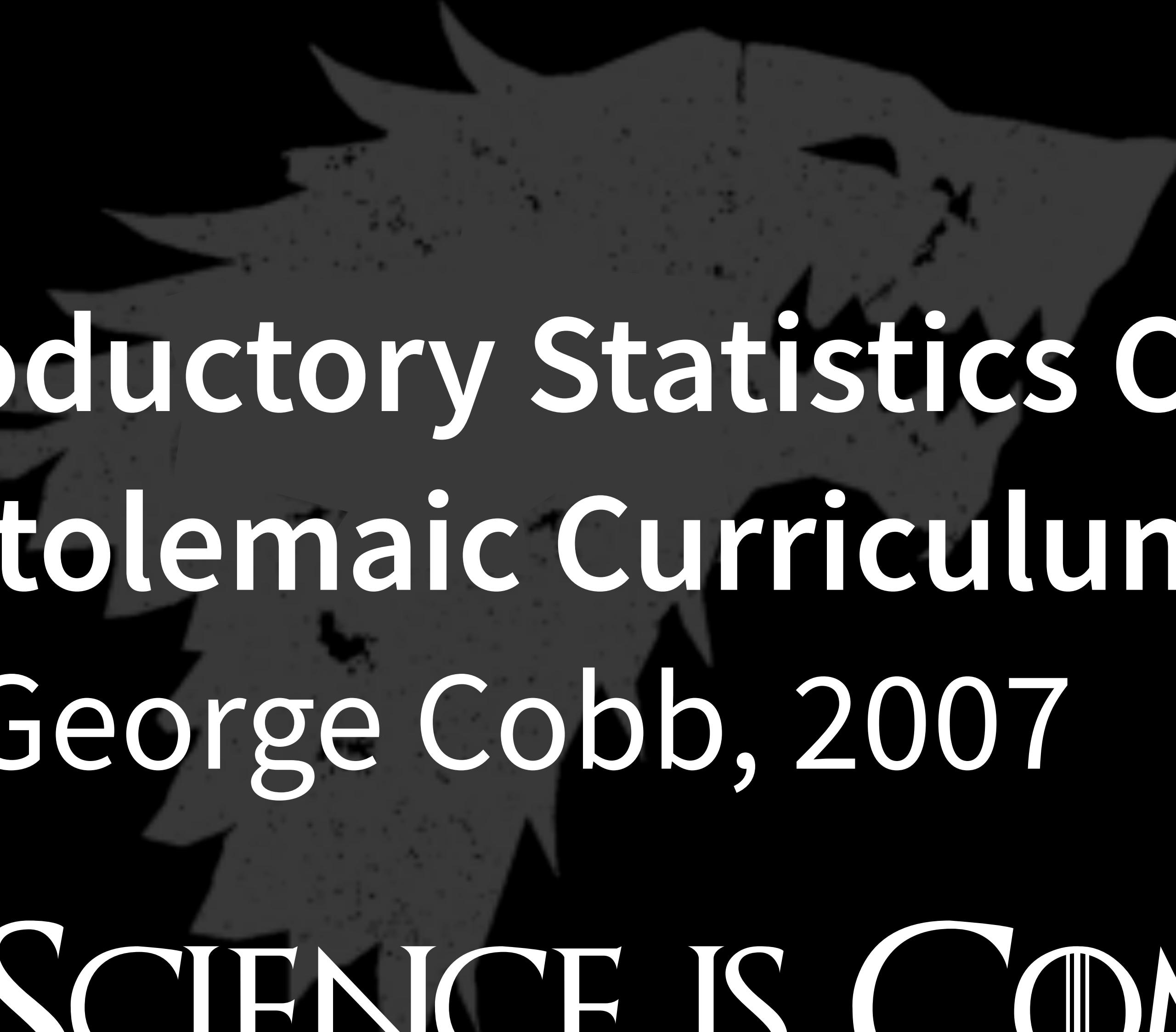
DATA SCIENCE IS COMING



The Future of Data Analysis

John Tukey, 1961

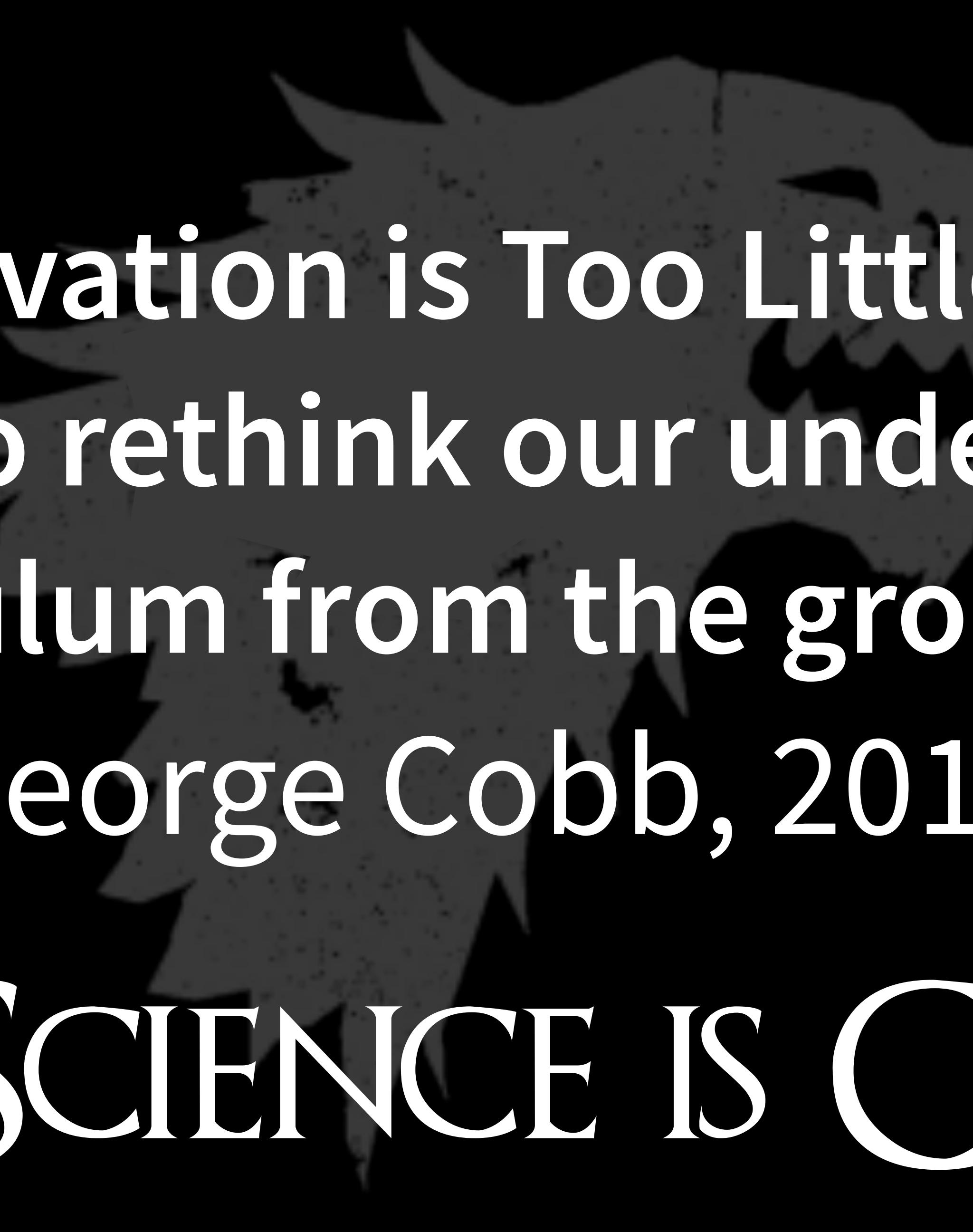
DATA SCIENCE IS COMING



The Introductory Statistics Course: A Ptolemaic Curriculum

George Cobb, 2007

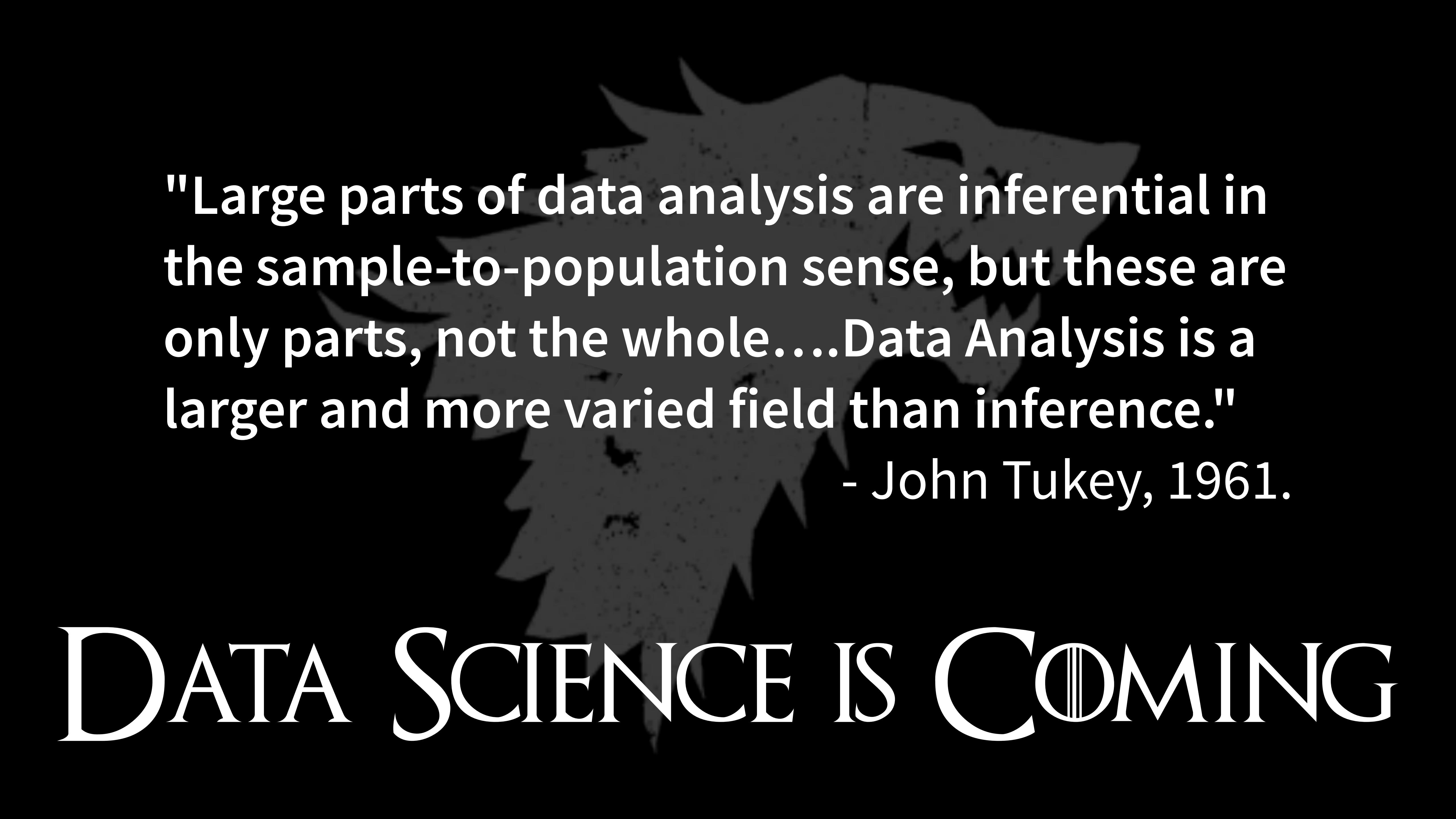
DATA SCIENCE IS COMING



Mere Renovation is Too Little Too Late:
We need to rethink our undergraduate
curriculum from the ground up

George Cobb, 2015

DATA SCIENCE IS COMING



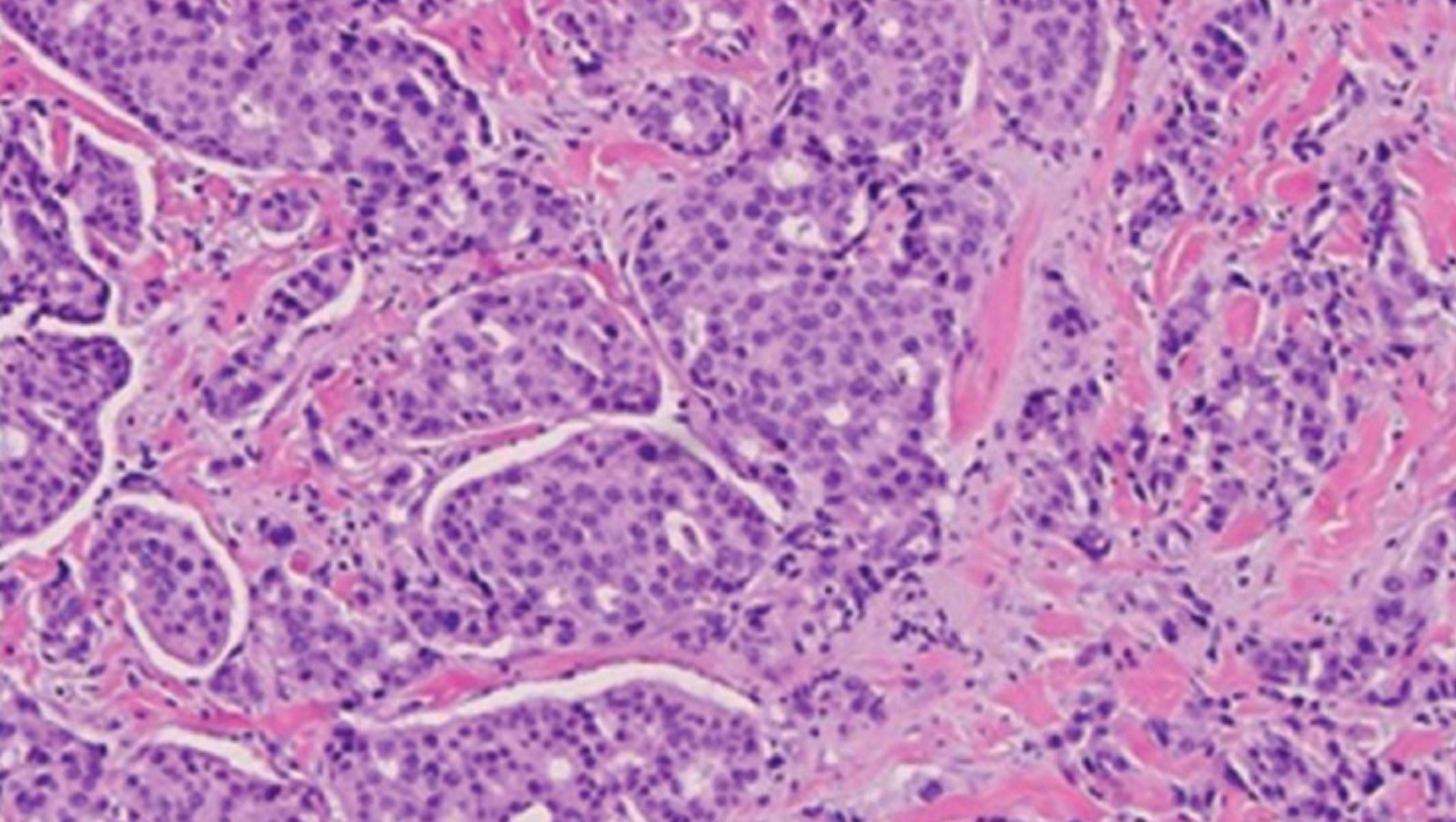
"Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole....Data Analysis is a larger and more varied field than inference."

