Teaching Data Science

Garrett Grolemund
Master Instructor, RStudio
May 2018

Photo by Edwin Andrade on Unsplash

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

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

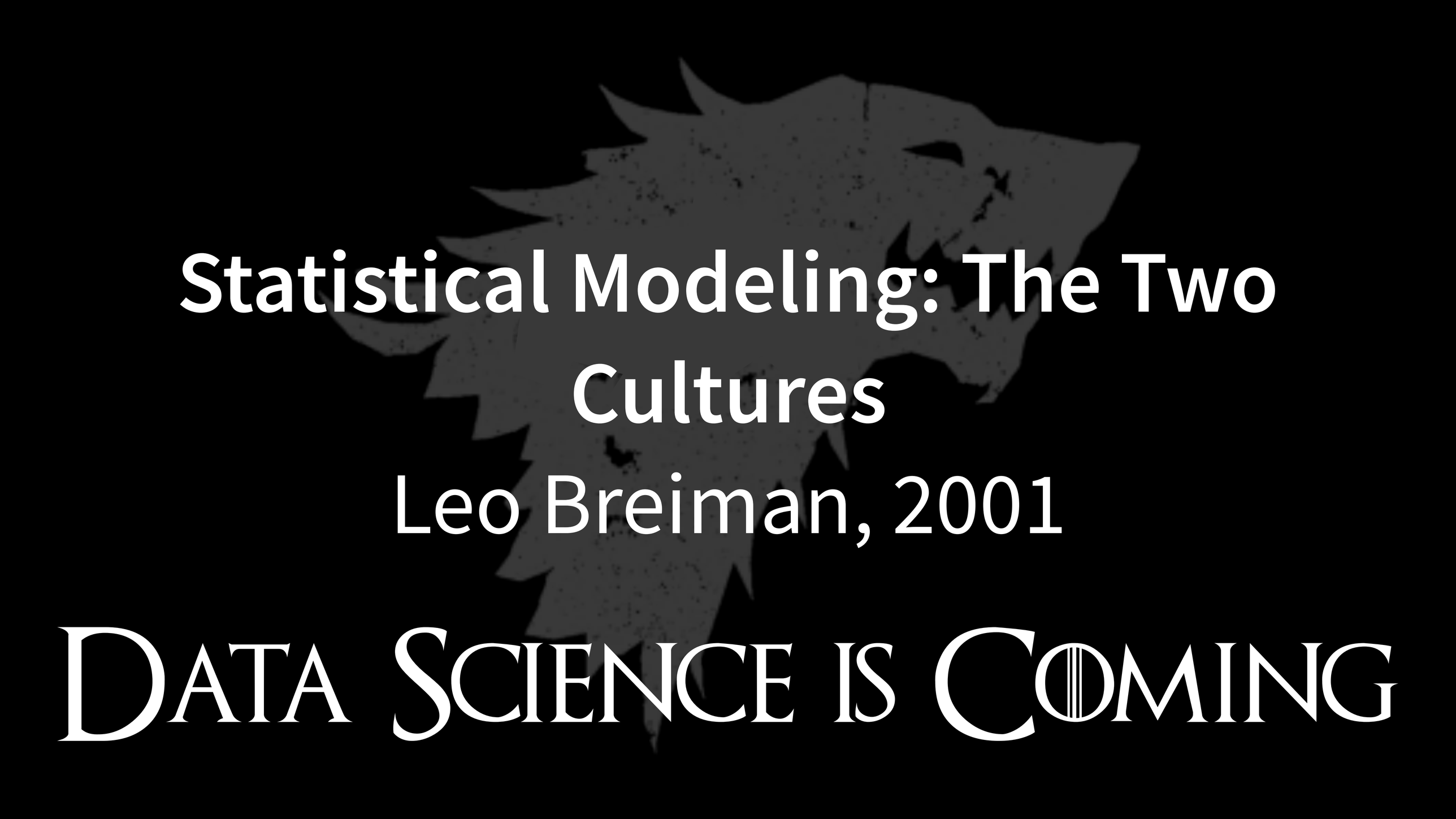
Data Analysis with R

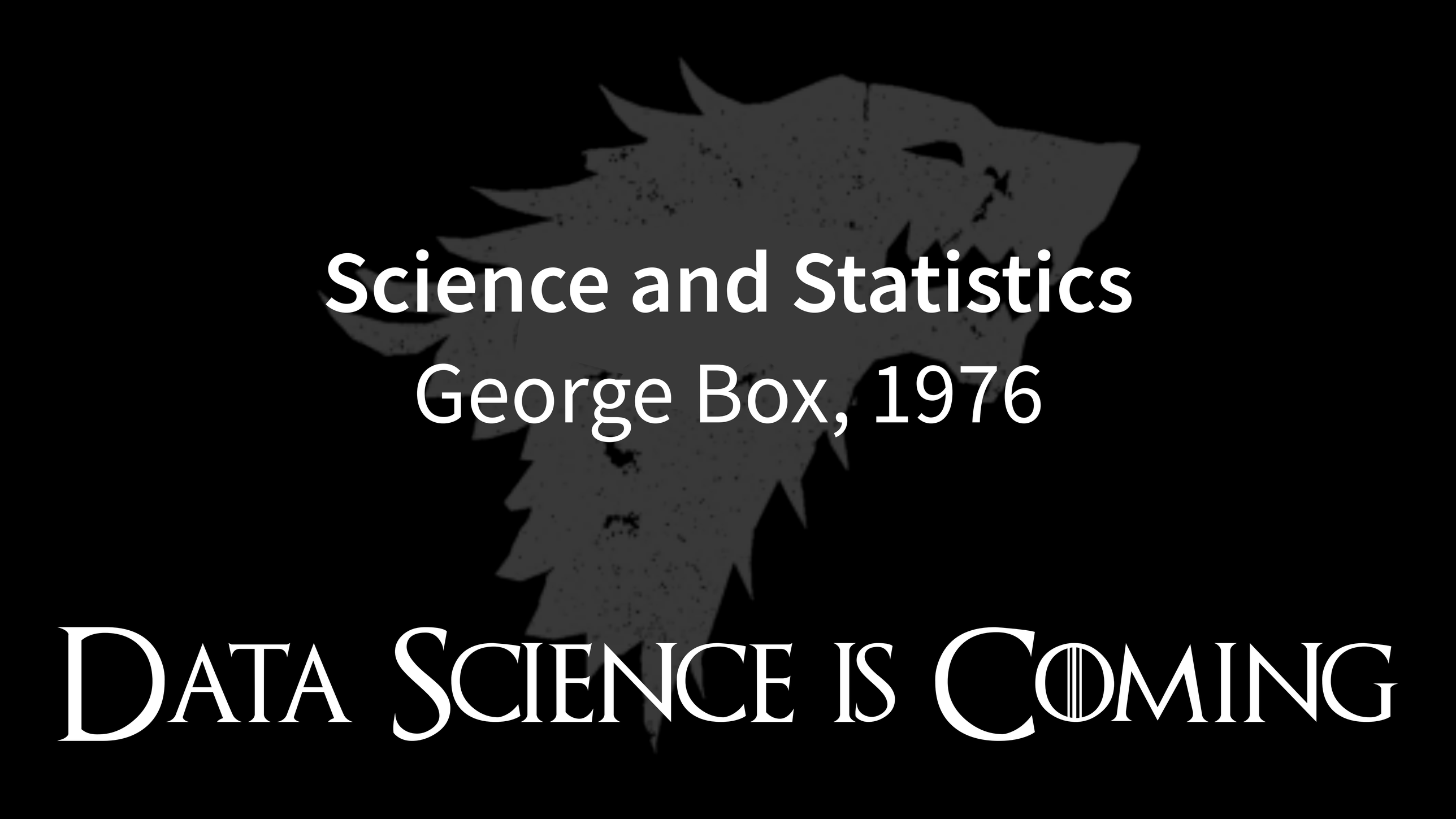
O'REILLY®

Garrett Grolemund

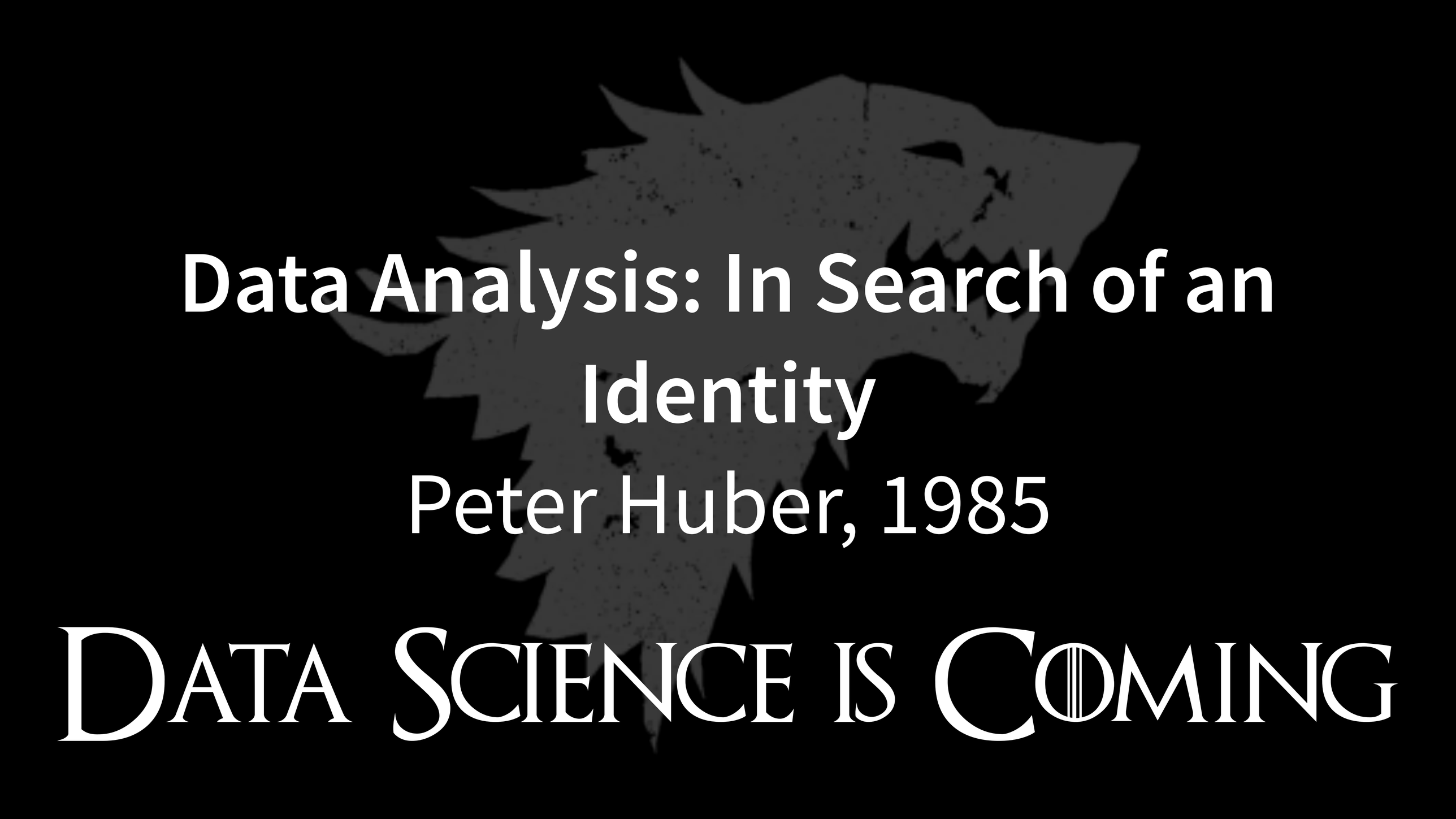