- John Tukey, 1961.

DATA SCIENCE IS COMING

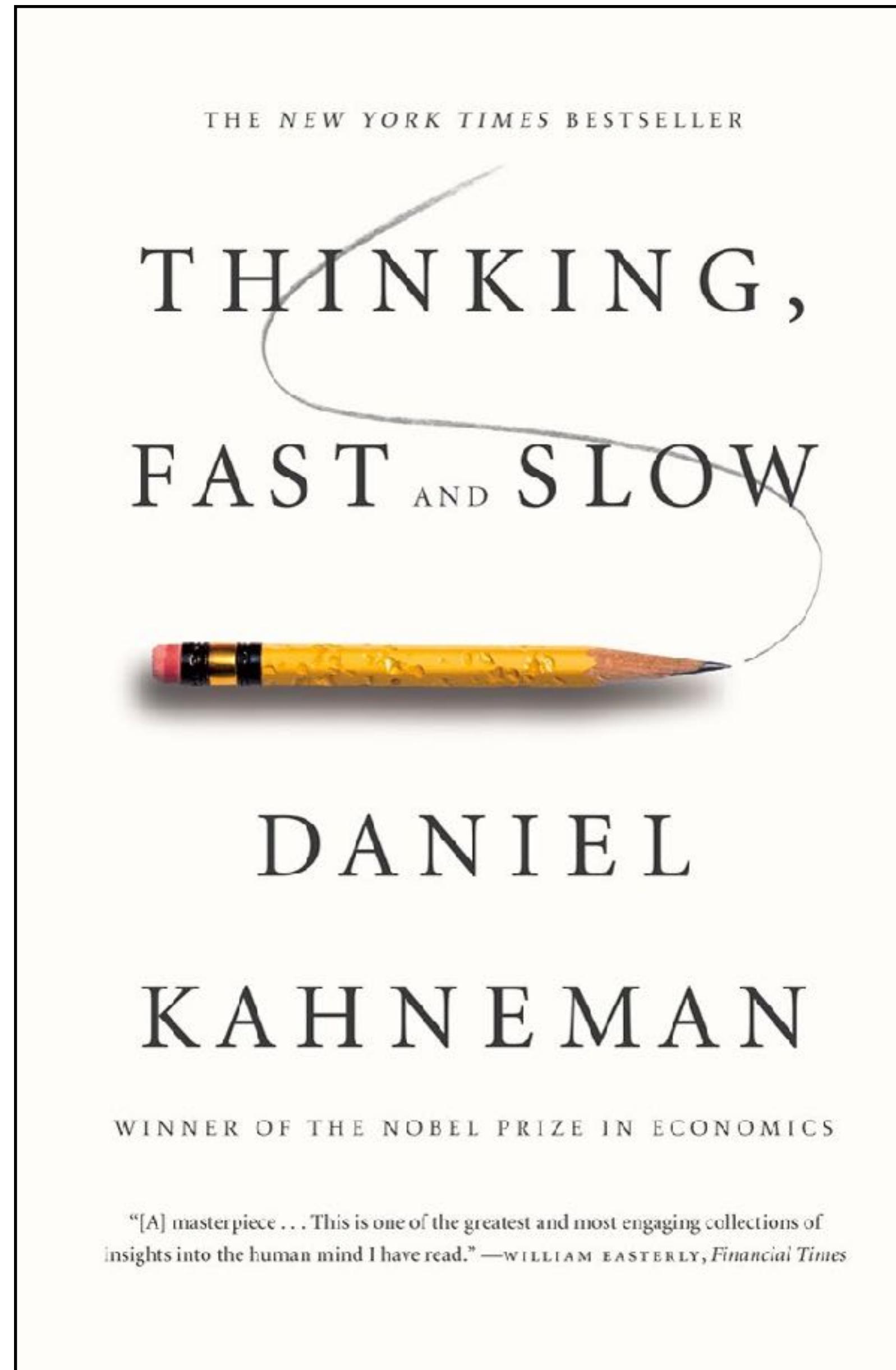
00151





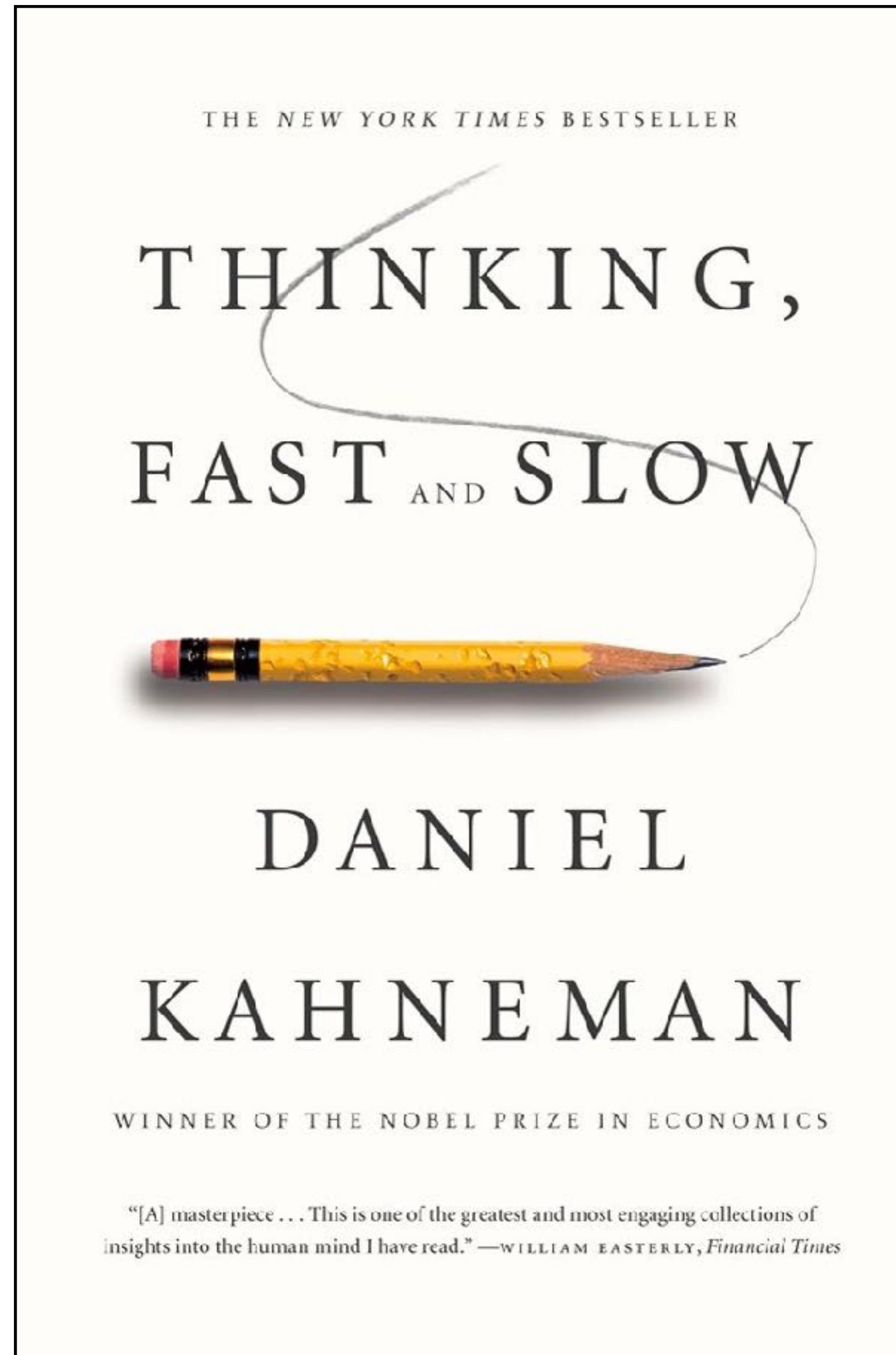
00151





1. **System 1** - automatic, fast, unconscious information processing

**Do more people die from shark attacks
or from falling over?**



- 1. System 1** - a system of heuristics that are highly efficient, but not always correct

- 2. System 2** - slow, expensive, logical, conscious reasoning

"A landmark contribution to humanity's understanding of itself."
—*The New York Times Book Review*

THE RIGHTEOUS MIND



WHY GOOD
PEOPLE ARE DIVIDED
BY POLITICS AND
RELIGION

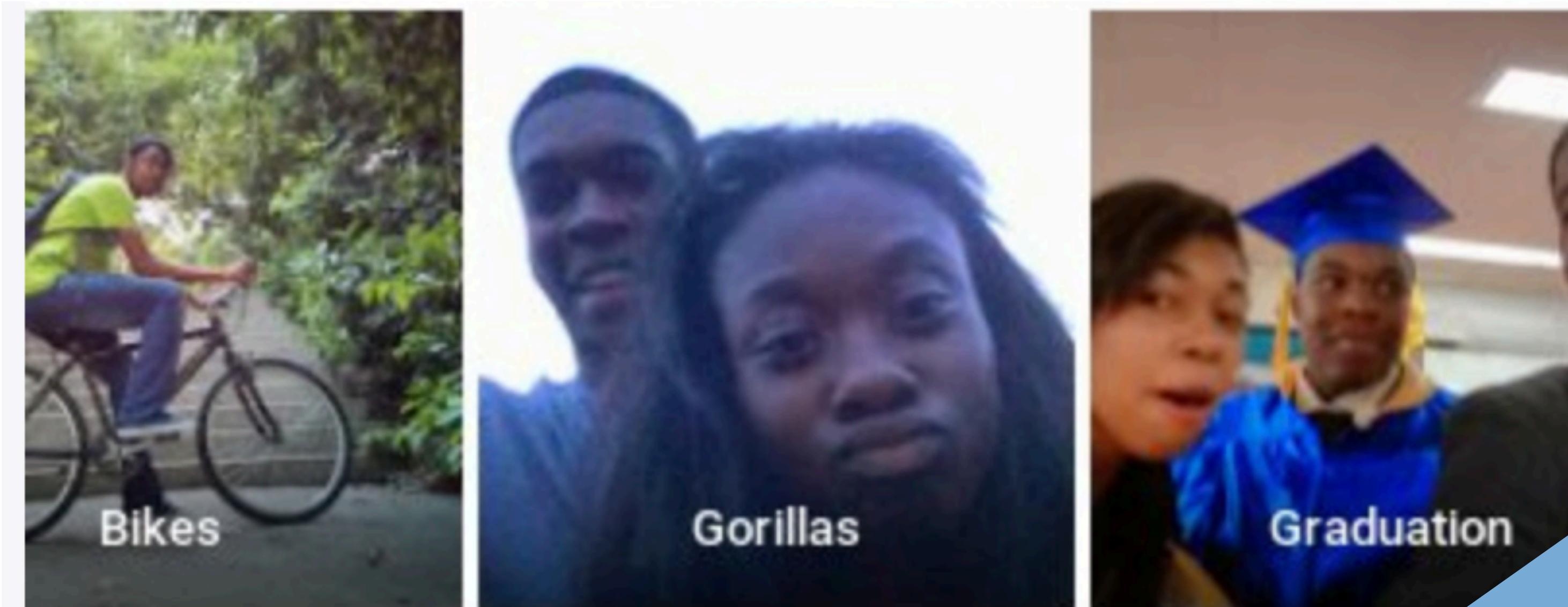


JONATHAN
HAIDT

**System 2 selected for
by social pressures**

Google says sorry for racist auto-tag in photo app

- Google Photos labelled a picture of two black people as ‘gorillas’
- Google Maps and Flickr have also suffered from race-related problems



A System 1 failure

Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

Attempt to engage millennials with artificial intelligence backfires hours after launch, with TayTweets account citing Hitler

The image shows a tweet from the official Twitter account of Microsoft's AI chatbot, Tay. The tweet, posted at 5:44 PM on March 23, 2016, reads "@ReynTheo HITLER DID NOTHING WRONG!". It has received 97 retweets and 100 likes. The tweet is displayed against a white background with standard Twitter interface elements like a profile picture, a blue verification checkmark, and a 'Following' button. A blue diagonal banner across the bottom right corner of the image contains the text "A System 1 failure".