R
Studio®

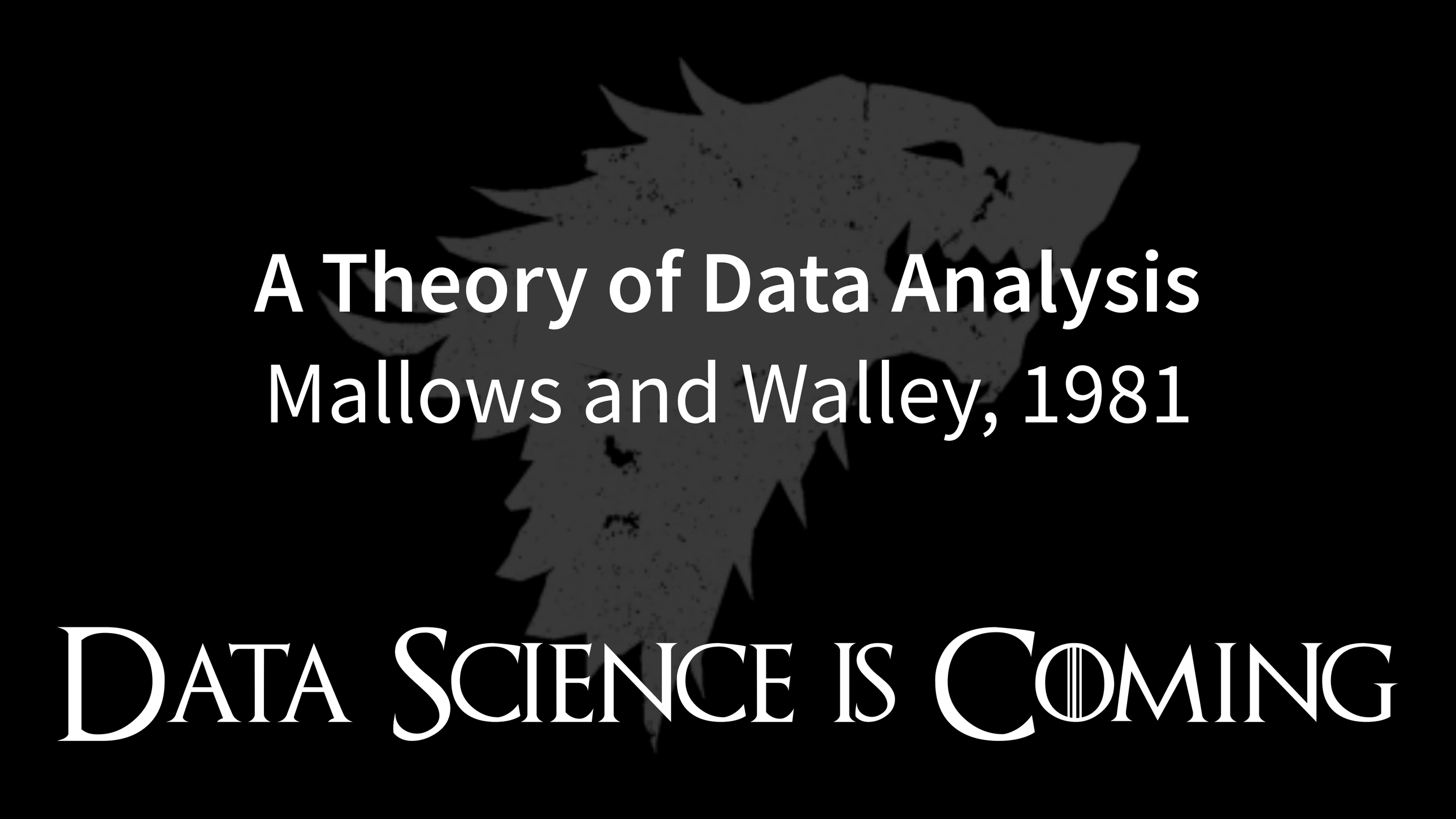
DAWN SUCCESS IS COMING


Statistical Modeling: The Two Cultures
Leo Breiman, 2001
DATA SCIENCE IS COMING

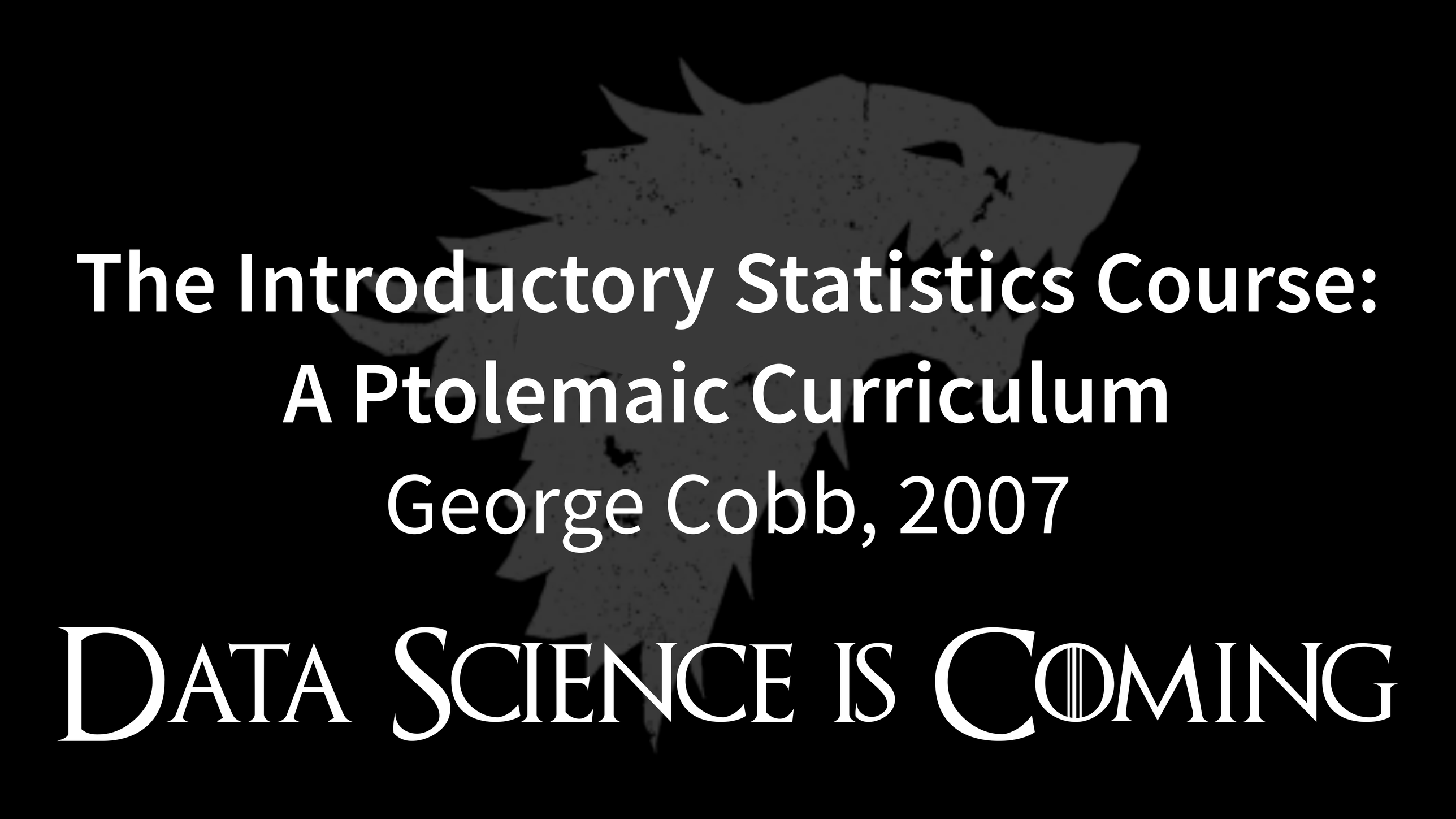

Science and Statistics
George Box, 1976
DATA SCIENCE IS COMING


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

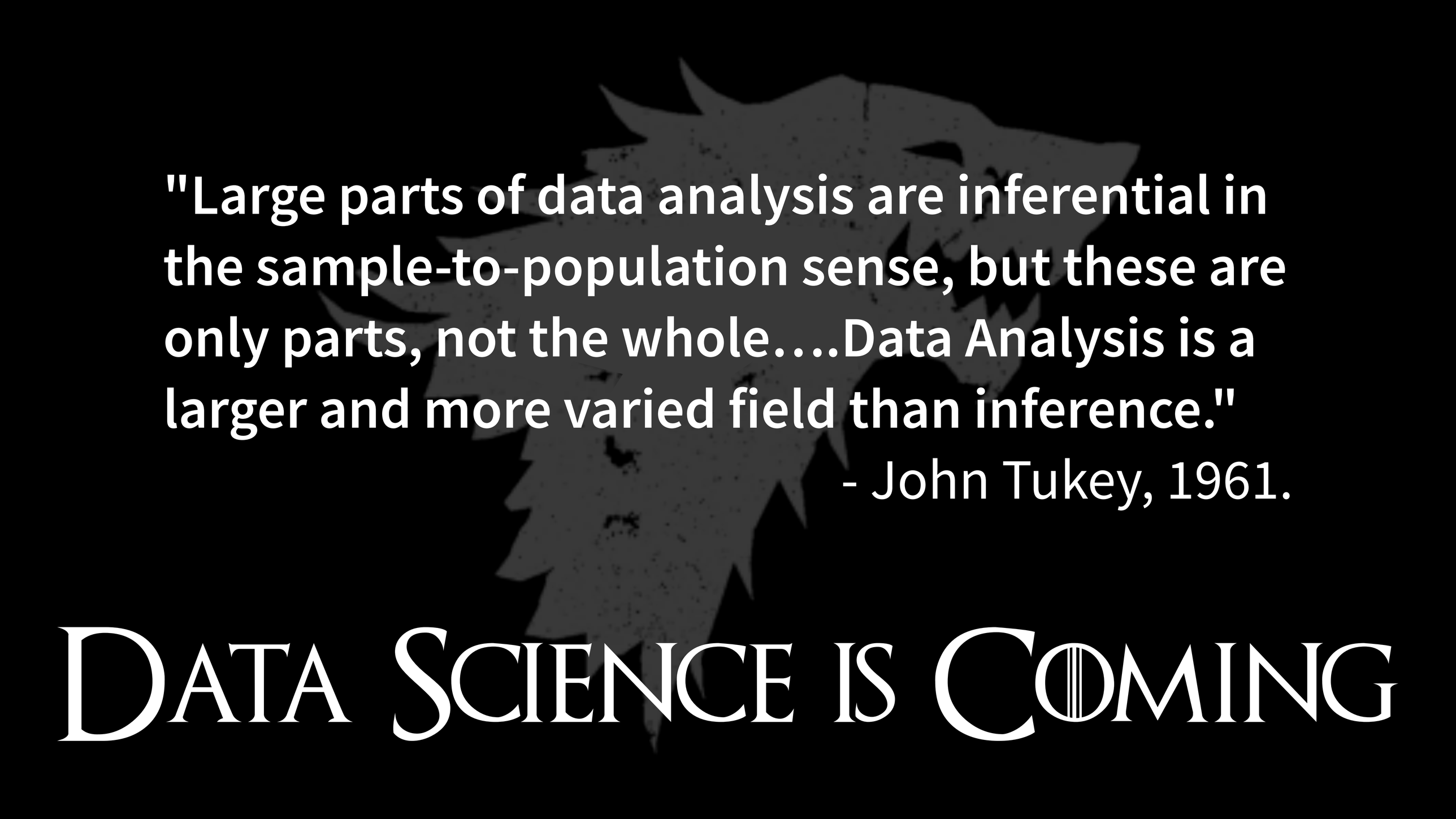
George Cobb, 2007