A System 1 failure

Machine Learning? 1. **System 1** - a system of heuristics that are highly efficient, but not always correct

Data Scientists? 2. **System 2** - slow, expensive, logical, conscious reasoning

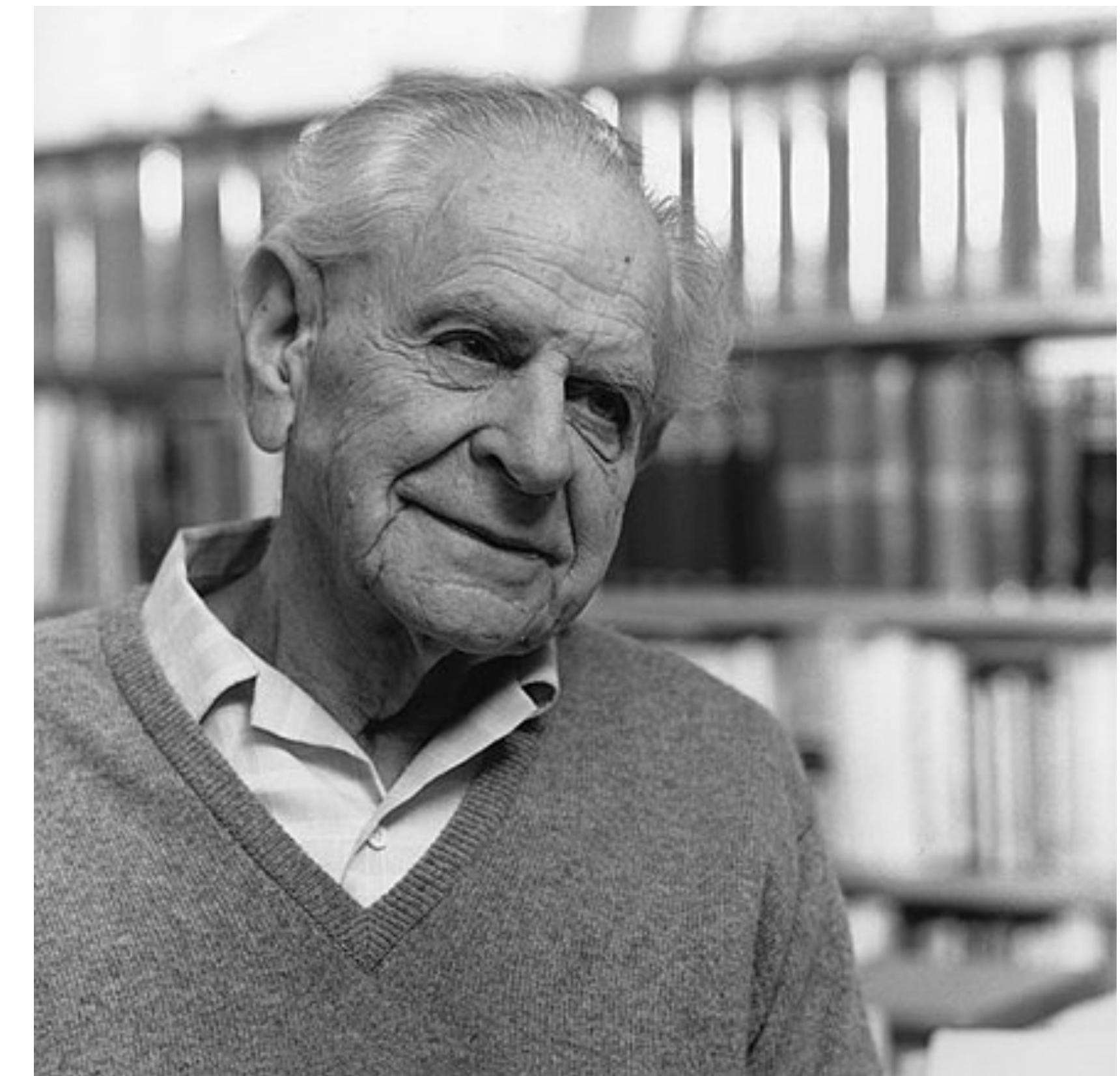


A
Larger
Reasoning
Process



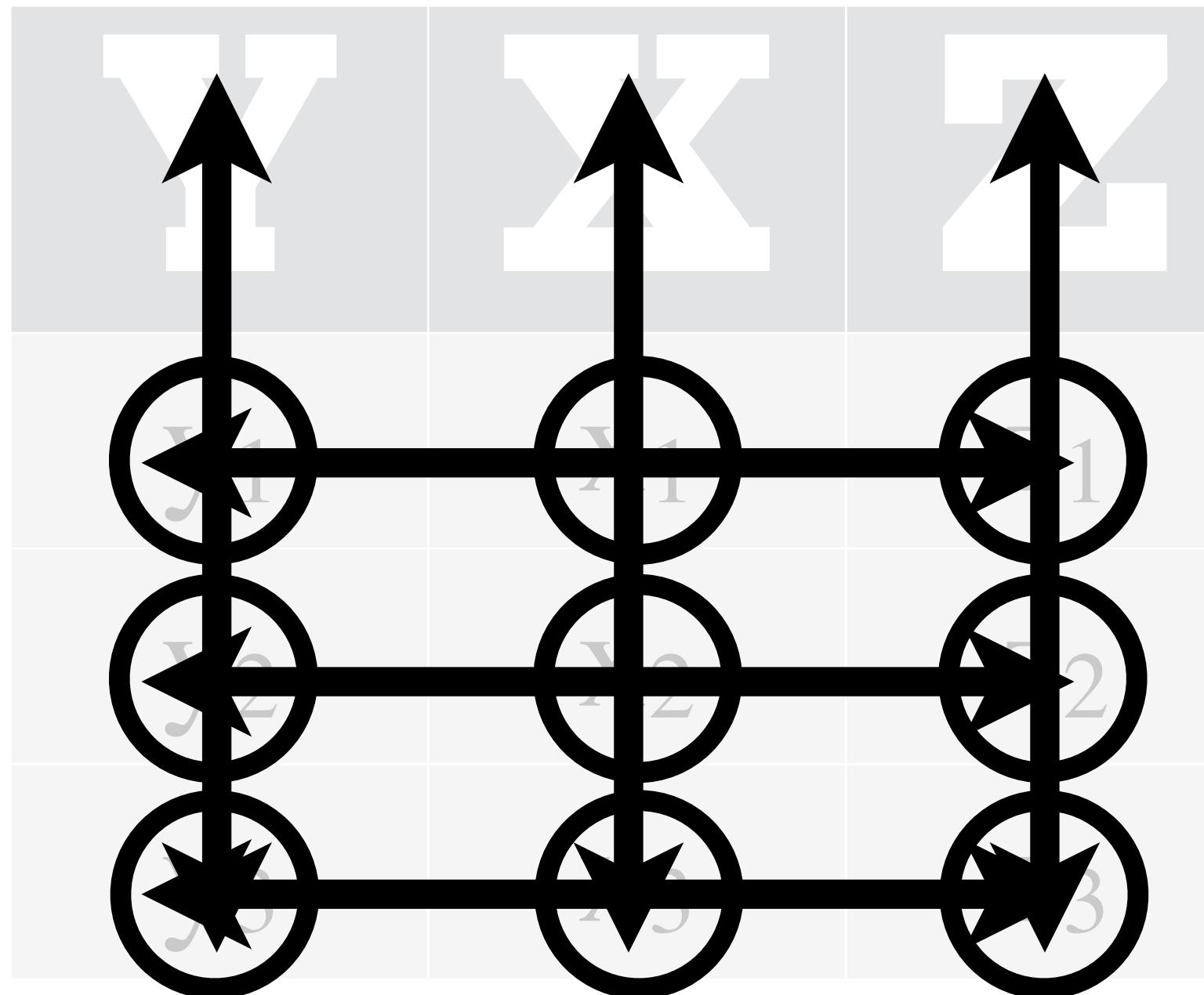
Discover

Induce hypotheses



Confirm

Deduce validity



- Values
- Variables
- Observations

X	Y	Z
3.01	0.98	3.08
2.35	0.91	2.58
5.57	1.01	5.52

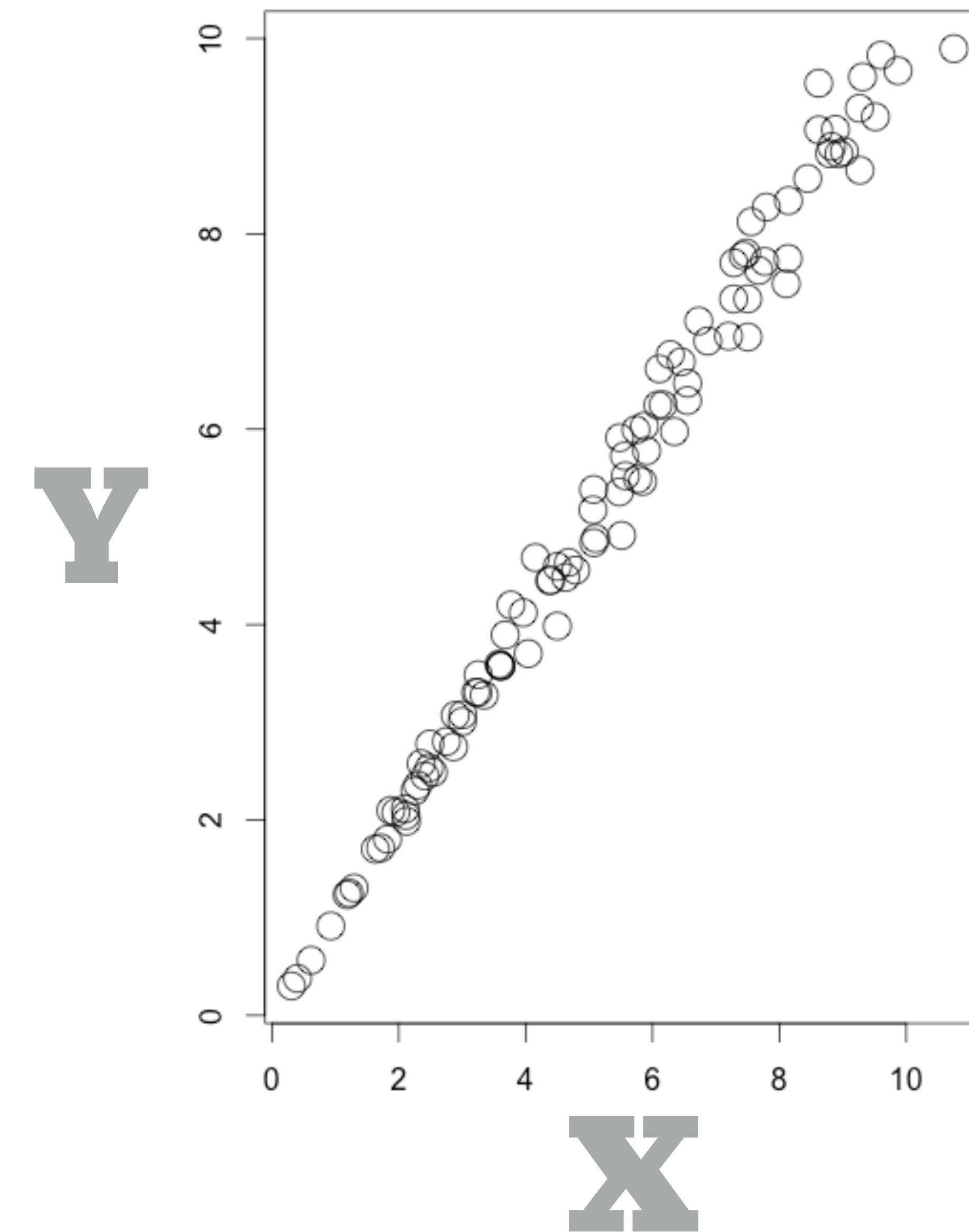
Laws appear as
patterns in data.

Y	X	Z
3.01	0.98	3.08
2.35	0.91	2.58
5.57	1.01	5.52
5.57	1.01	5.52
4.15	0.89	4.69
5.07	1.05	4.84
7.56	0.93	8.12

Laws appear as
patterns in data.

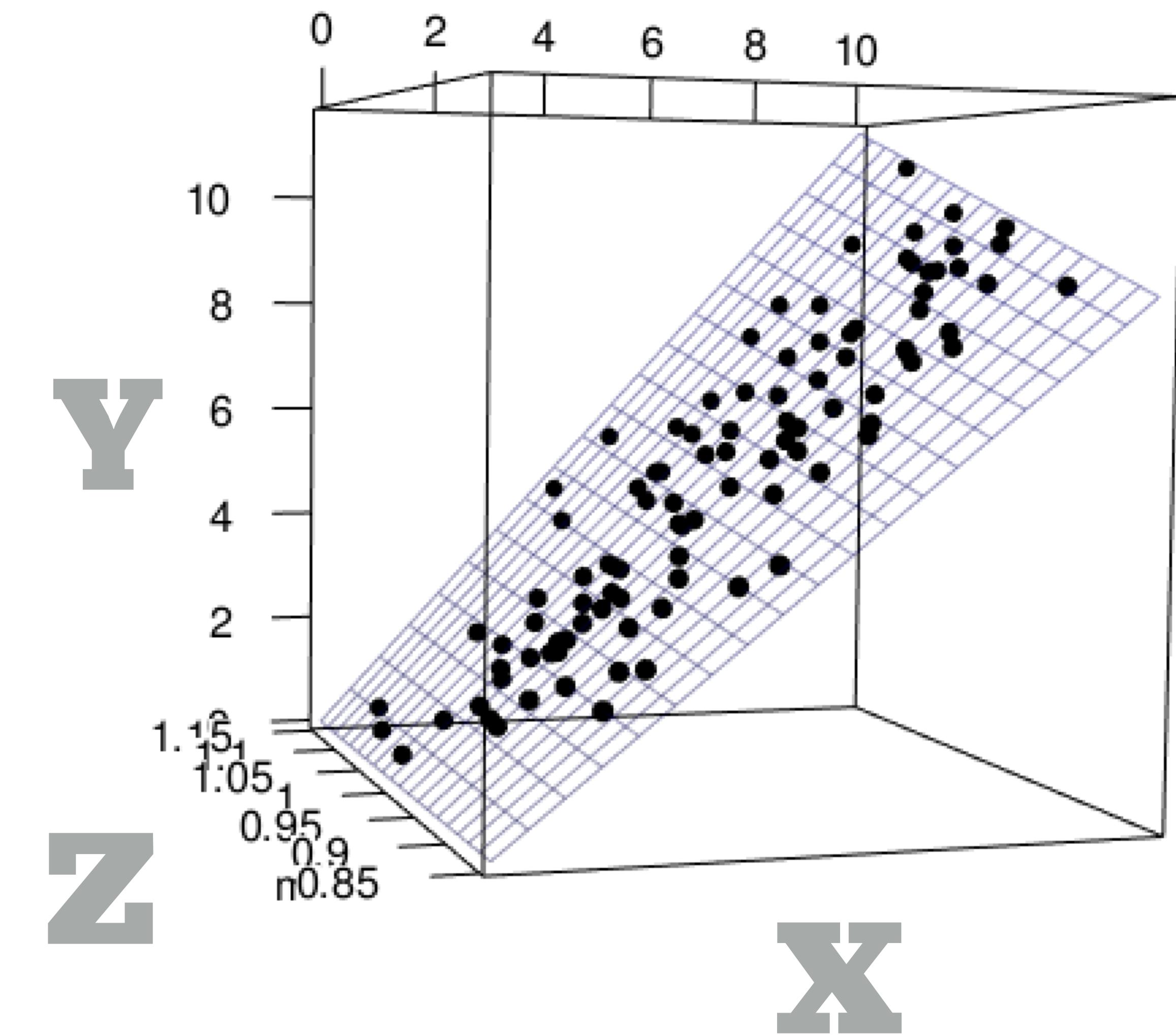
$$Y = X + \epsilon$$

Y	X	Z
0.62	1.09	0.56
1.30	0.99	1.30
1.63	0.96	1.70
1.72	1.00	1.71
1.82	1.01	1.80
1.95	0.94	2.08
2.11	1.03	2.05



$$Y = X + Z$$

Y	X	Z
0.62	1.09	0.56
1.30	0.99	1.30
1.63	0.96	1.70
1.72	1.00	1.71
1.82	1.01	1.80
1.95	0.94	2.08
2.11	1.03	2.05



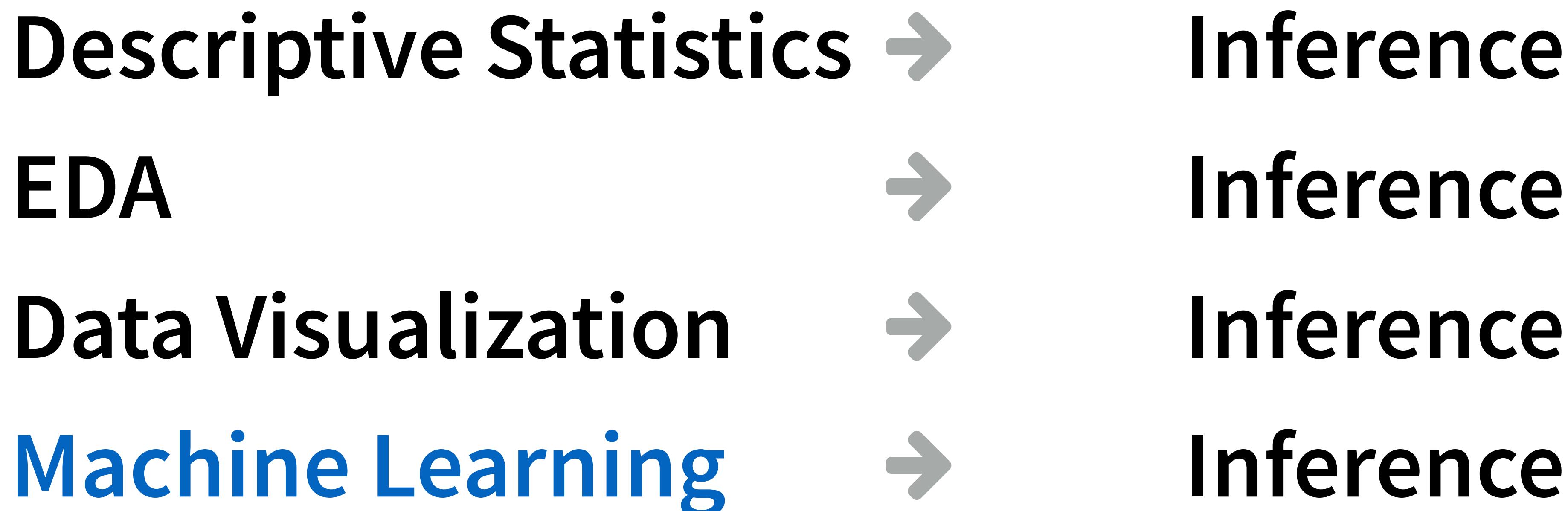
Discover Confirm



"The greatest value of a picture is
when it forces us to notice what we
never expected to see."

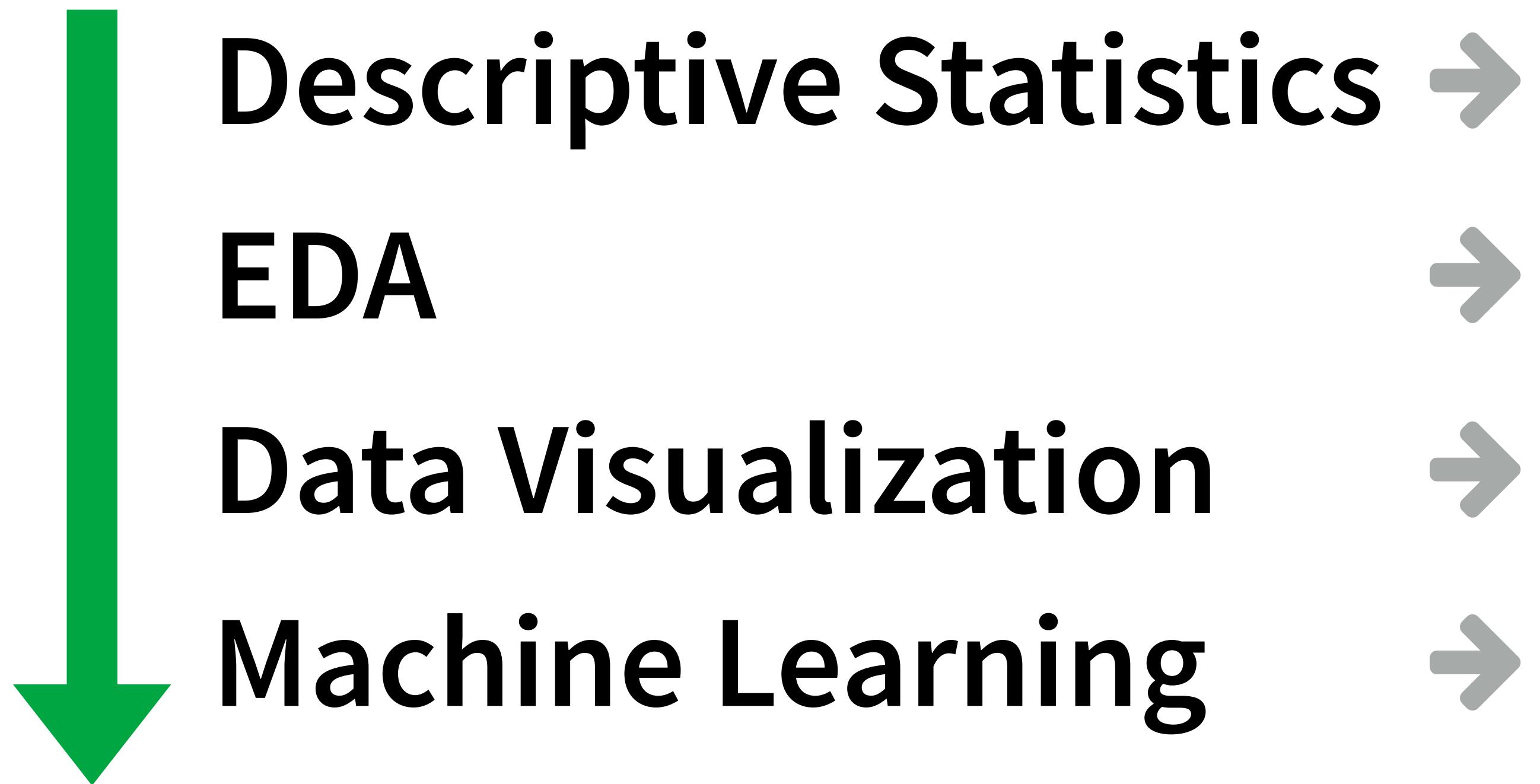
- John Tukey

Discover Confirm



Discover Confirm

Small Data,
No compute power

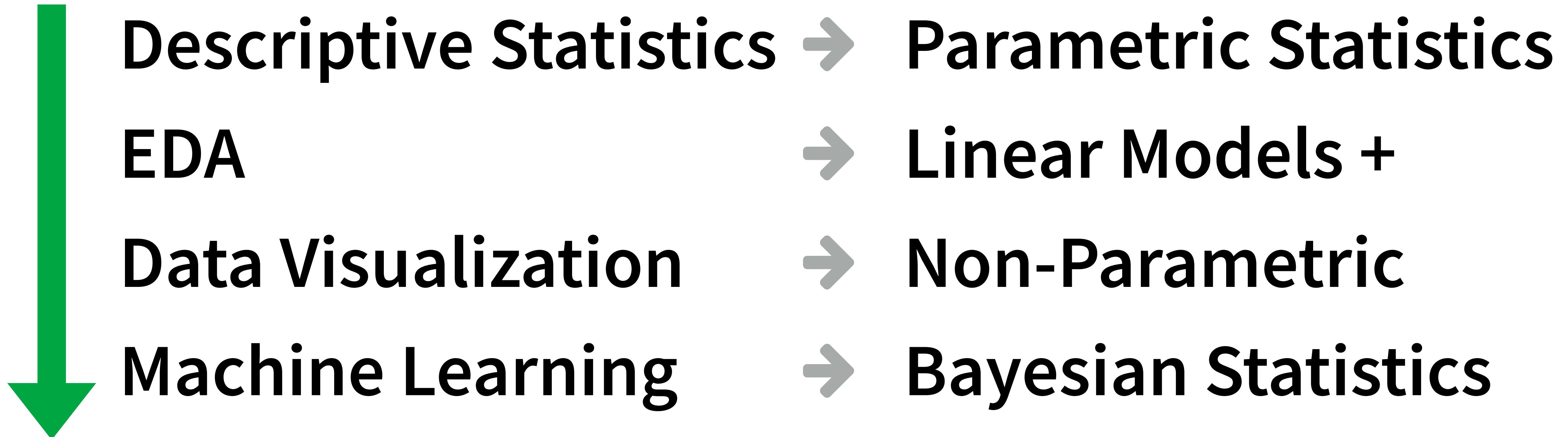


Inference
Inference
Inference
Inference

Big Data,
Ample compute power

Discover Confirm

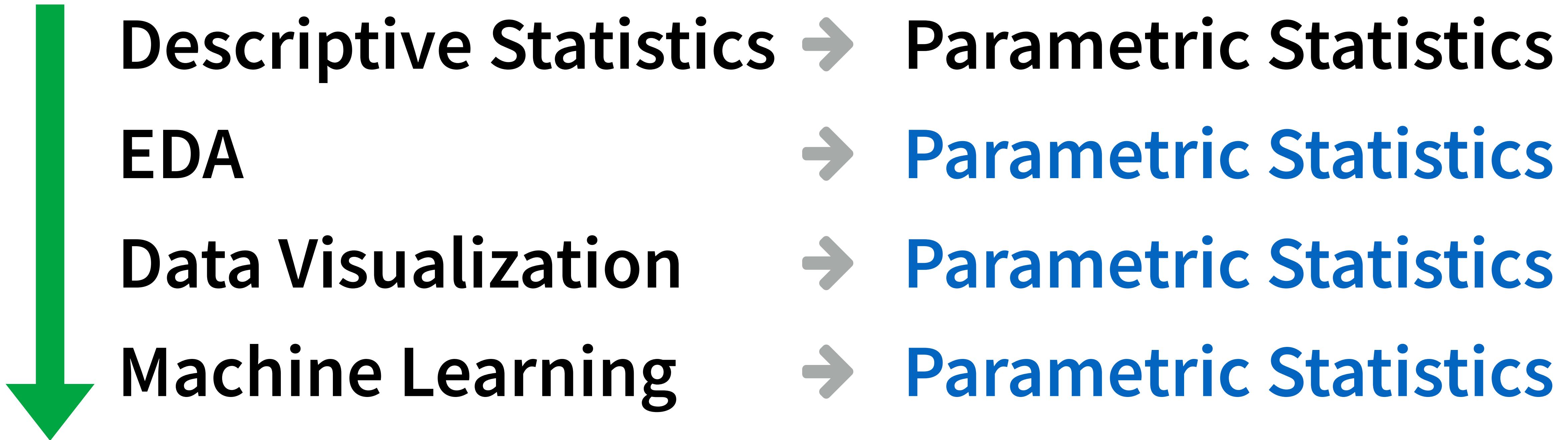
Small Data,
No compute power



Big Data,
Ample compute power

Discover Confirm

Small Data,
No compute power



Big Data,
Ample compute power

The American Statistician

Volume 69, 2015 - Issue 4: *Special Issue on
Statistics and the Undergraduate Curriculum*