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

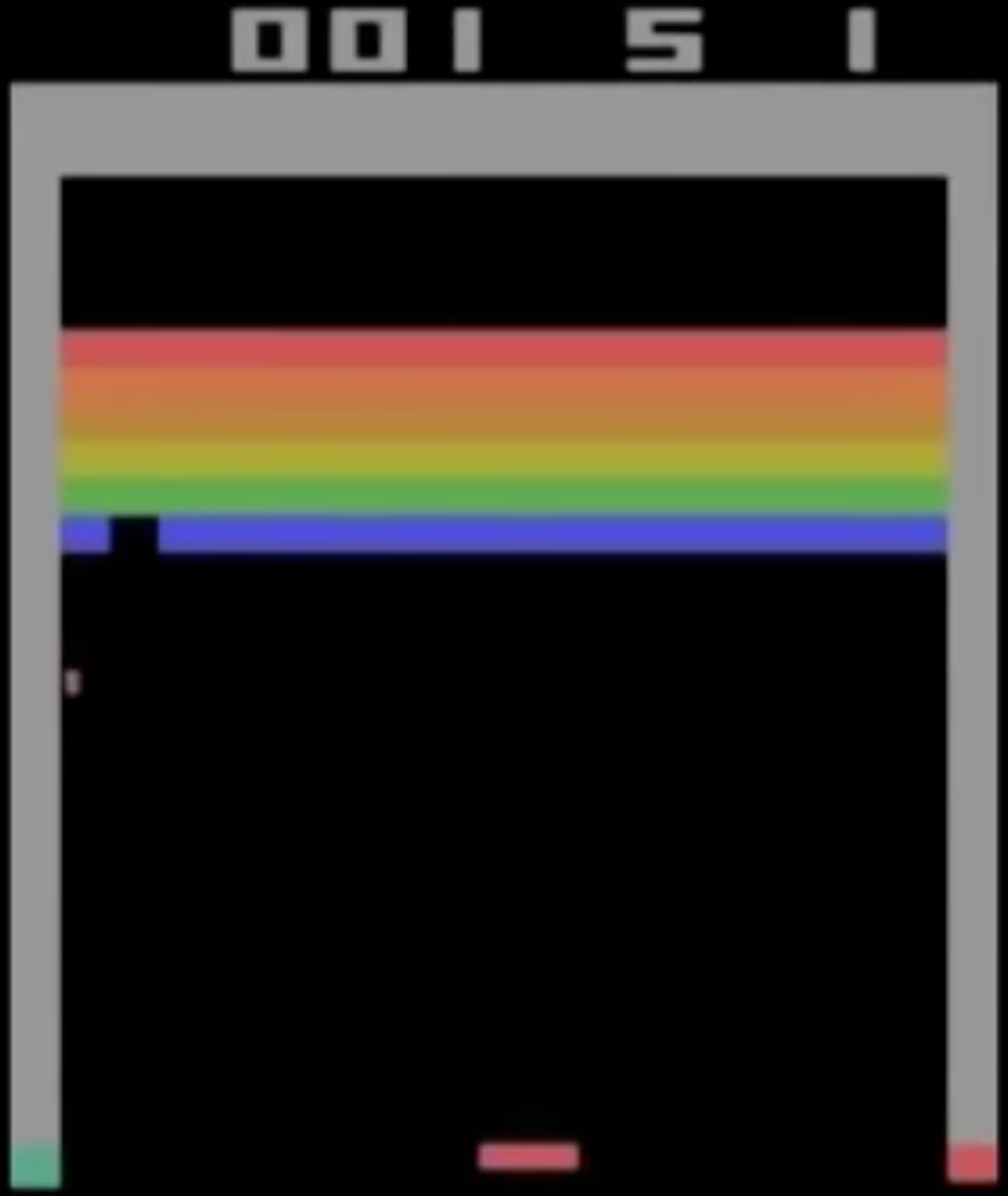
Data Science is Coming

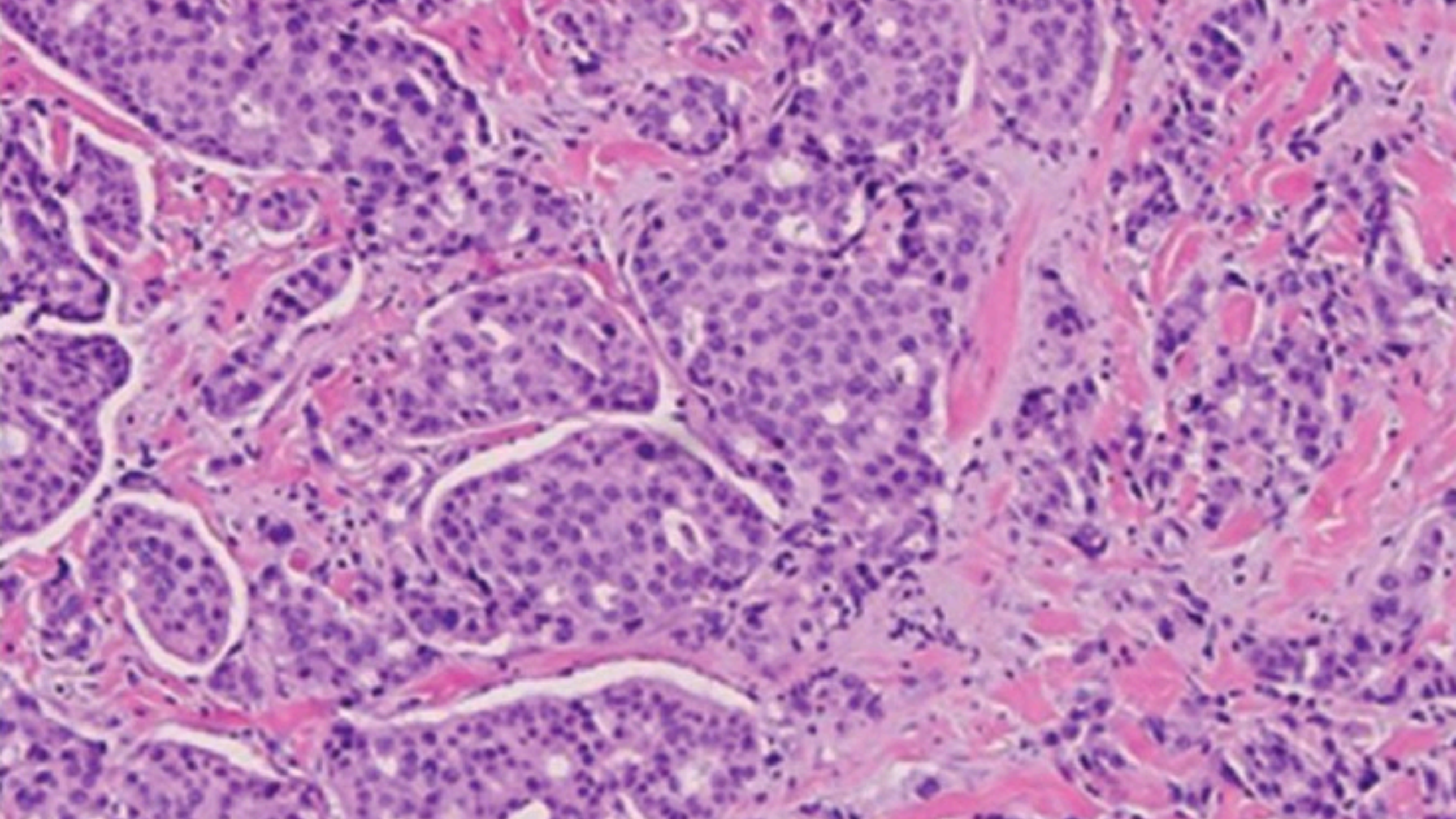


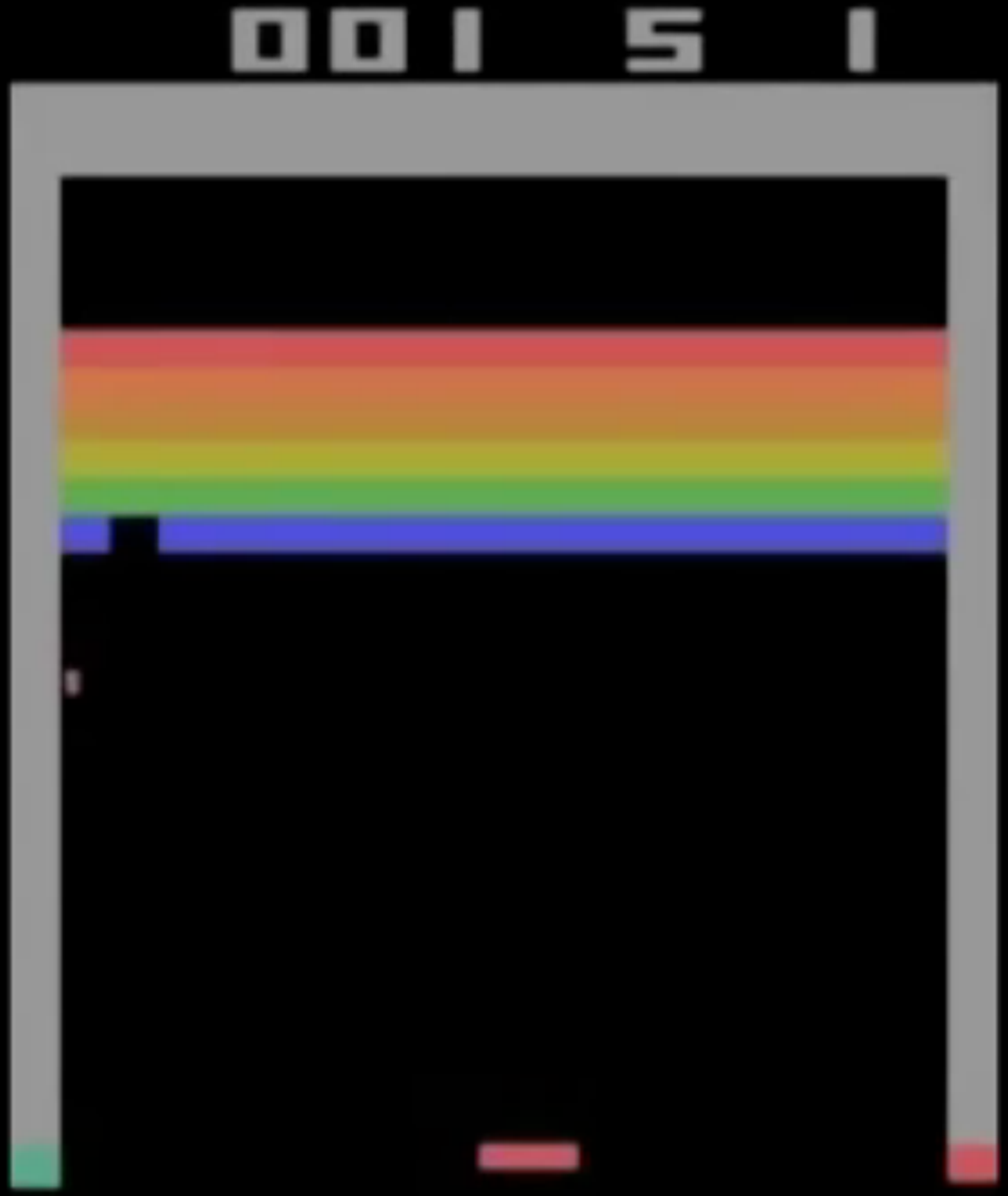
"Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole….Data Analysis is a larger and more varied field than inference."

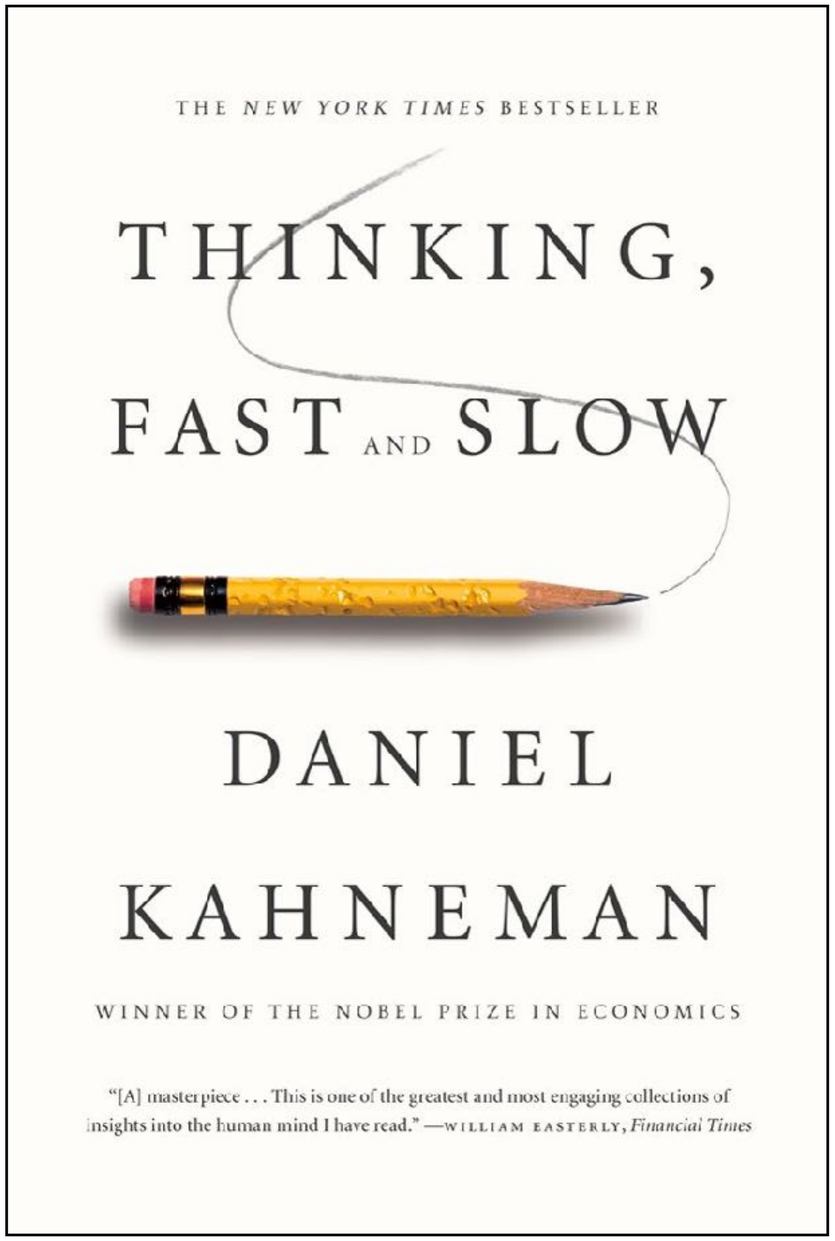
- John Tukey, 1961.

Data Science is Coming


001 5 1

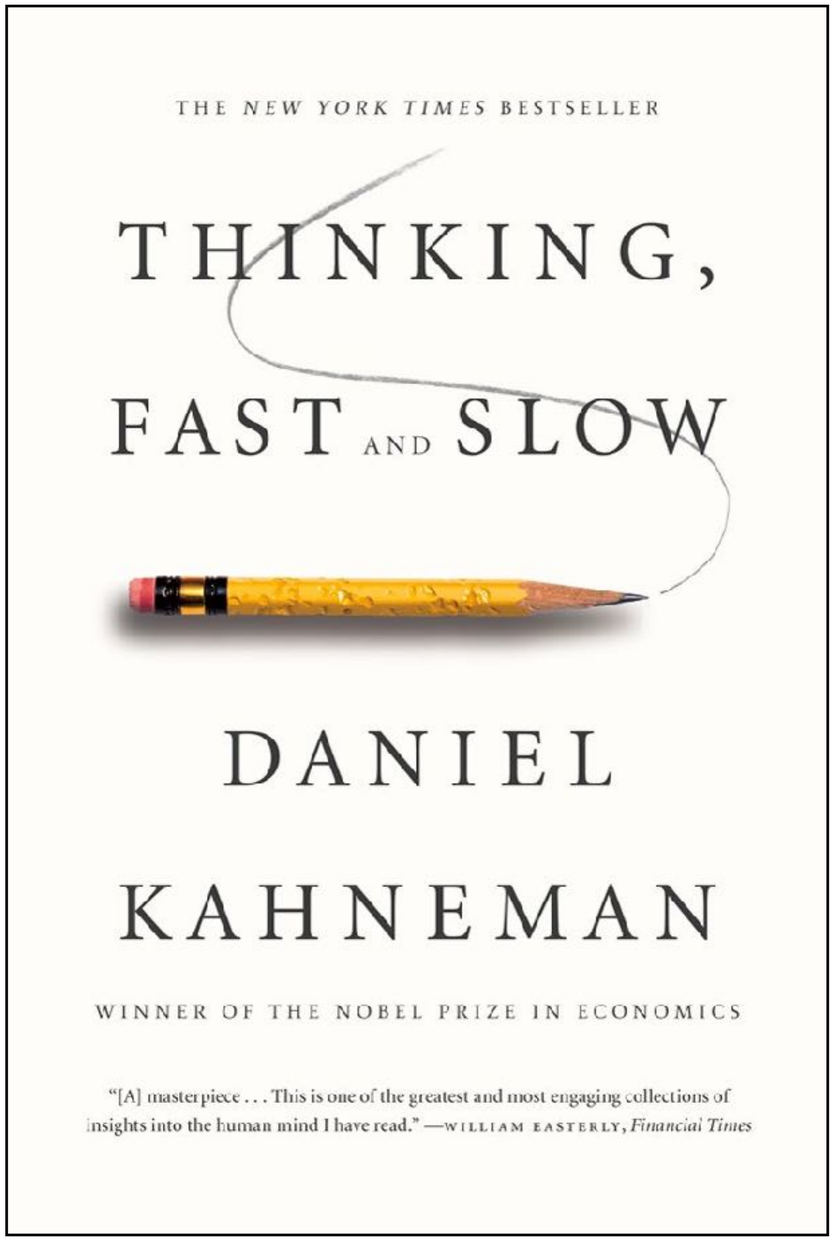



001 5 1



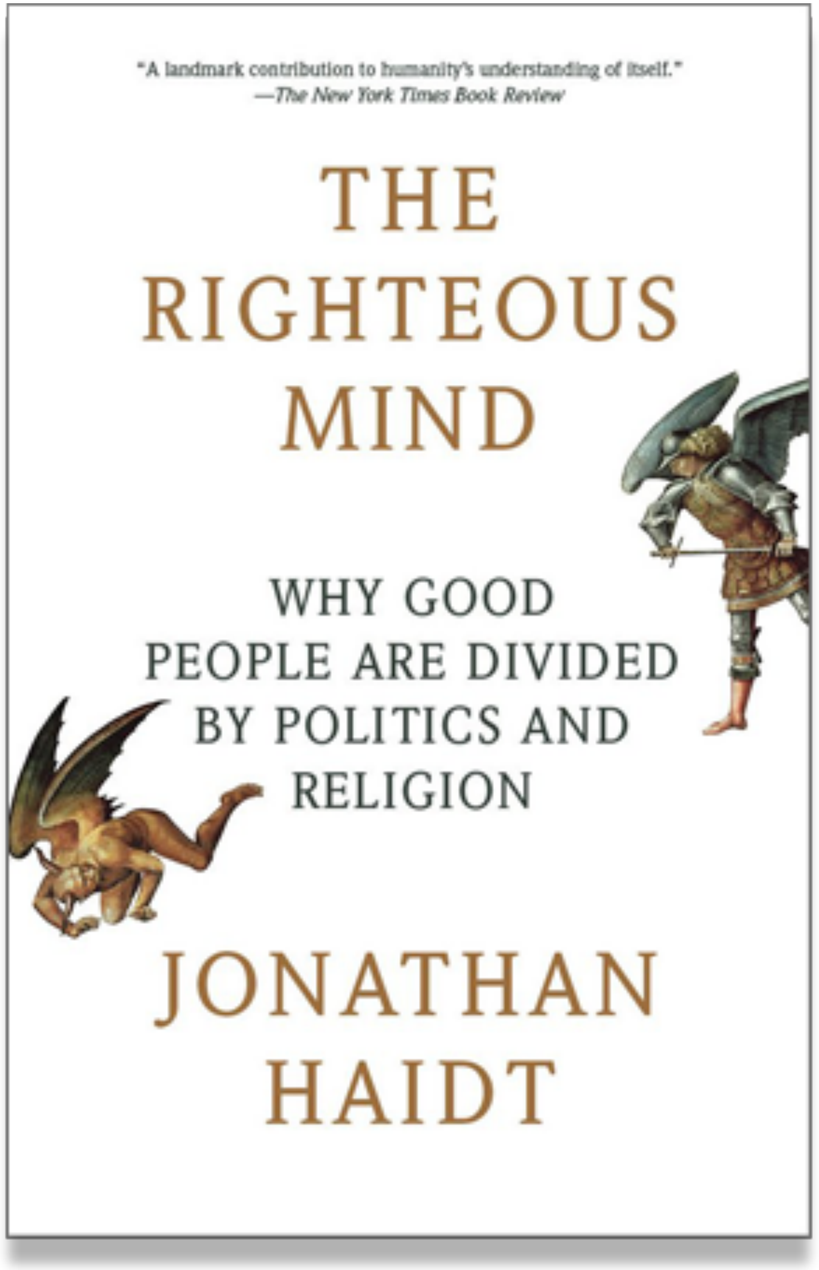
1. **System 1** - automatic, fast, unconscious information processing

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

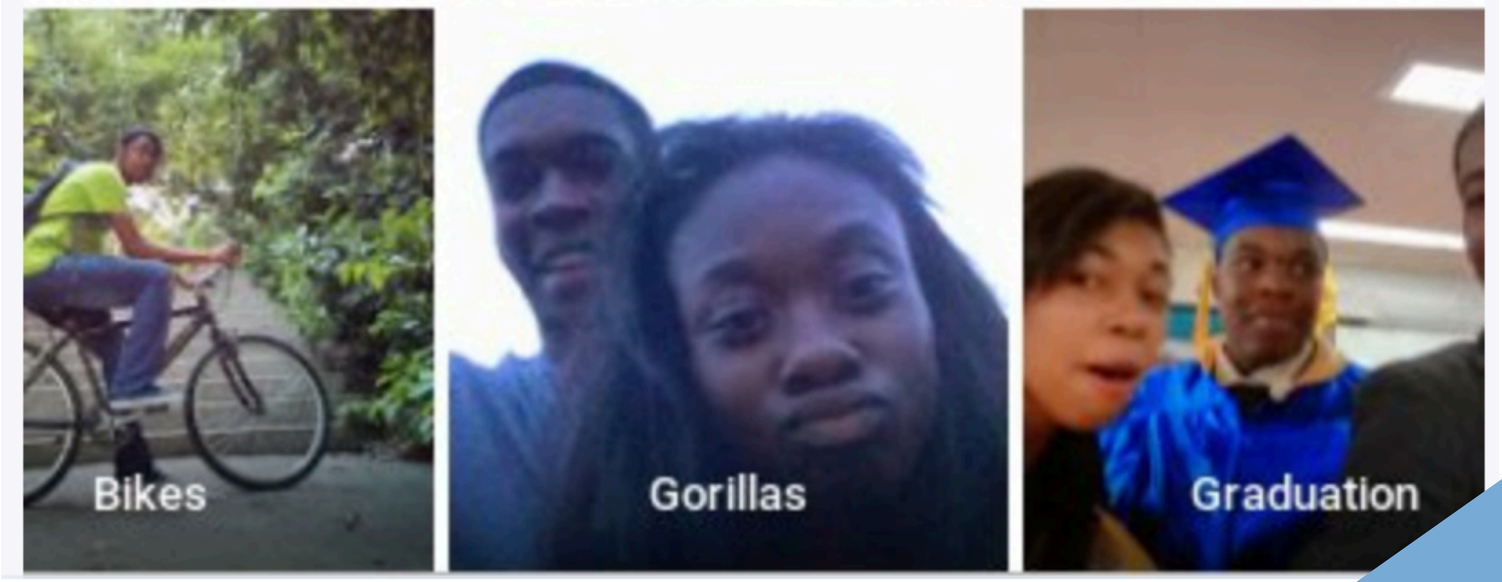
2. **System 2** - slow, expensive, logical, conscious reasoning



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

A System 1 failure

Machine Learning?

Data Scientists?

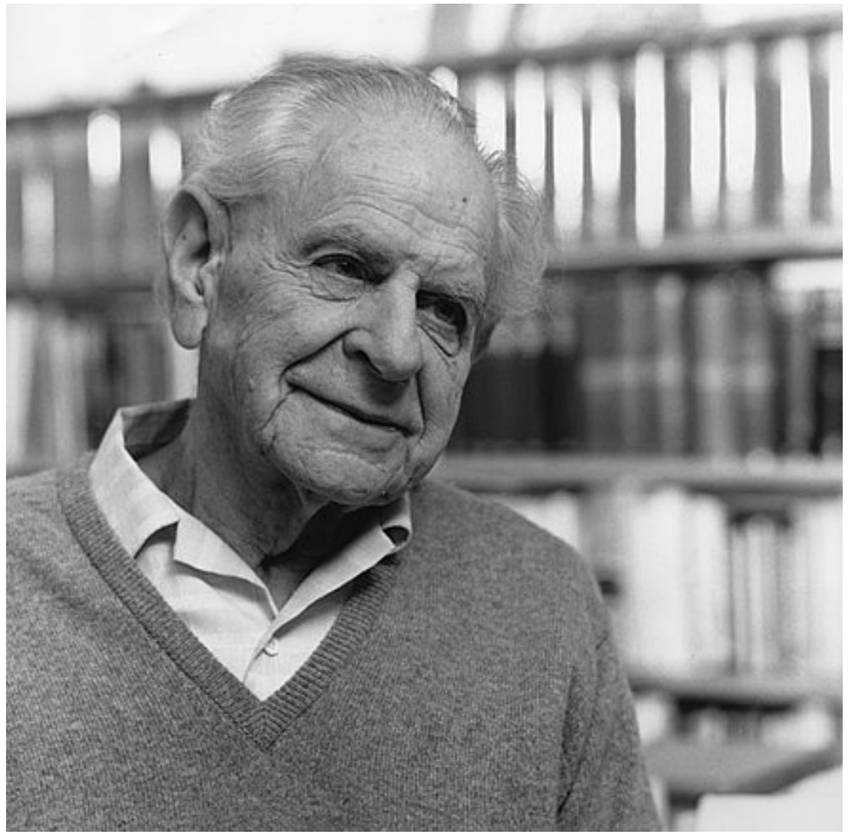
1. **System 1** - a system of heuristics that are highly efficient, but not always correct

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


A
Larger
Reasoning
Process

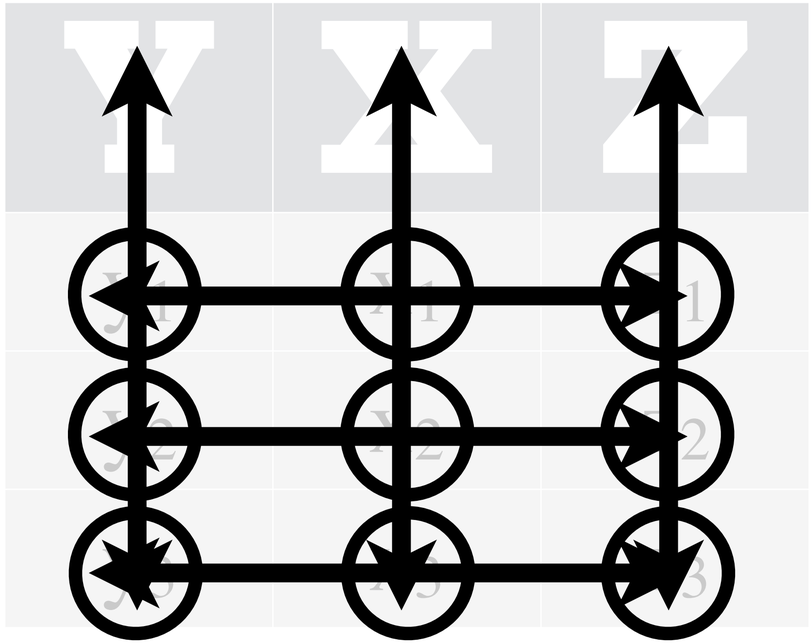
Discover

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

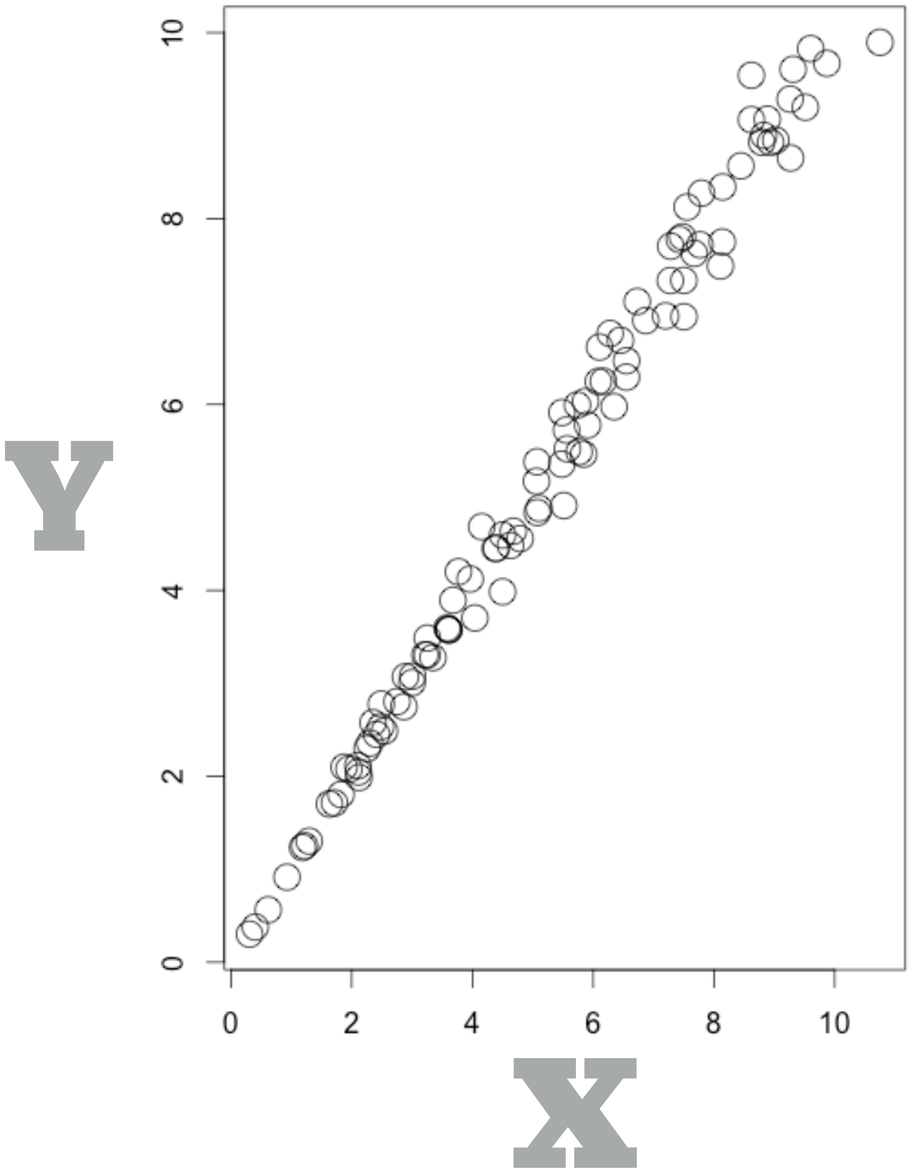
Laws appear as **patterns** in data.