Discover Confirm

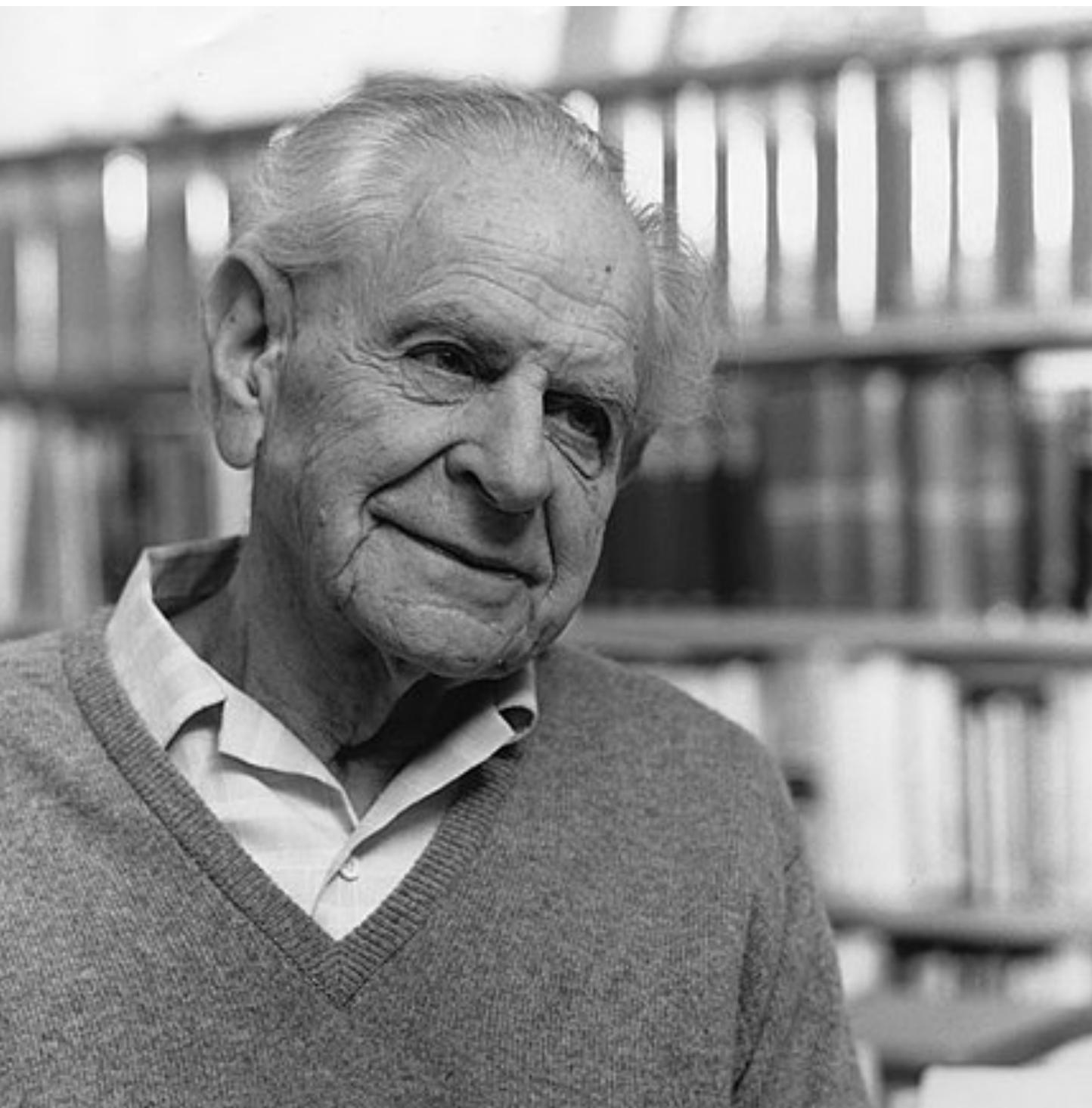
Descriptive Statistics → Parametric Statistics

EDA → Linear Models +

Data Visualization → Non-Parametric

Machine Learning → Bayesian Statistics

Teach the craft
not the tools



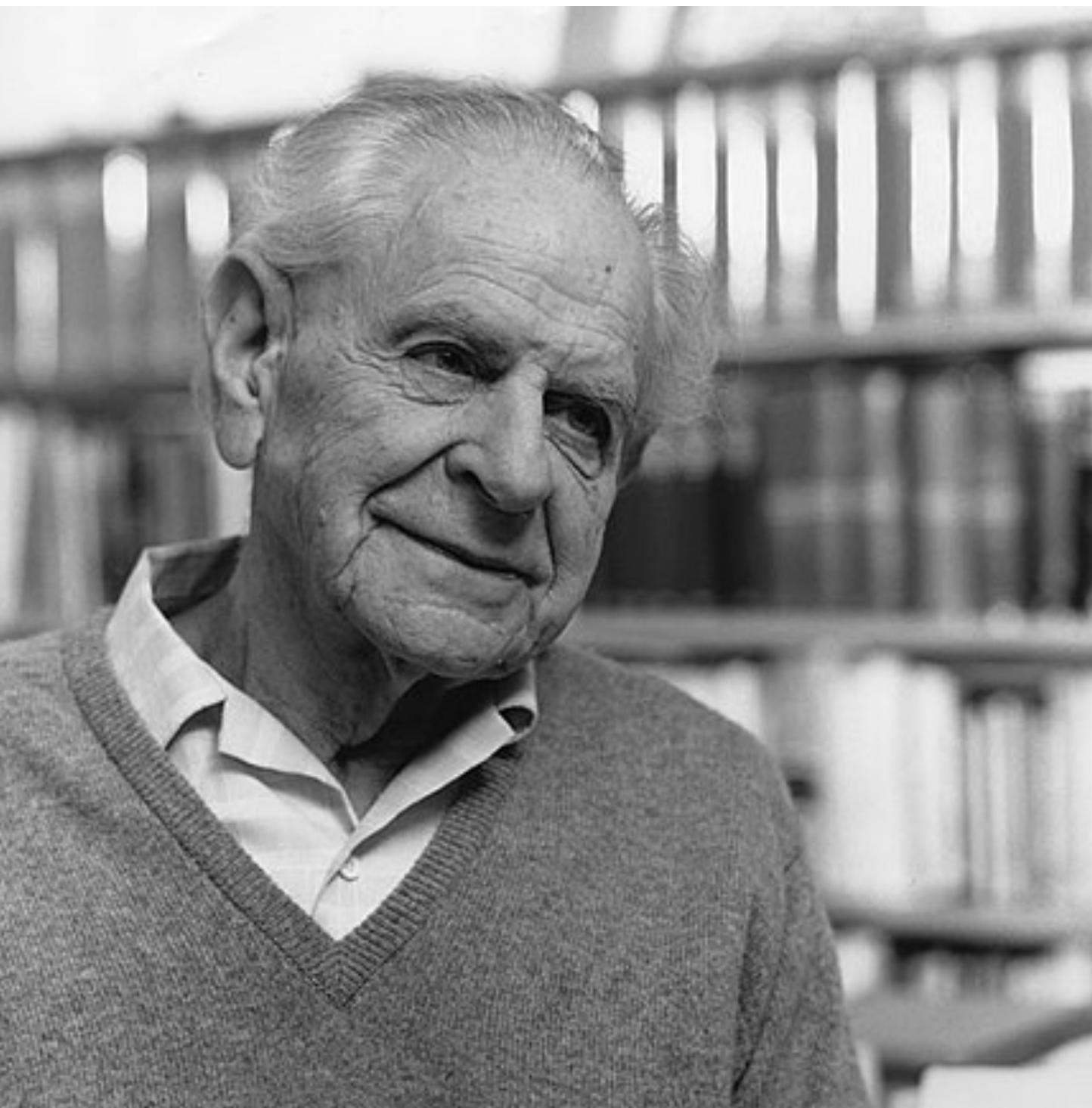
Discover Confirm

Induce hypothesis

Bound

Deduce validity

Define applicability



Discover Confirm

Induce hypothesis

Deduce validity

Infer

Define applicability

Clinical Trial

(randomized treatment, random sampling)

Discover Confirm Infer

Descriptive
Statistic

Permutation
Test

Applies to
population, qualify
with bootstrapped
confidence interval

Clinical Trial

(randomized treatment, ~~random sampling~~)

Discover Confirm Infer

Descriptive
Statistic

Permutation
Test

Does not apply to
population

Clinical Trial

(~~randomized treatment, random sampling~~)

Discover	Confirm	Infer
Descriptive Statistic	Cannot confirm	Applies to population, qualify with bootstrapped confidence interval

Machine Learning

(that yields prediction)

Discover	Confirm	Infer
Model algorithm selected through cross validation	Nothing to Confirm	Applies to events that are similar to training events

Regression Model

(that yields coefficients)

Discover Confirm Infer

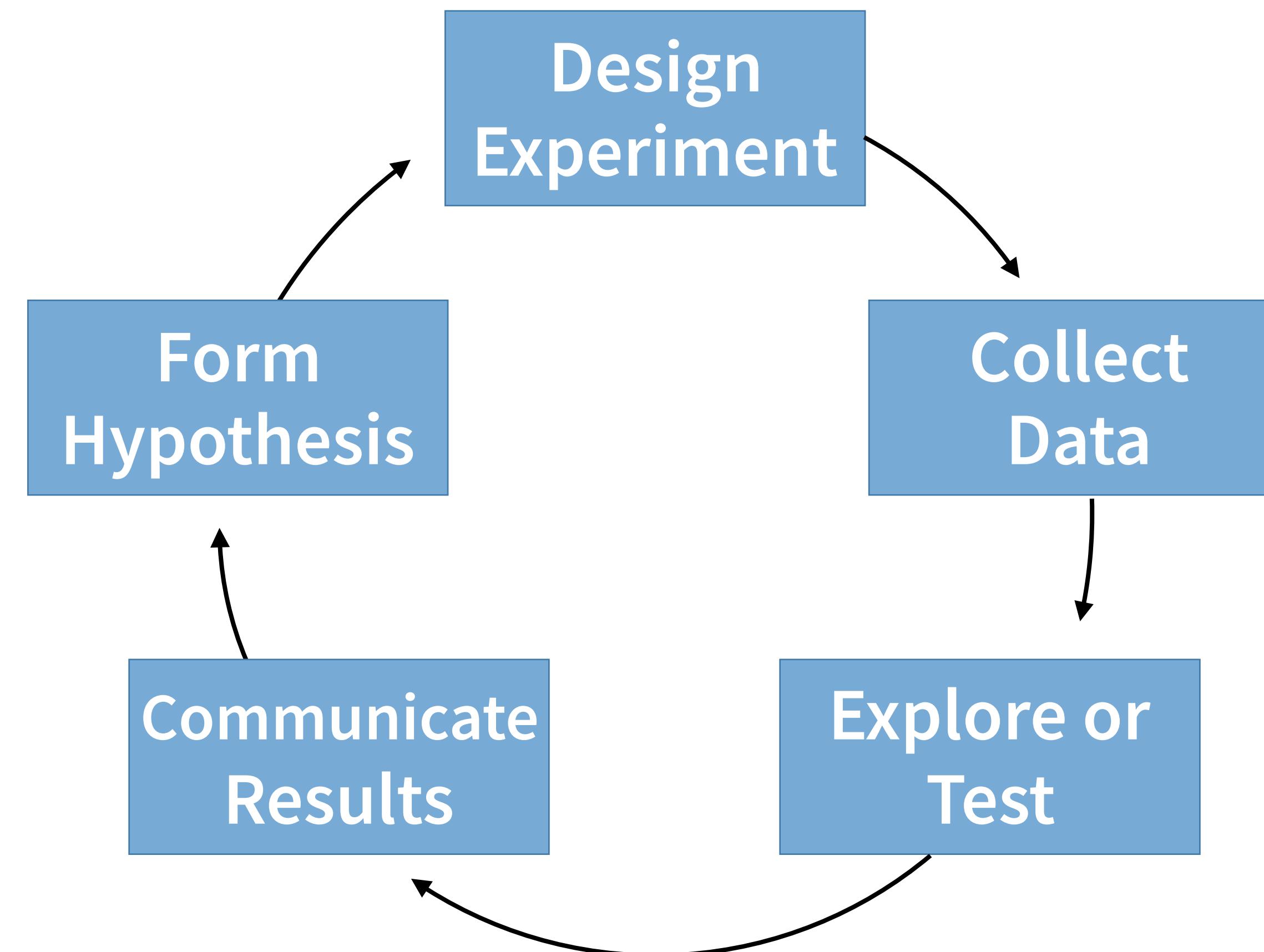
Model fit with
OLS, MLE, etc.

Permutation
test

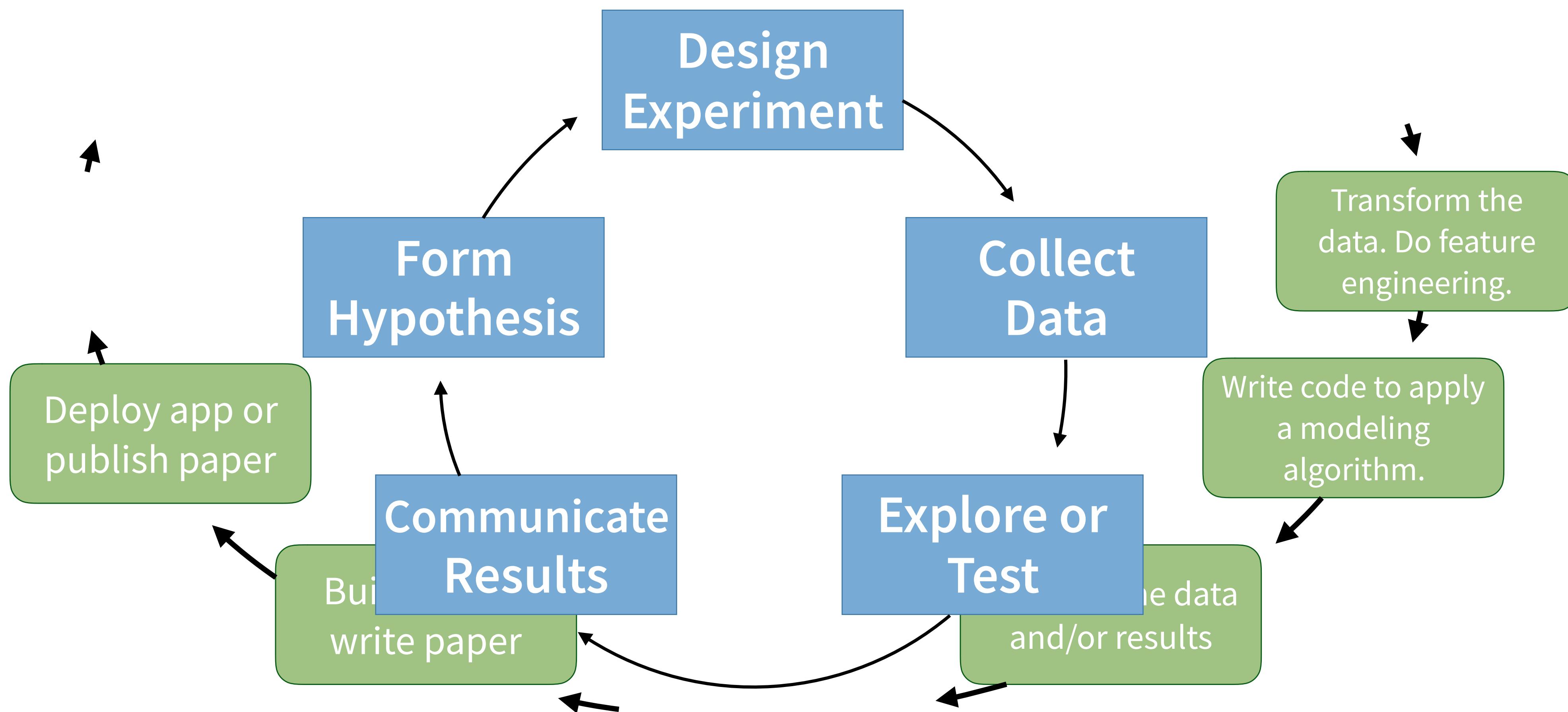
Applies to events
with similar sources
of unexplained
variation

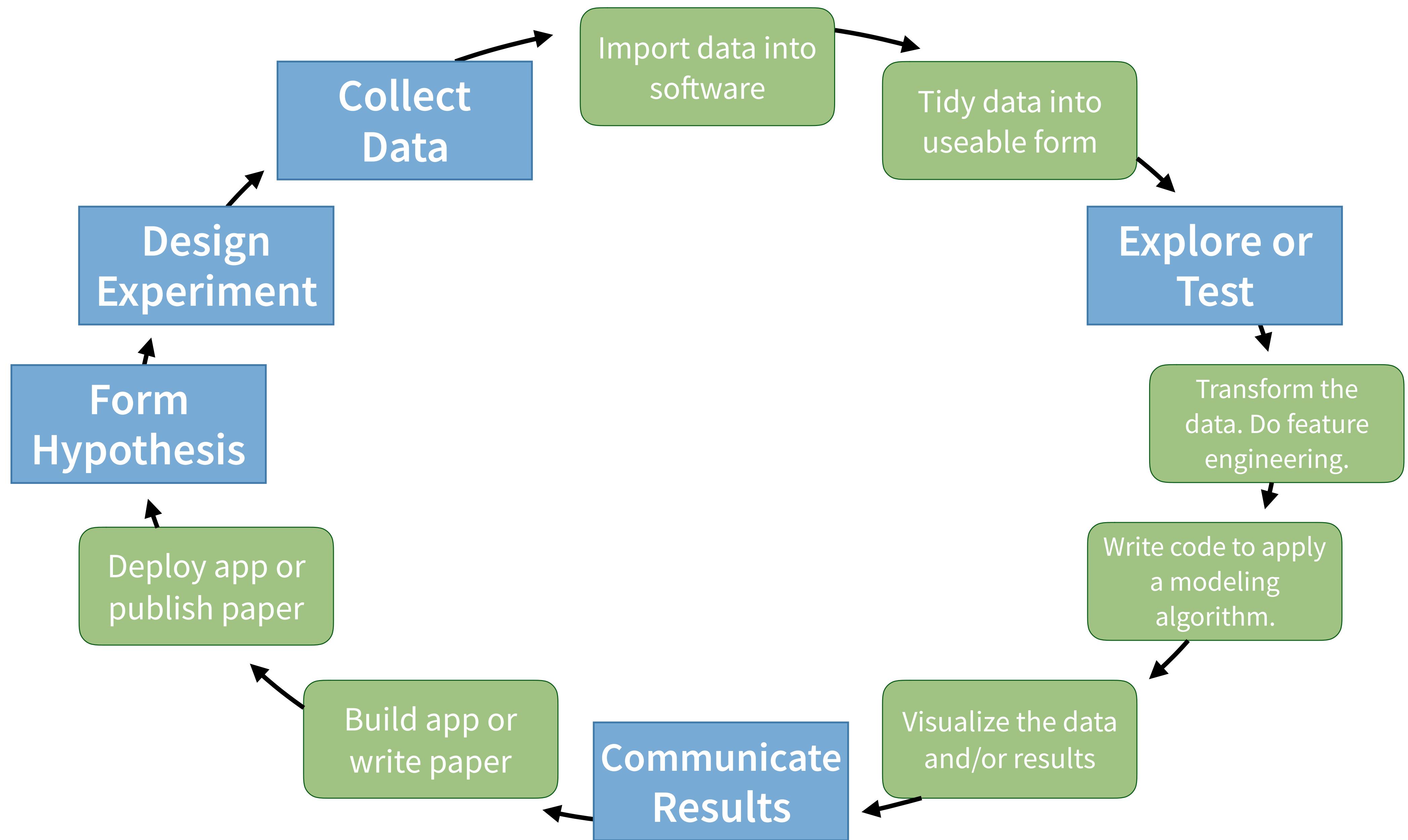
Teach the practicum

"Science with Data"



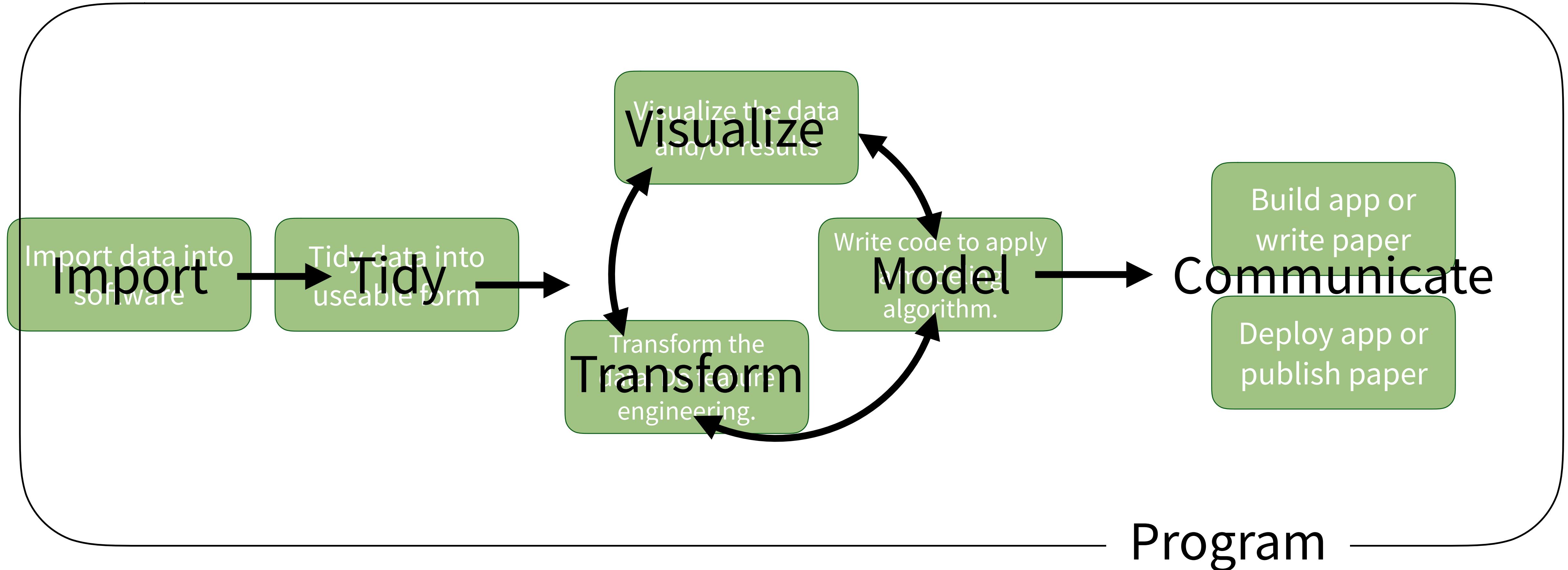
"Science with Data"!







(Applied) Data Science



tidyverse.org

The screenshot shows the homepage of tidyverse.org. At the top, there's a navigation bar with links for Packages, Articles, Learn, Help, and Contribute. Below the navigation, there's a large heading "Tidyverse". To the right of the heading, there's a brief description: "R packages for data science. The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying philosophy and common APIs." Below this text, there's a section titled "Install the complete tidyverse with:" followed by a code block: `install.packages("tidyverse")`. On the left side of the page, there's a decorative graphic consisting of several hexagonal icons representing different R packages: dplyr (orange, with a pliers icon), ggplot2 (grey, with a line plot icon), readr (blue, with a document icon), purrr (white with a cat icon), tibble (black, with a grid icon), and tidyr (orange, with a circular arrow icon).

www.rstudio.com/resources/cheatsheets

The screenshot shows the RStudio Cheat Sheets page at <https://www.rstudio.com/resources/cheatsheets/>. The page features the R Studio logo and navigation links for rstudio::conf, Products, Resources, Pricing, About Us, and Blogs. A search icon is also present. The main title is "RStudio Cheat Sheets". Below it, a text block says: "The cheat sheets below make it easy to learn about and use some of our favorite packages. From time to time, we will add new cheat sheets to the gallery. If you'd like us to drop you an email when we do, let us know by clicking the button to the right." To the right, there is a blue button labeled "SUBSCRIBE TO CHEAT SHEET UPDATES HERE". Further down, a "Deep Learning with Keras Cheat Sheet" is displayed, featuring a diagram of the Keras API flow and code snippets for training a neural network on MNIST data.

The Deep Learning with Keras Cheat Sheet includes the following sections:

- Intro:** Keras is a high-level neural networks API developed with a focus on enabling fast experimentation. It supports multiple backends, including TensorFlow, CNTK, and Theano. TensorFlow is a lower-level interface for building deep neural network architectures. The Keras package makes it easy to use TensorFlow and Keras together.
- PREDICT:** `predict()` Generates predictions from a Keras model.
`predict_proba()` or `predict_log_proba()` Generates probability or class probability predictions for the input samples.
- WORKING WITH KERAS MODELS:** `compile()` Configures a Keras model for training.
`fit()` Trains a Keras model for a fixed number of epochs (iterations).
- FIT A MODEL:** `fit(x=None, y=None, batch_size=None, epochs=10, validation_split=0.1, callbacks=[TensorBoard])` Trains a Keras model for a fixed number of epochs (iterations).
- OTHER MODEL OPERATIONS:** `summary()` Prints a summary of a Keras model.
`get_layer(index)` Returns a layer based on its index.
`pop()` Removes the last layer in a model.
`save_model(filepath, model, save_format='h5')` Save a Keras model using HDF5.
`load_model(filepath, custom_objects=None)` Load a Keras model using a custom object.
`serializable_model(model)` Create a model that's serializable.
- EVALUATE A MODEL:** `evaluate(x=None, y=None, batch_size=None, verbose=1, sample_weight=None)` Evaluate a Keras model over one batch of samples.
- TRAINING AN IMAGE RECOGNIZER ON MNIST DATA:** A detailed code snippet for training a neural network on the MNIST dataset.

Deep Learning with Keras :: CHEAT SHEET

INSTALLATION: The Keras package uses the Python keras library. You can install all the prerequisites directly from R: <https://keras.io/installation/>. Library (keras) See [here](#); `install_keras()`.

THE KERAS API IS ALREADY LOADING.