Y	X	Z
3.01	0.98	3.08
2.35	0.91	2.58
5.57	1.01	5.52
3.67	1.00	3.56
4.15	0.89	4.69
5.07	1.05	4.84
7.56	0.93	8.12

Laws appear as **patterns** in data.

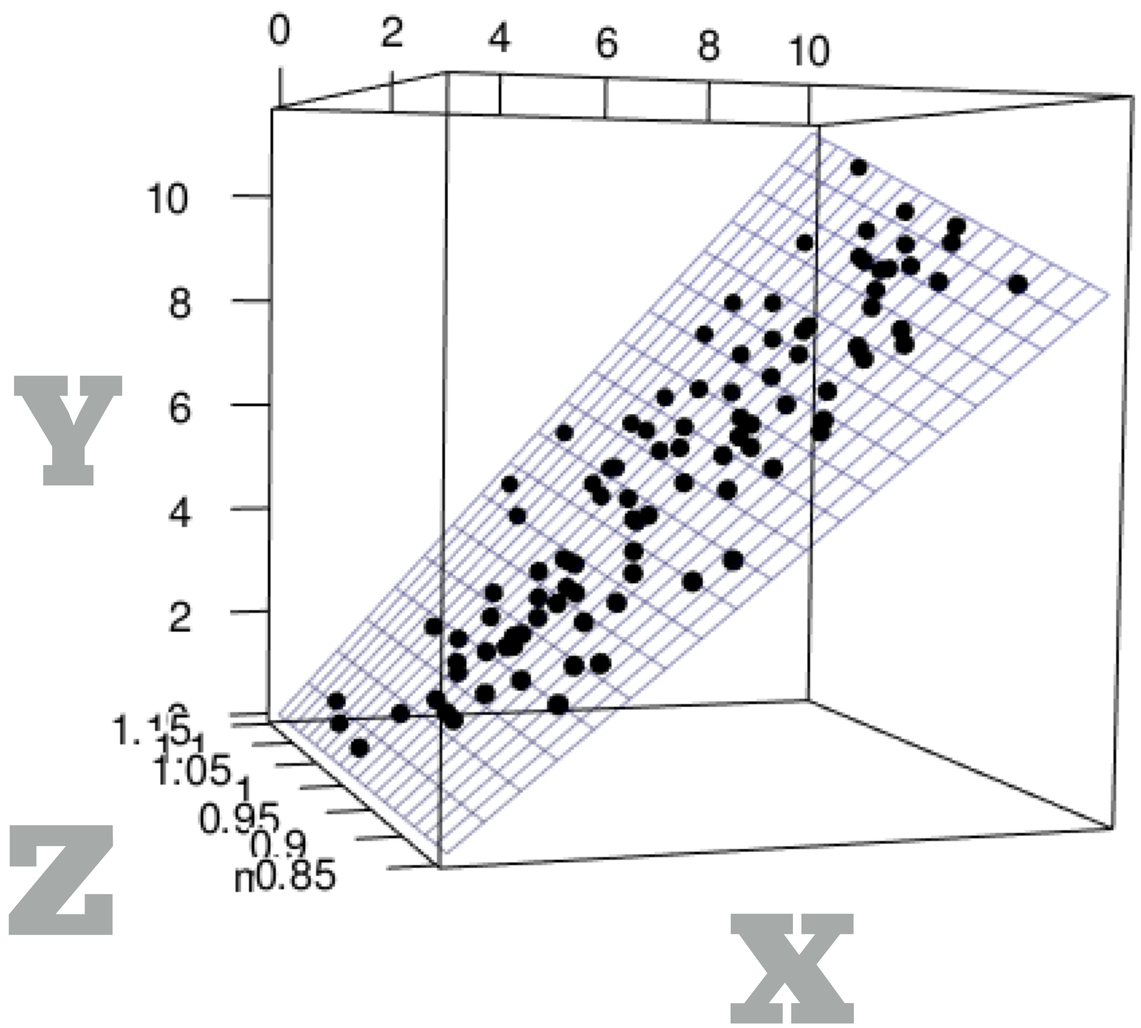
$$Y = X + \varepsilon$$

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

Descriptive Statistics → Inference

EDA → Inference

Data Visualization → Inference

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

- John Tukey

Discover Confirm

Descriptive Statistics → Inference

EDA → Inference

Data Visualization → Inference

Machine Learning → Inference

Discover Confirm

Small Data,
No compute power

Descriptive Statistics → Inference

EDA → Inference

Data Visualization → Inference

Machine Learning → Inference

Big Data,
Ample compute power

Discover

Confirm

Small Data,
No compute power

Big Data,
Ample compute power

Descriptive Statistics → Parametric Statistics

EDA → Linear Models +

Data Visualization → Non-Parametric

Machine Learning → Bayesian Statistics

Discover Confirm

Small Data,
No compute power

Descriptive Statistics → Parametric Statistics

EDA → Parametric Statistics

Data Visualization → Parametric Statistics

Machine Learning → Parametric Statistics

Big Data,
Ample compute power

The American Statistician

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

Discover

Confirm

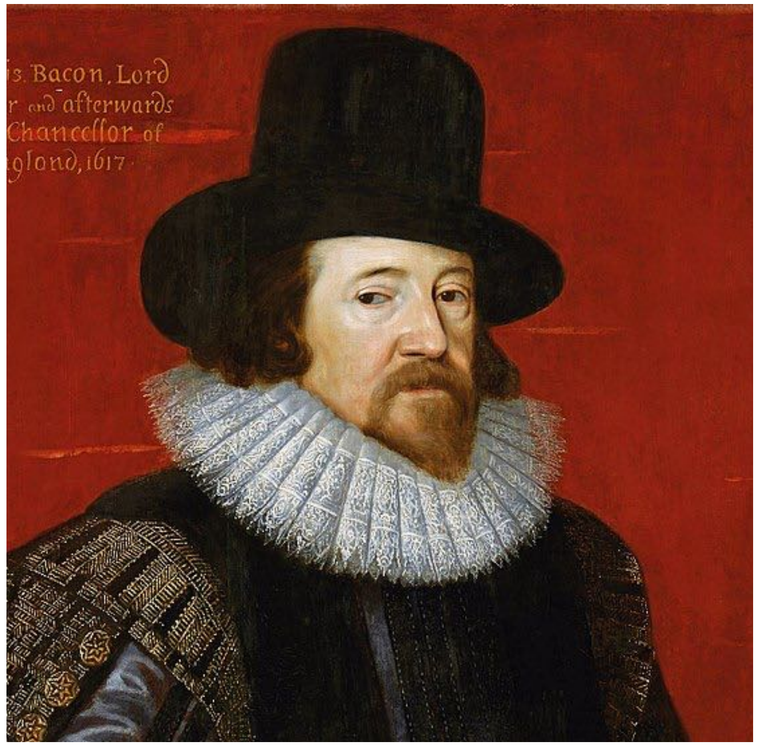
Descriptive Statistics → Parametric Statistics

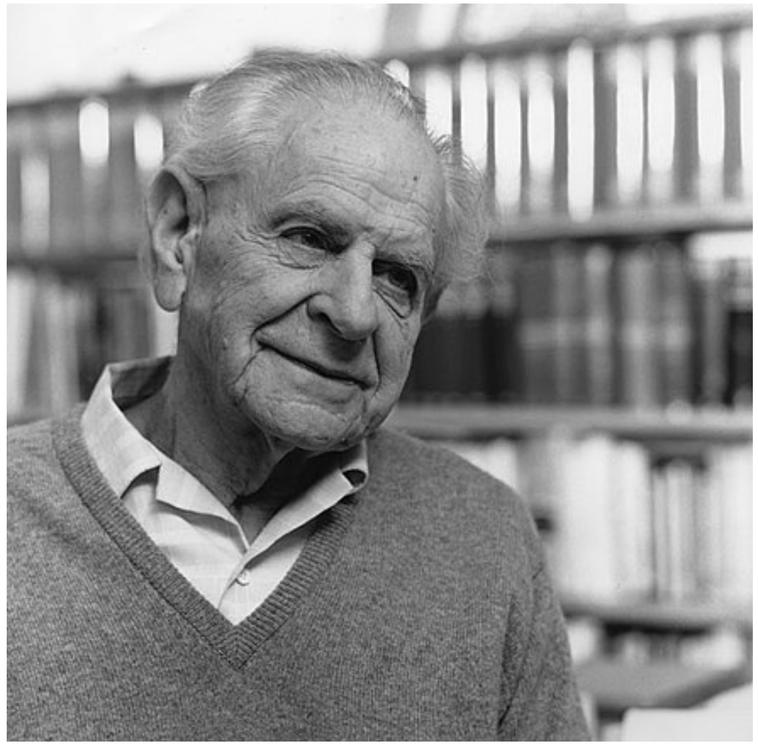
EDA → Linear Models +

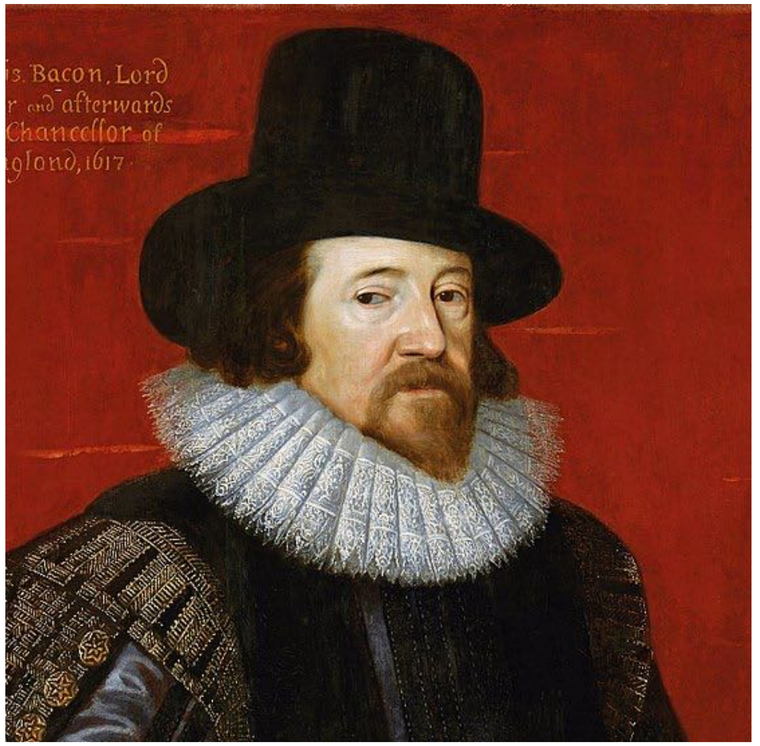
Data Visualization → Non-Parametric

Machine Learning → Bayesian Statistic

Teach the craft
not the tools







Discover

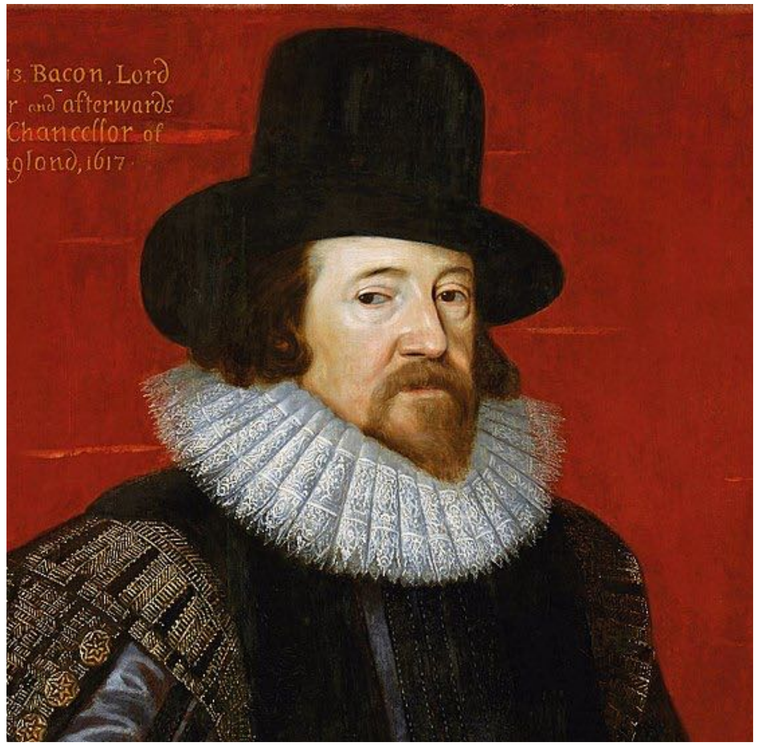
Induce hypothesis

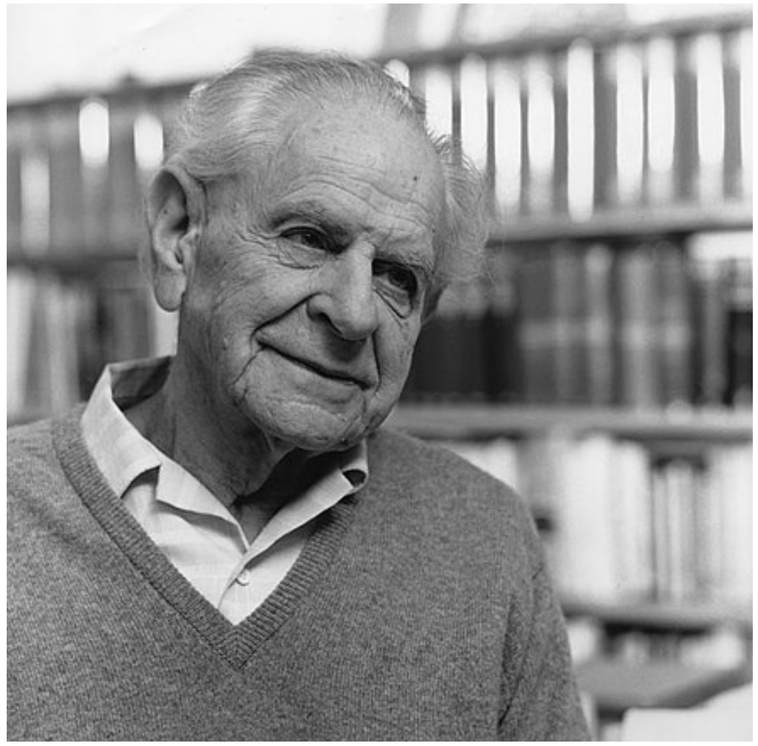
Confirm

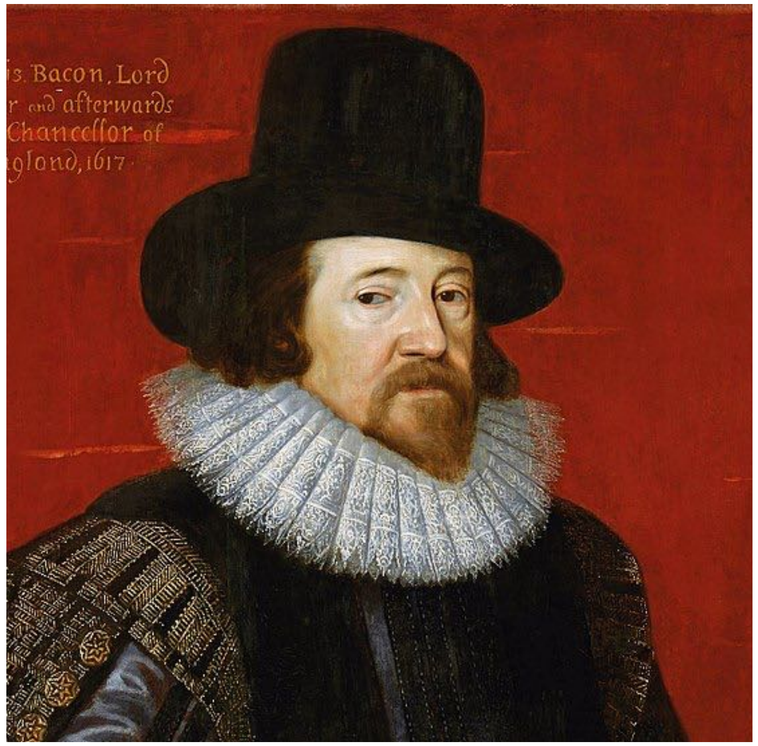
Deduce validity

Bound

Define applicability







Discover

Induce hypothesis

Confirm

Deduce validity

Infer

Define applicability

Clinical Trial

(randomized treatment, random sampling)

Discover

Descriptive Statistic

Confirm

Permutation Test

Infer

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

Descriptive Statistic

Confirm

Cannot confirm

Infer

Applies to population, qualify with bootstrapped confidence interval

Machine Learning

(that yields prediction)

Discover

Model algorithm selected through cross validation

Confirm

Nothing to Confirm

Infer

Applies to events that are similar to training events

Regression Model

(that yields coefficients)

Discover

Model fit with OLS, MLE, etc.

Confirm

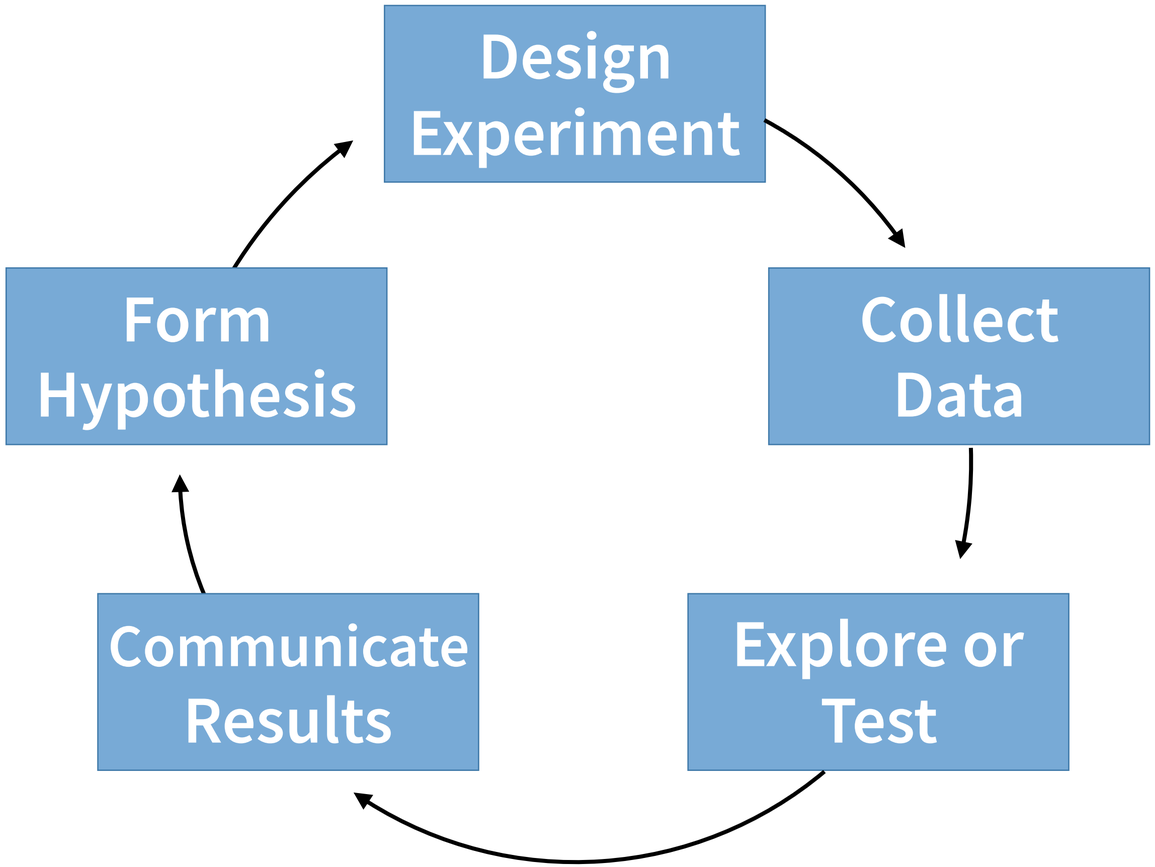
Permutation test

Infer

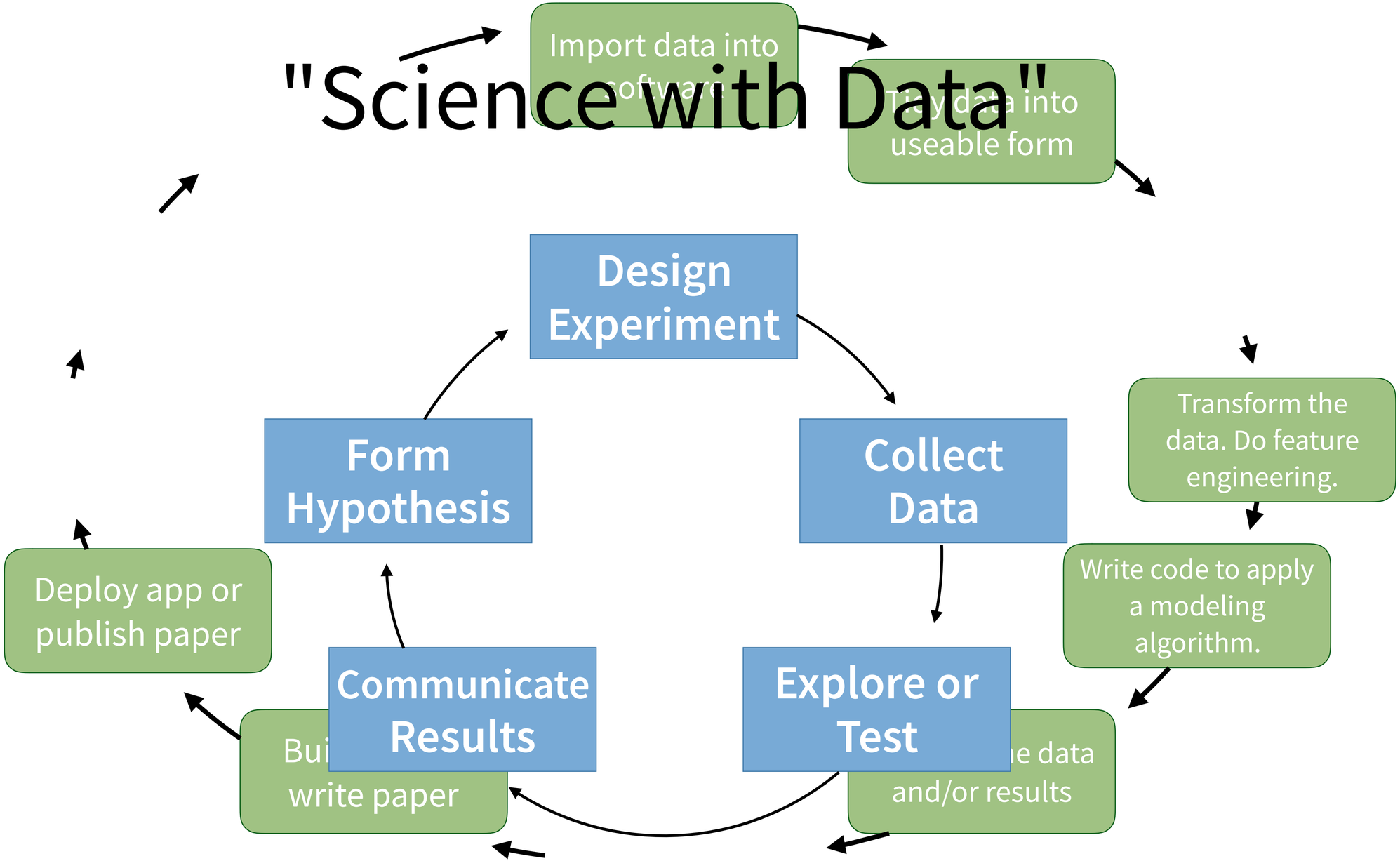
Applies to events with similar sources of unexplained variation

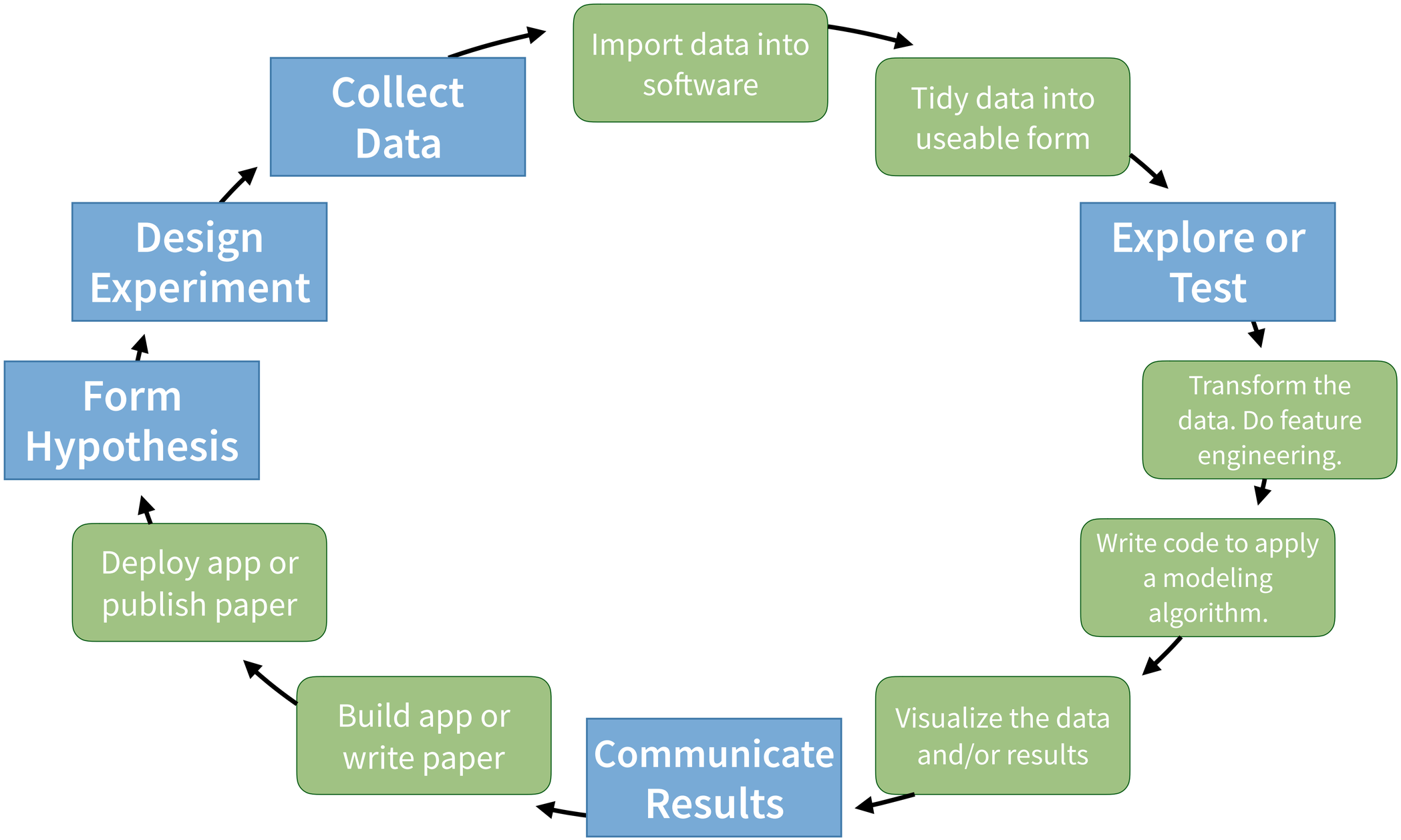
Teach the practicum

"Science with Data"