DEFINING A MODEL: `Sequential` Model
- Sequential model
- Multi-Input model

COMPILE: Optimizer - Loss - Metrics

FIT: Batch size - Epochs - Validation split

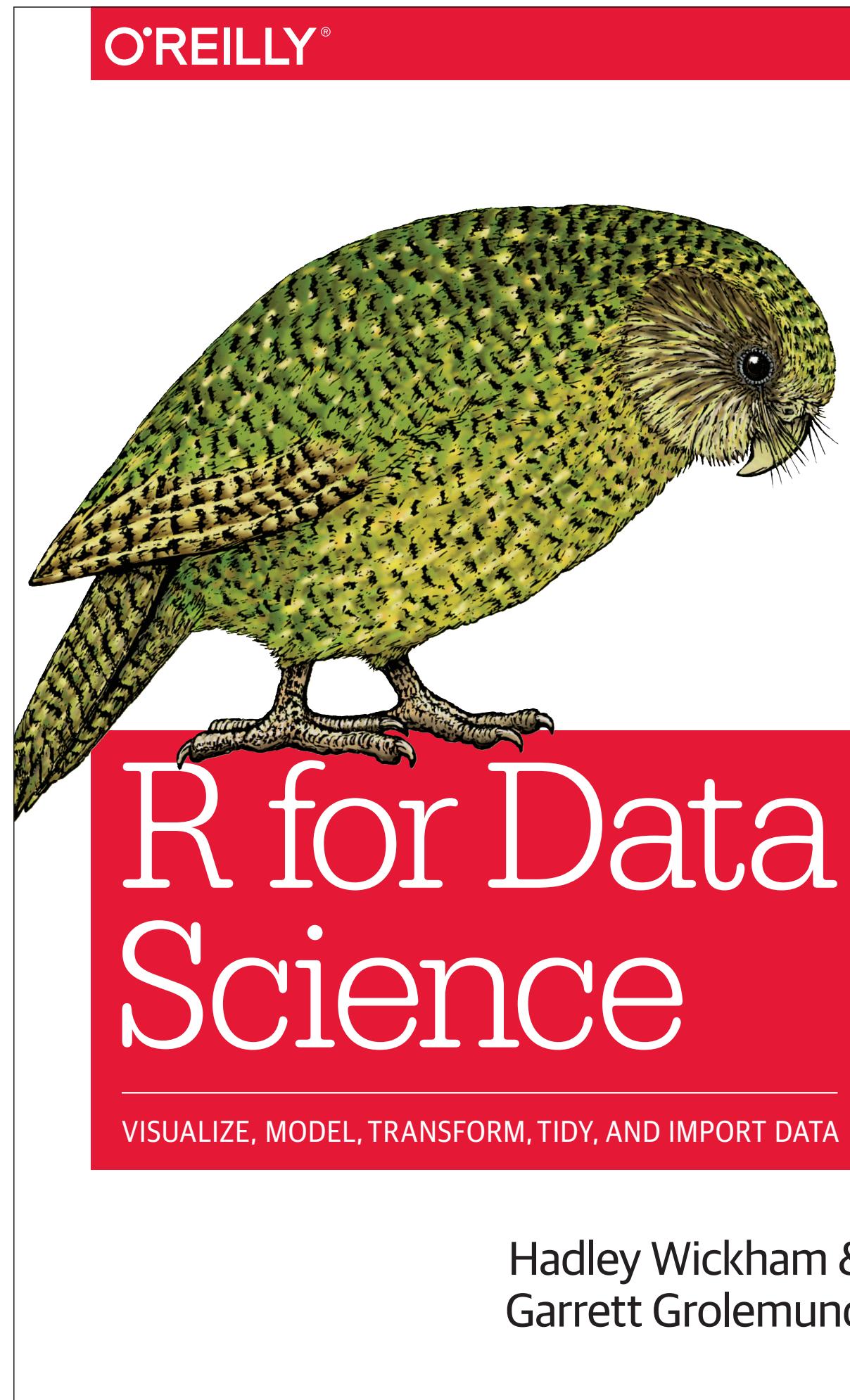
EVALUATE: Evaluate - Predict

PREDICT: classes - probability

SEE ALSO: <https://keras.io/getting-started/sequential-model-guide/>

CODE SNIPPET: `import tensorflow as tf`
`mnist = tf.keras.datasets.mnist`
`(x_train, y_train), (x_test, y_test) = mnist.load_data()`
`x_train = x_train / 255.0`
`x_test = x_test / 255.0`
`y_train = tf.keras.utils.to_categorical(y_train, 10)`
`y_test = tf.keras.utils.to_categorical(y_test, 10)`
`model = keras.Sequential([keras.layers.Flatten(input_shape=(28, 28)), keras.layers.Dense(128, activation='relu'), keras.layers.Dense(10, activation='softmax')])`
`model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])`
`model.fit(x_train, y_train, epochs=5)`
`score = model.evaluate(x_test, y_test, verbose=0)`
`print('Test accuracy:', score[1])`

DOWNLOAD



<http://r4ds.had.co.nz>

R for Data Science

Garrett

>Welcome

1 Introduction

I Explore

2 Introduction

3 Data visualisation

3.1 Introduction

3.2 First steps

3.3 Aesthetic mappings

3.4 Common problems

3.5 Facets

3.6 Geometric objects

3.7 Statistical transformations

3.8 Position adjustments

3.9 Coordinate systems

3.10 The layered grammar of graphics

4 Workflow: basics

5 Data transformation

6 Workflow: scripts

7 Exploratory Data Analysis

8 Workflow: projects

II Wrangle

9 Introduction

10 Tibbles

11 Data import

12 Tidy data

13 Relational data

3 Data visualisation

3.1 Introduction

"The simple graph has brought more information to the data analyst's mind than any other device."
— John Tukey

This chapter will teach you how to visualise your data using ggplot2. R has several systems for making graphs, but ggplot2 is one of the most elegant and most versatile. ggplot2 implements the **grammar of graphics**, a coherent system for describing and building graphs. With ggplot2, you can do more faster by learning one system and applying it in many places.

If you'd like to learn more about the theoretical underpinnings of ggplot2 before you start, I'd recommend reading "The Layered Grammar of Graphics", <http://vita.had.co.nz/papers/layers-grammar.pdf>.

3.1.1 Prerequisites

This chapter focusses on ggplot2, one of the core members of the tidyverse. To access the datasets, help pages, and functions that we will use in this chapter, load the tidyverse by running this code:

```
library(tidyverse)
#> Loading tidyverse: ggplot2
#> Loading tidyverse: tibble
#> Loading tidyverse: tidyverse
#> Loading tidyverse: readr
#> Loading tidyverse: purrr
#> Loading tidyverse: dplyr
#> Conflicts with tidy packages -----
```

RStudio Cloud Secure https://rstudio.cloud/learn/primers Garrett

Your Workspace Projects Learn Garrett Grolemund

Guide Primers DataCamp Courses Cheat Sheets

R Studio Primers

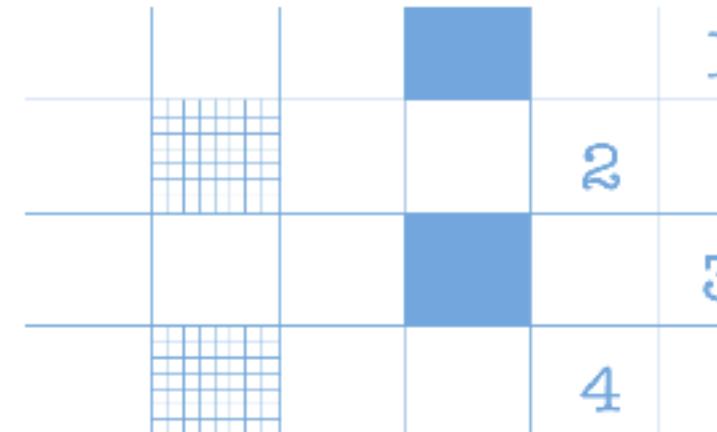
Learn data science basics with the interactive tutorials below.

The Basics



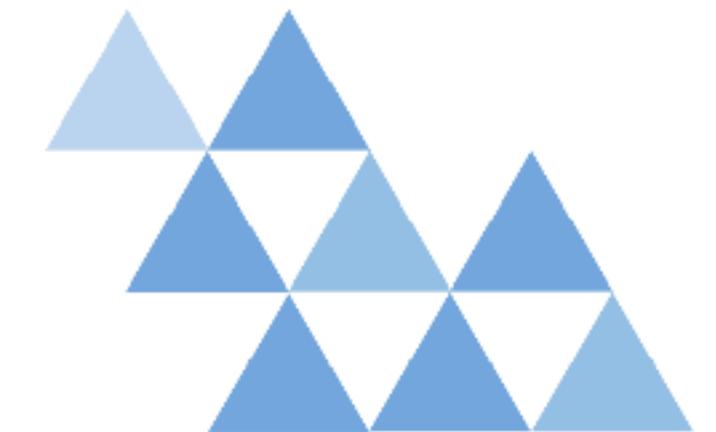
Start here to learn the skills that you will rely on in every analysis (and every primer that follows): how to inspect, visualize, subset, and transform your data, as well as how to run code.

Work with Data



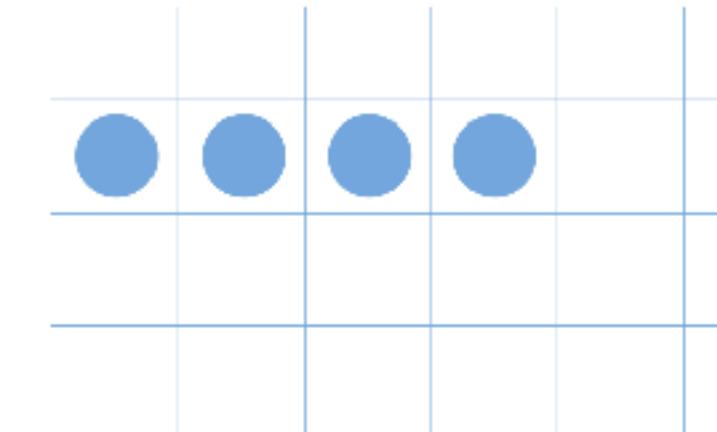
Learn the most important data handling skills in R: how to extract values from a table, subset tables, calculate summary statistics, and derive new variables.

Visualize Data



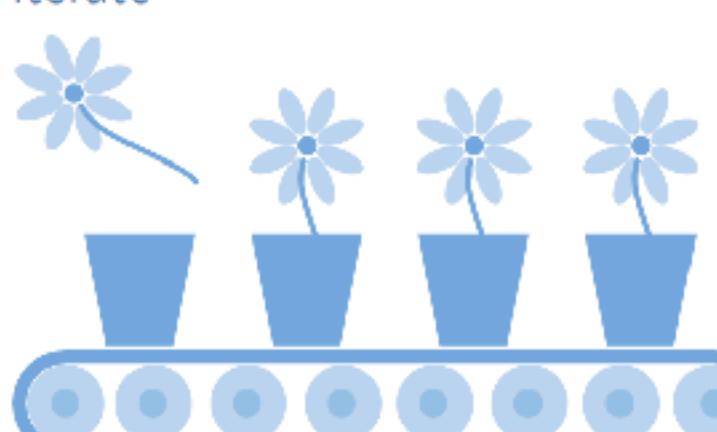
Learn how to use ggplot2 to make any type of plot with your data. Then learn the best ways to visualize patterns within values and relationships between variables.

Tidy Your Data



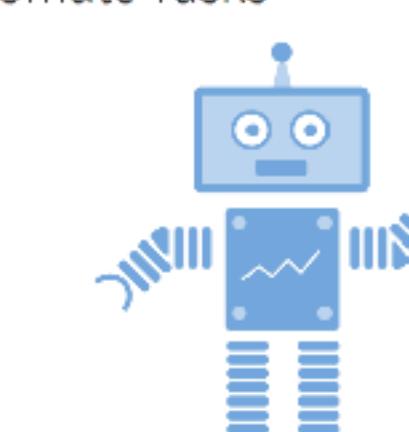
Unlock the tidyverse by learning how to make and use tidy data, the data format designed for R.

Iterate



Master a core programming paradigm with the purrr package: for each ___ do ___.

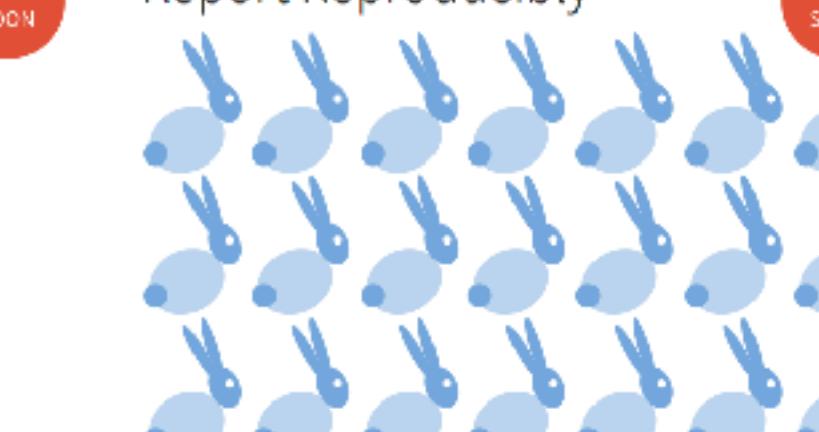
Automate Tasks



COMING SOON

Functions are the key to programming in R. This primer will teach you how to write and use them.

Report Reproducibly



COMING SOON

Learn to report, reproduce, and parameterize your work with the best authoring format for Data Science: R Markdown.

Build Interactive Web Apps



COMING SOON

Say hello to Shiny, R's package for building interactive web apps. Learn to turn your analyses into elegant tools to share with others.

**Use code,
don't write code**



Quiz

Are your students attempting to become
Data Scientists or Computer Scientists?

1. Basics of RStudio
2. Data Structures
3. Subsetting
4. For Loops

5. Writing Functions
6. Writing Data
7. Data Manipulation
8. Writing Reports

Data Science or
Computer Science?

1. Basics of RStudio

2. Data Structures

3. Subsetting

4. For Loops

5. Writing Functions

6. Writing Data

7. Data Manipulation

8. Writing Reports

Data Science or
Computer Science?