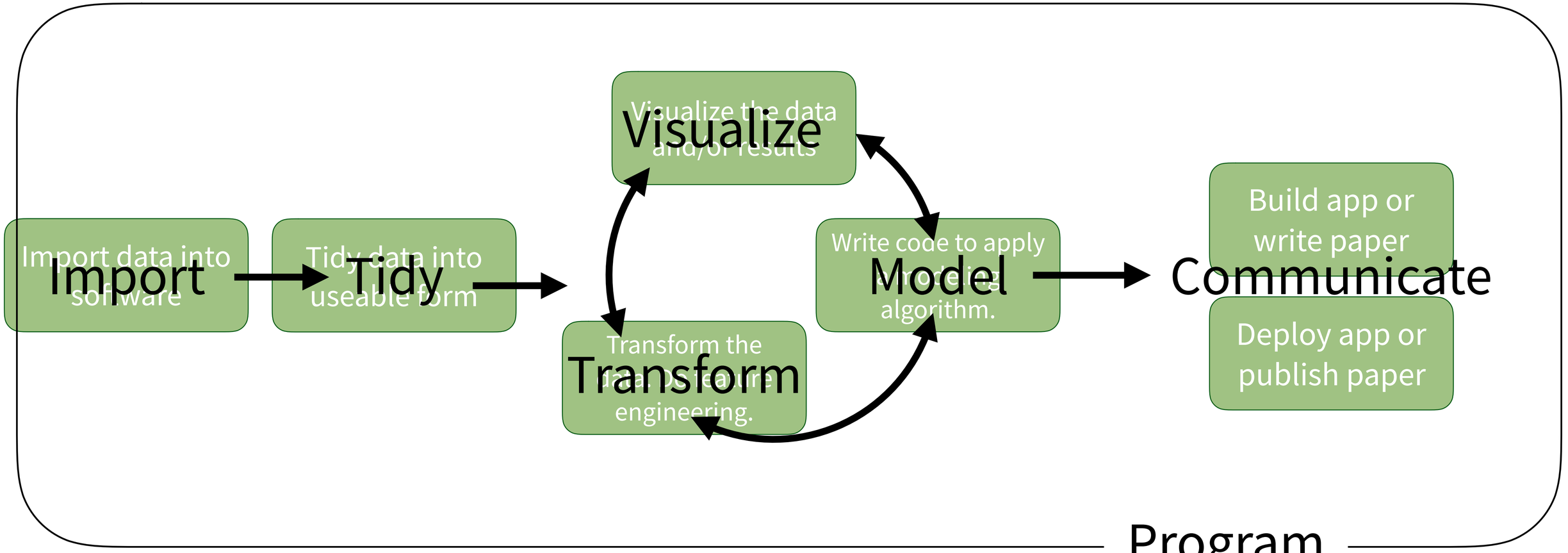


"Science with Data"

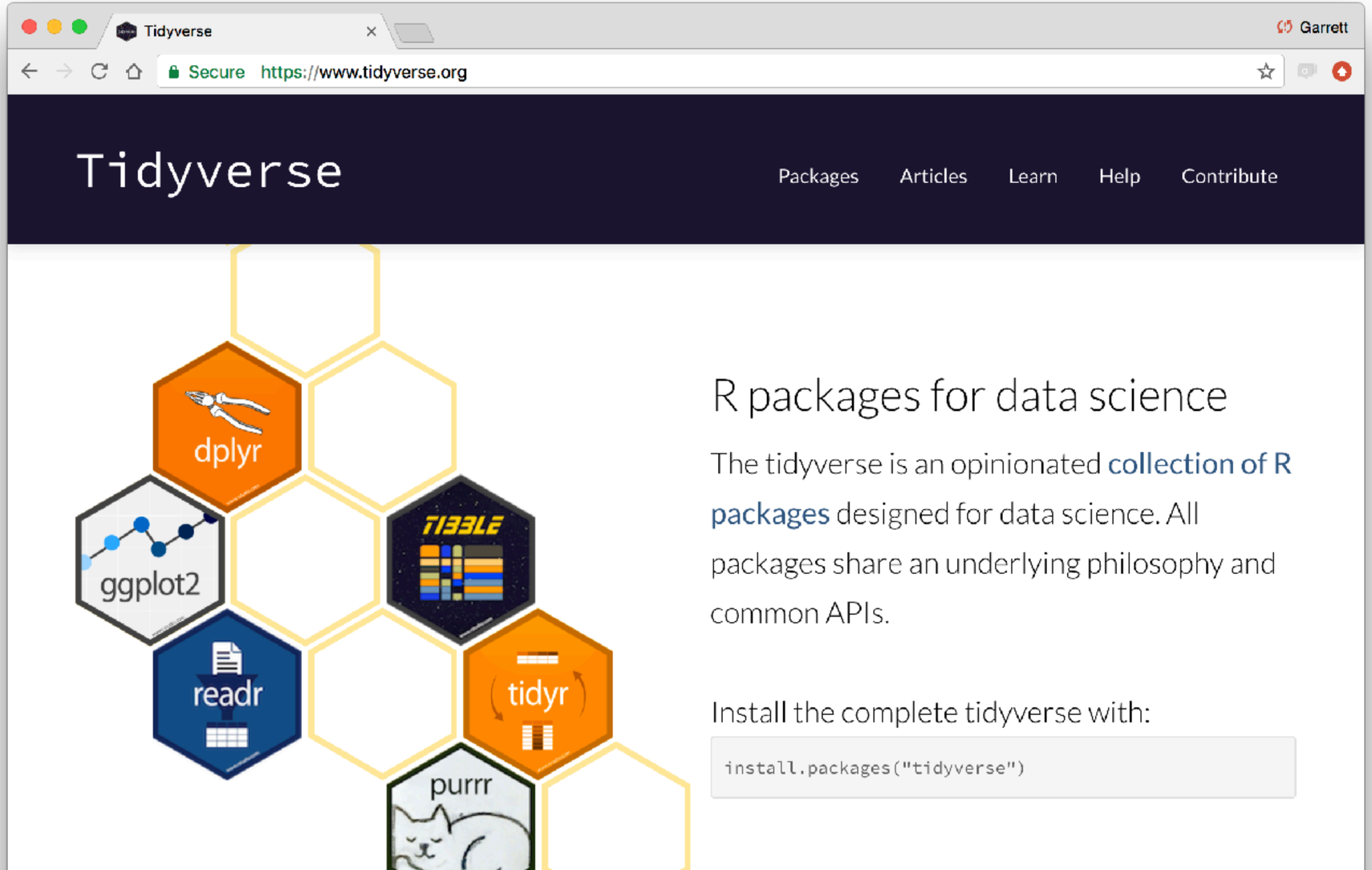



Collect Data
Import data into software
Tidy data into useable form
Explore or Test
Transform the data. Do feature engineering.
Write code to apply a modeling algorithm.
Visualize the data and/or results
Communicate Results
Build app or write paper
Deploy app or publish paper
Form Hypothesis
Design Experiment

(Applied) **Data Science**



tidyverse.org



www.rstudio.com/resources/cheatsheets

Cheatsheets – RStudio

Garrett

Secure | https://www.rstudio.com/resources/cheatsheets/

RStudio

rstudio::conf Products Resources Pricing About Us Blogs

RStudio Cheat Sheets

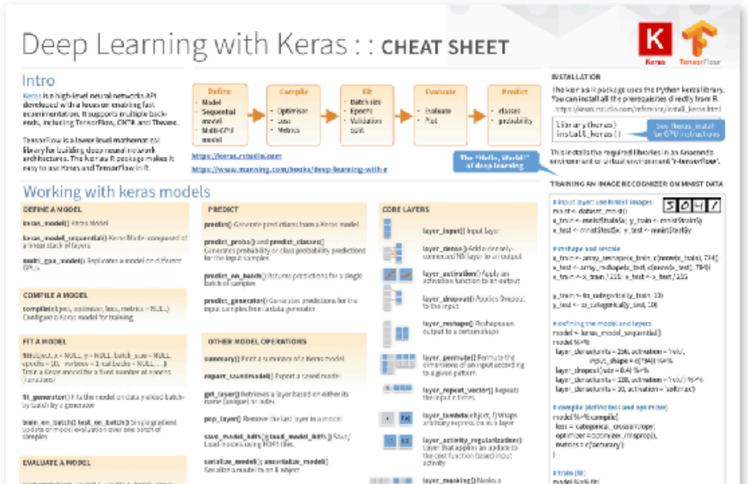
The cheat sheets below make it easy to learn about and use some of our favorite packages. From time to time, we will add new cheat sheets to the gallery. If you'd like us to drop you an email when we do, let us know by clicking the button to the right.