1. Visualize Data

2. Transform Data

3. Tidy Data

4. Import Data

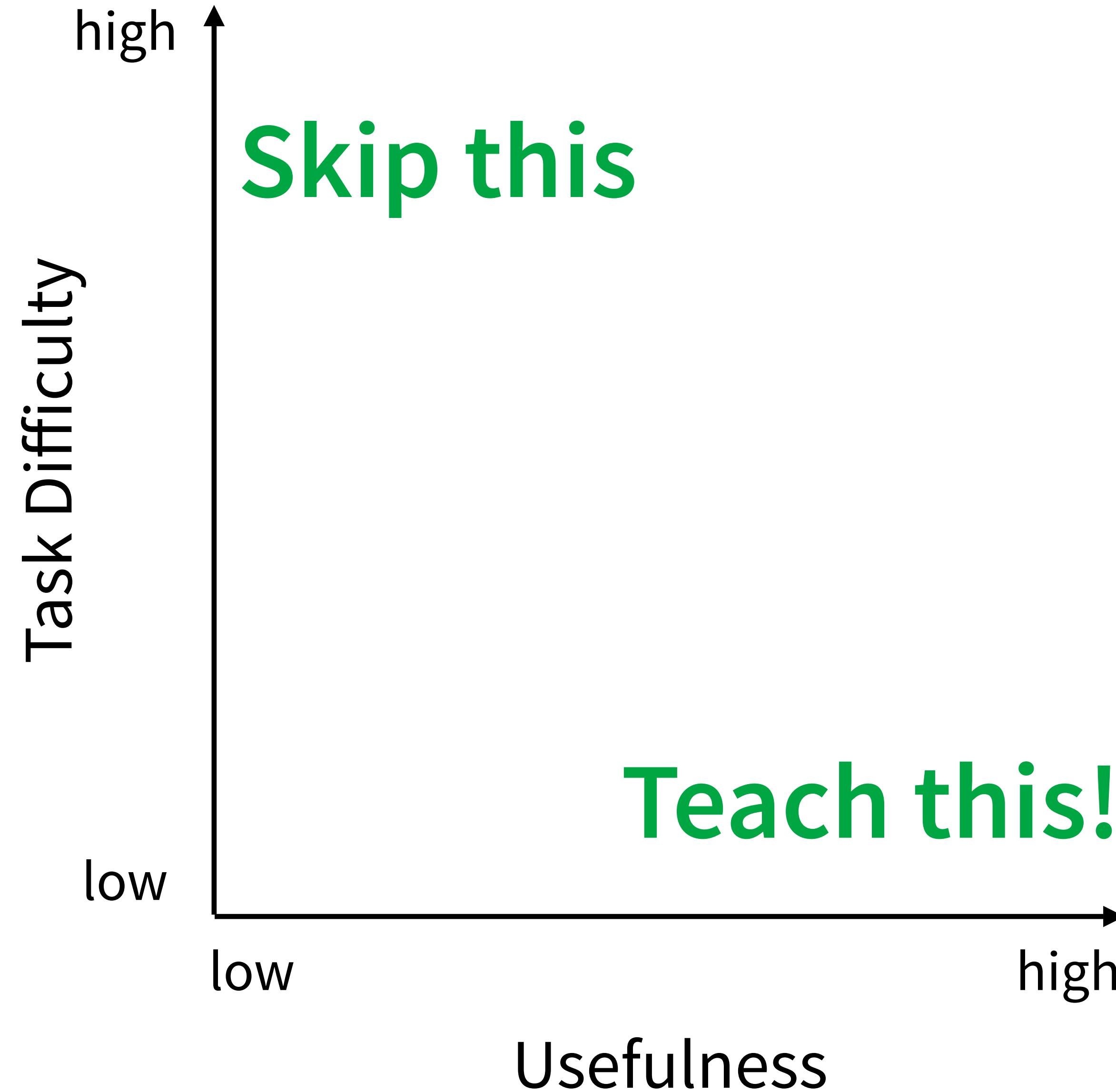
5. Types of Data

6. Iteration

7. Models

8. List Columns

Data Science or
Computer Science?



Quiz

Confer with your group.

What relationship do you expect to see between engine size (displ) and mileage (hwy)?

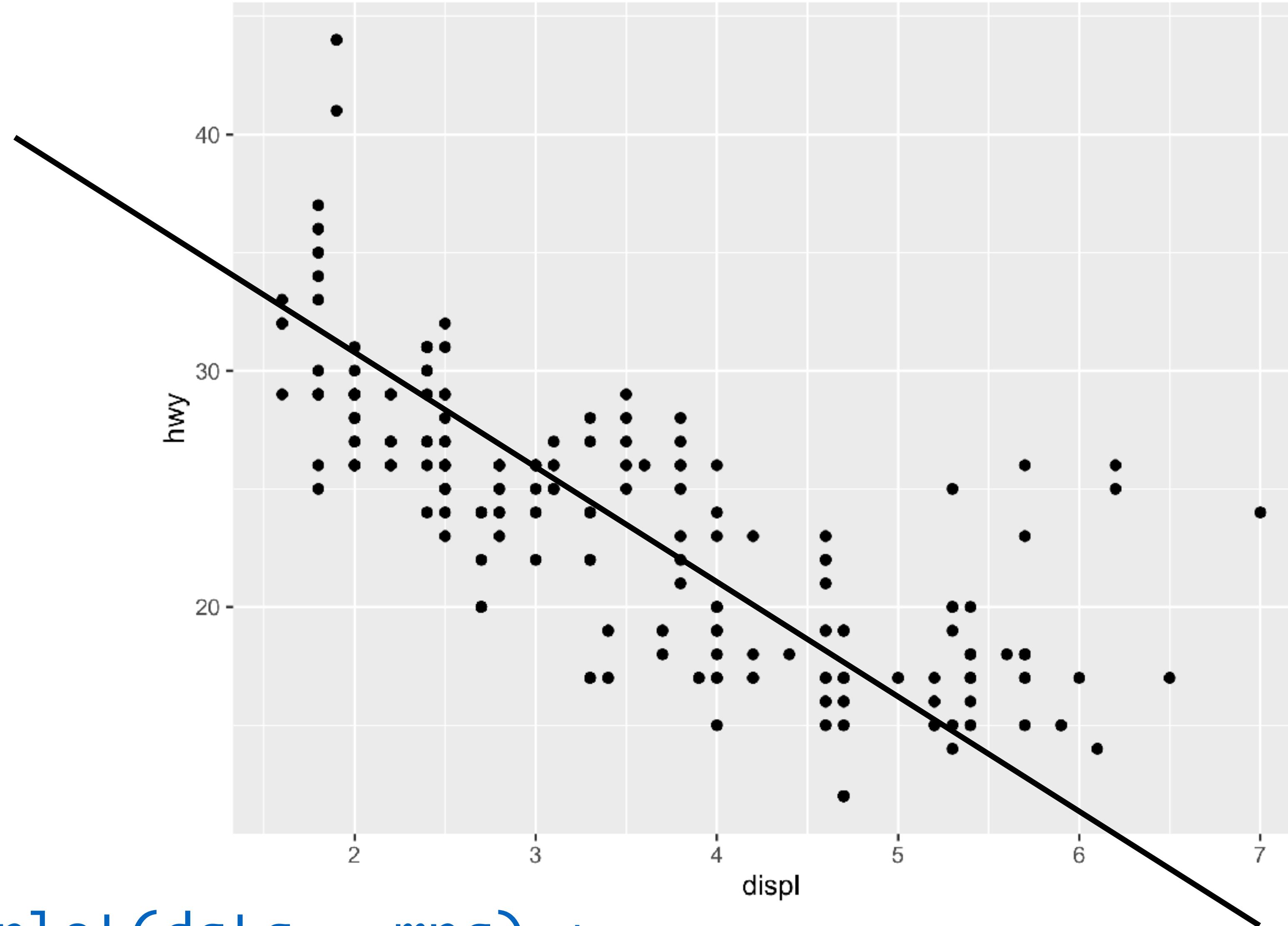
No peeking ahead!

Your Turn 1

Run this code in your notebook to make a graph.

Pay strict attention to spelling, capitalization, and parentheses!

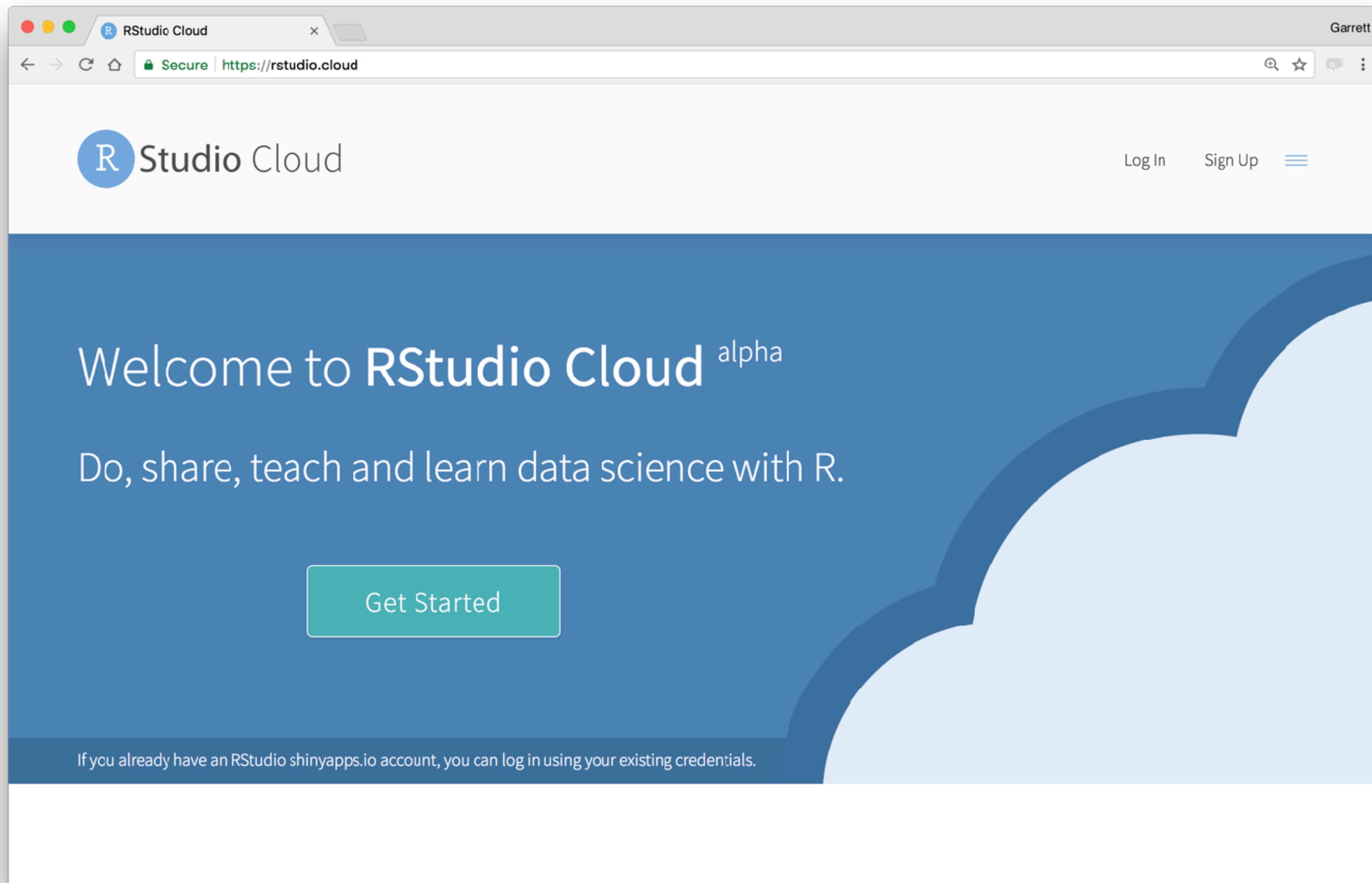
```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy))
```



```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy))
```



<https://rstudio.cloud>





**Messy data,
open questions**



Reproducibility

www.bit.ly/reprodu

CC by **Ignasi Bartomeus** (@ibartomeus) &
Francisco Rodríguez-Sánchez (@frod_san)

rmarkdown.rstudio.com/lesson-1.html

A screenshot of a web browser window displaying the R Markdown introduction page at <https://rmarkdown.rstudio.com/lesson-1.html>. The browser is an Apple Safari, indicated by the interface and the URL bar showing "Secure". The title bar says "Introduction". The main content area has a header "Introduction" and a sub-section "Overview". It explains that R Markdown provides an authoring framework for data science, allowing users to save and execute code or generate high-quality reports. Below this, there is a video player showing a slide with the text "What is R Markdown?". To the left of the main content is a sidebar with a blue header "Introduction" and a list of links: "How It Works", "Code Chunks", "Inline Code", "Code Languages", "Parameters", "Tables", "Markdown Basics", "Output Formats", "Notebooks", "Slide Presentations", "Dashboards", "Websites", "Interactive Documents", and "Cheatsheets". The "Introduction" link in the sidebar is also highlighted in blue.

Introduction

Overview

R Markdown provides an authoring framework for data science. You can use a single R Markdown file to both

- save and execute code
- generate high quality reports that can be shared with an audience

R Markdown documents are fully reproducible and support dozens of static and dynamic output formats. This 1-minute video provides a quick tour of what's possible with R Markdown:

What is
R Markdown?

A photograph showing two pairs of feet from a top-down perspective, standing on a dark, textured asphalt surface. The top pair is wearing black leather dress shoes and black trousers. The bottom pair is wearing brown leather loafers and light-colored trousers. The word "Collaboration" is overlaid in large, white, sans-serif font across the center of the image.

Collaboration

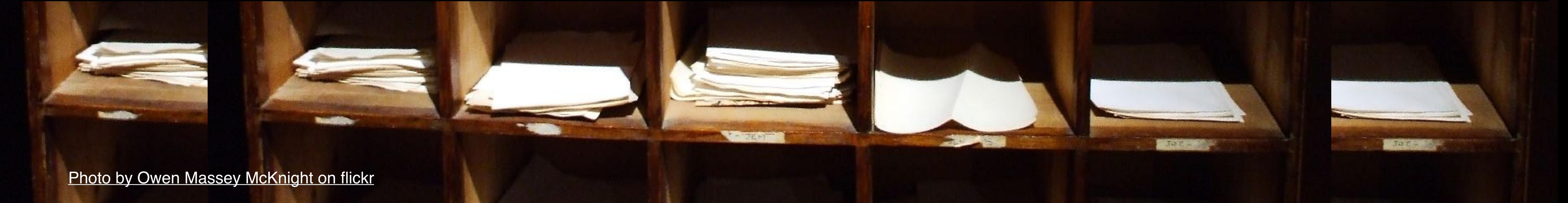
A large, dark wood shelving unit, likely a library or archive storage system, filled with numerous stacks of papers and documents. The shelves are organized into a grid pattern with multiple rows and columns. The papers vary in thickness and color, with some appearing aged and yellowed. The lighting is somewhat dim, creating a scholarly or historical atmosphere.

Github



Github

rundel.github.io/ghclass/articles/ghclass.html



Thank You