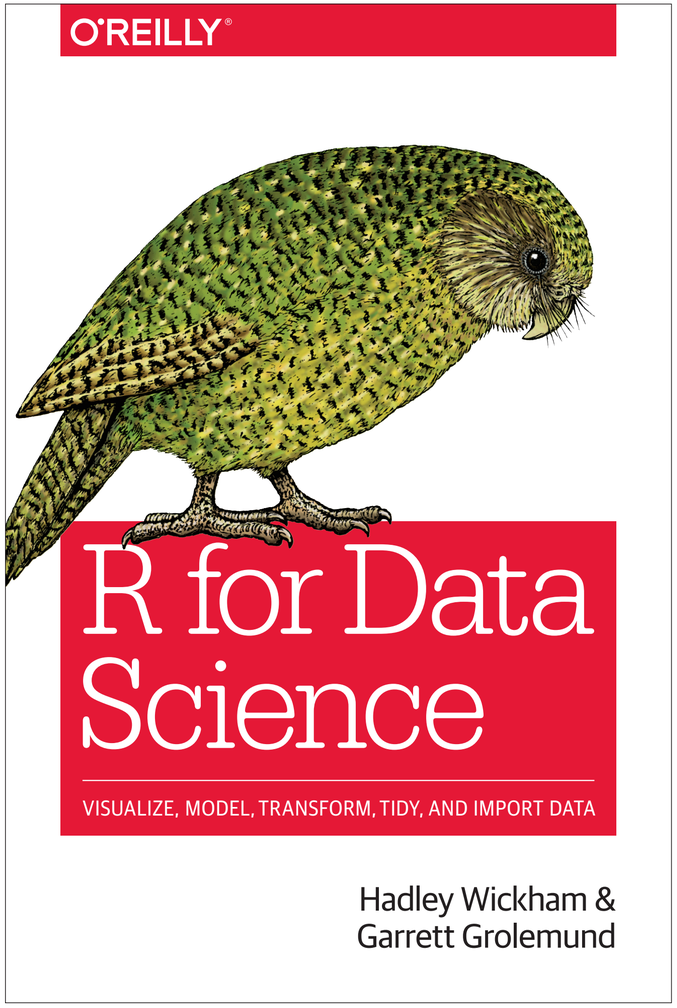
SUBSCRIBE TO CHEAT SHEET UPDATES HERE

Deep Learning with Keras Cheat Sheet

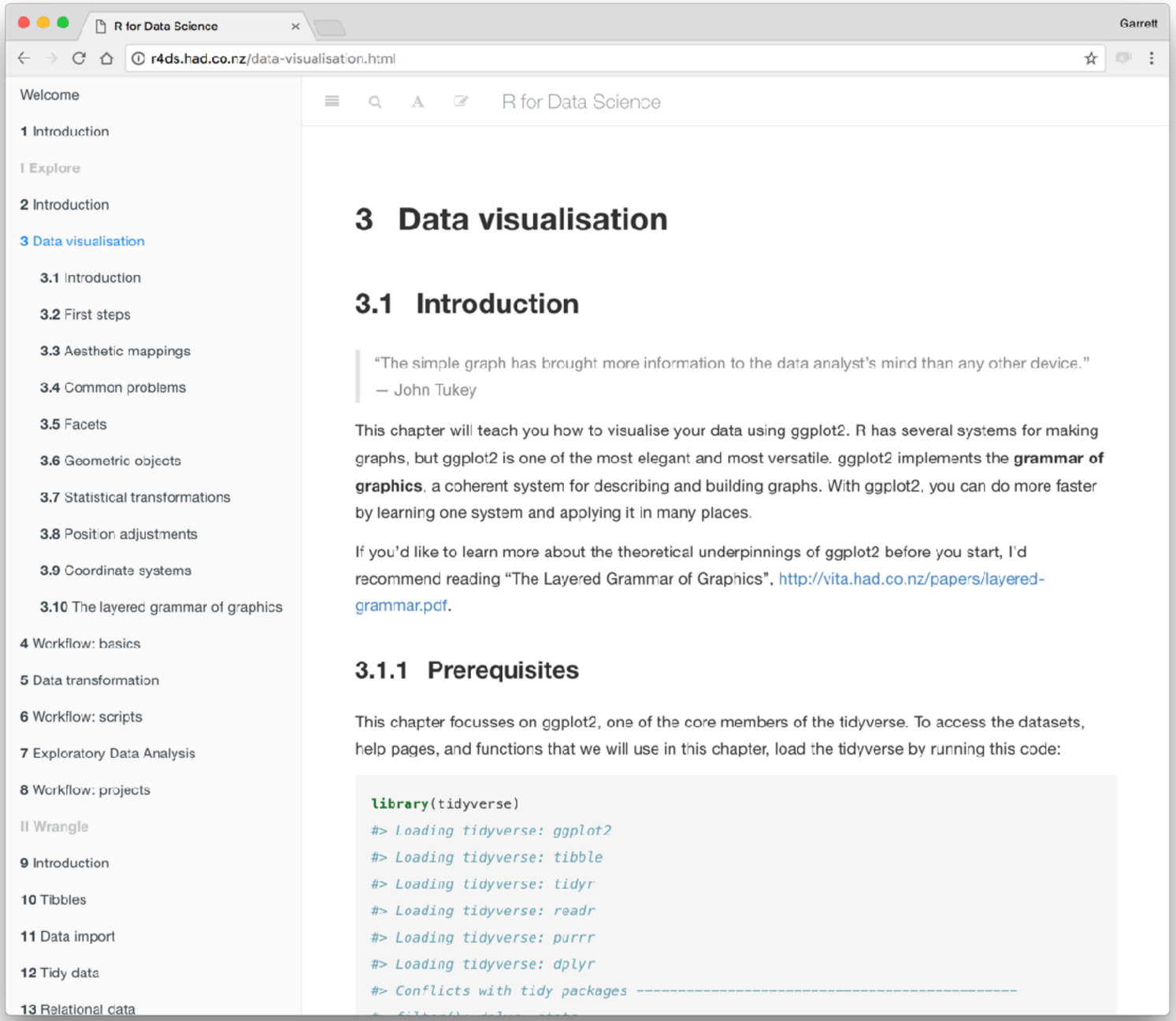
Keras is a high-level neural networks API developed with a focus on enabling fast experimentation. Keras supports both convolution based networks and recurrent networks (as well as combinations of the two), runs seamlessly on both CPU and GPU devices, and is capable of running on top of multiple back-ends including TensorFlow, CNTK, and Theano. Updated 12/17.

DOWNLOAD





http://r4ds.had.co.nz



RStudio Cloud

Garrett

Secure https://rstudio.cloud/learn/primers

Your Workspace | Projects | Learn

Garrett Grolemund

Guide | Primers | DataCamp Courses | Cheat Sheets

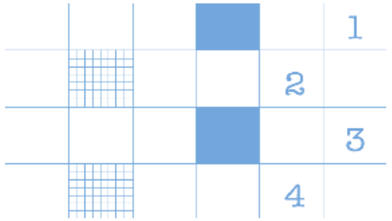
RStudio Primers

Learn data science basics with the interactive tutorials below.

The Basics

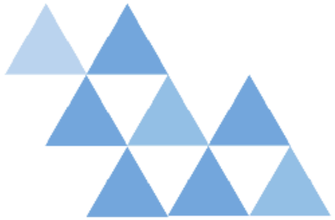
Start here to learn the skills that you will rely on in every analysis (and every primer that follows): how to inspect, visualize, subset, and transform your data, as well as how to run code.

Work with Data



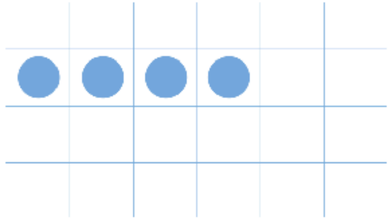
Learn the most important data handling skills in R: how to extract values from a table, subset tables, calculate summary statistics, and derive new variables.

Visualize Data



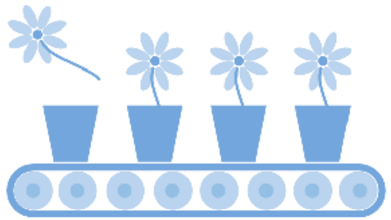
Learn how to use ggplot2 to make any type of plot with your data. Then learn the best ways to visualize patterns within values and relationships between variables.

Tidy Your Data



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

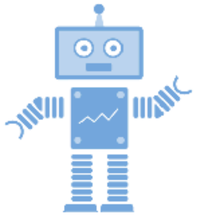
Iterate



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

Automate Tasks

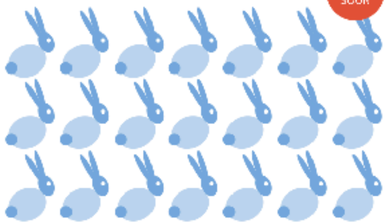
COMING SOON



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

Photo by Jesus Kiteque on Unsplash

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

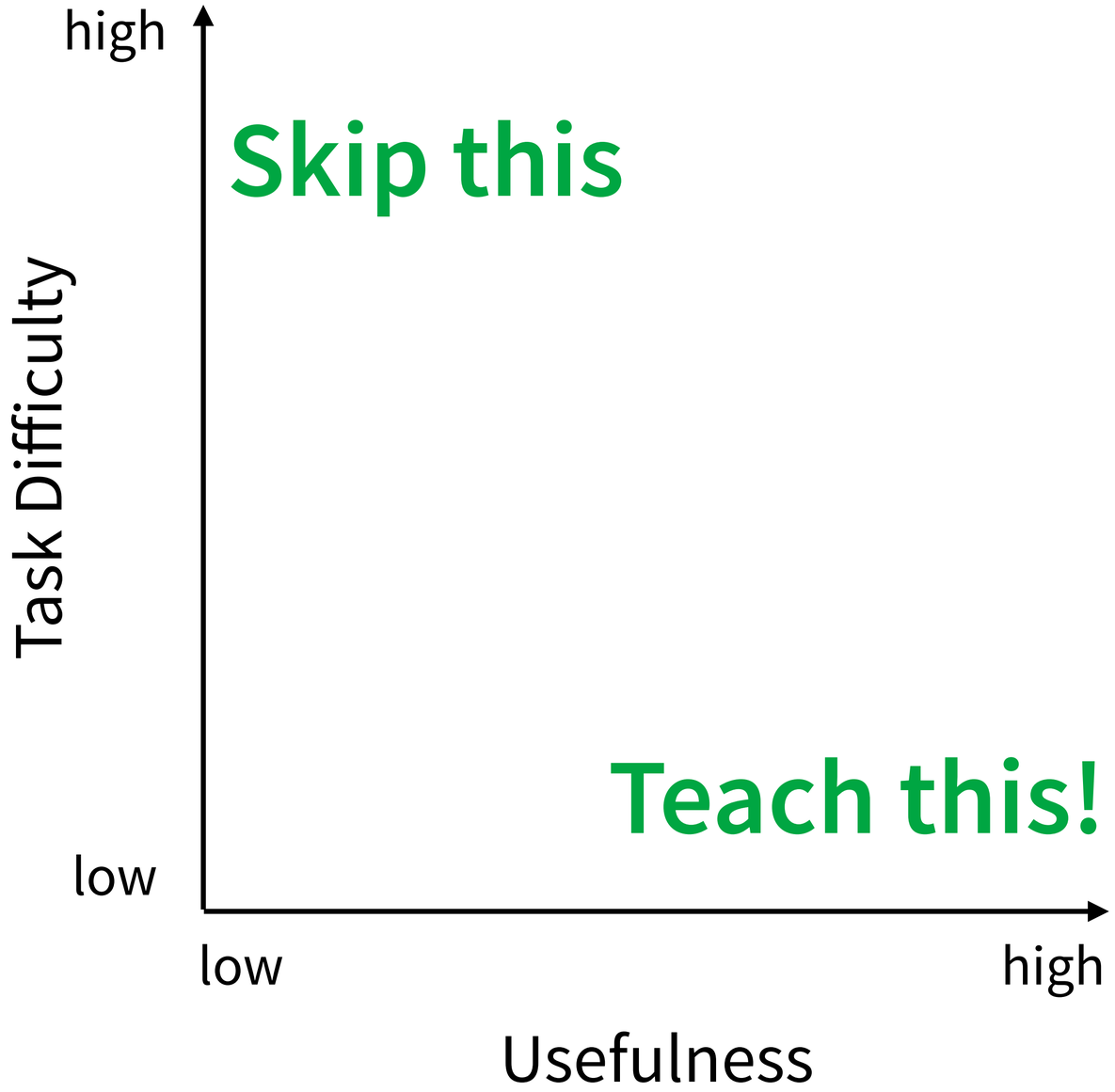
Data Science or Computer Science?

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?


high
Skip this
Task Difficulty
Teach this!
low
low
high
Usefulness

Quiz

Confer with your group.

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

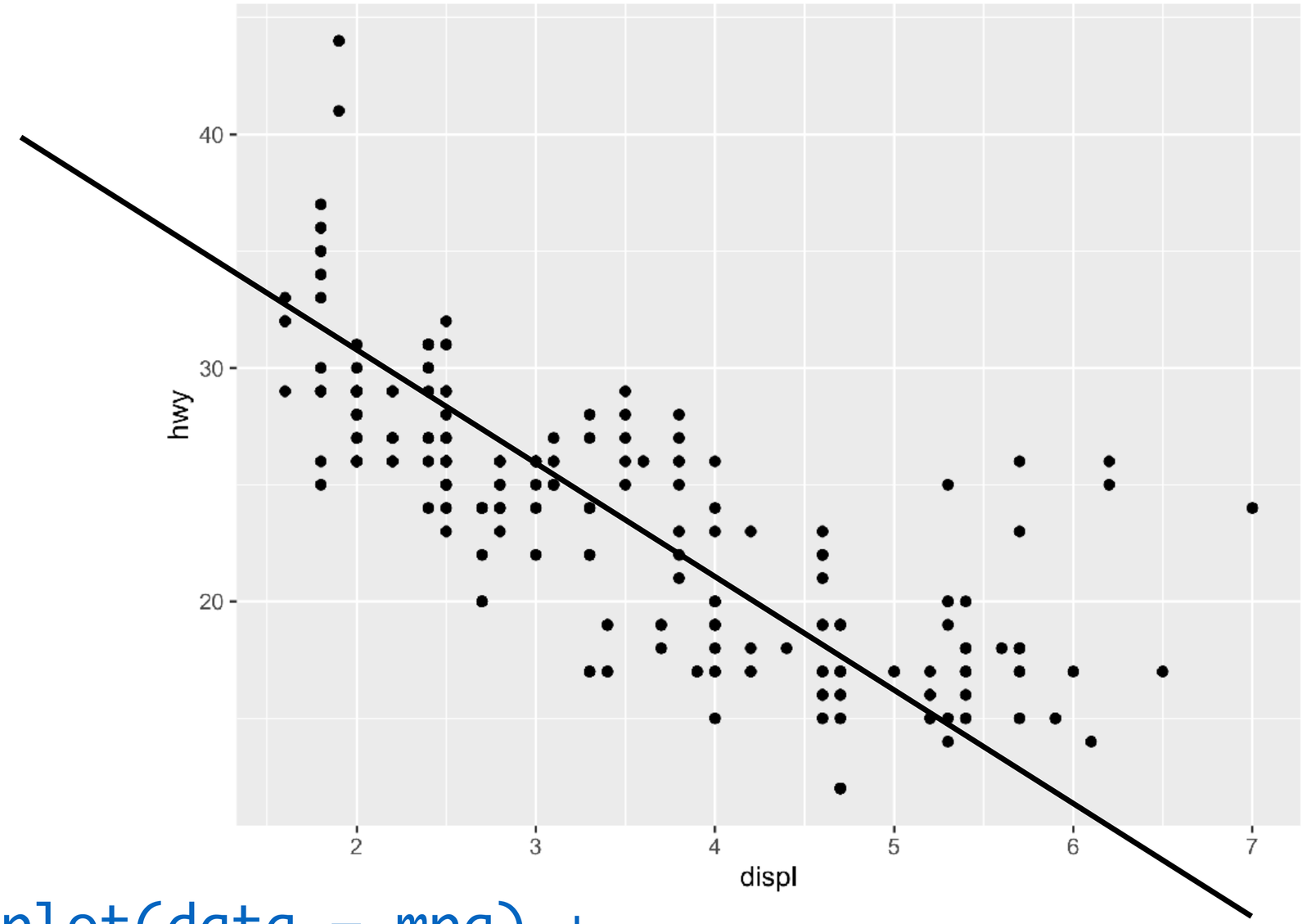
No peeking ahead!

Your Turn 1

Run this code in your notebook to make a graph.

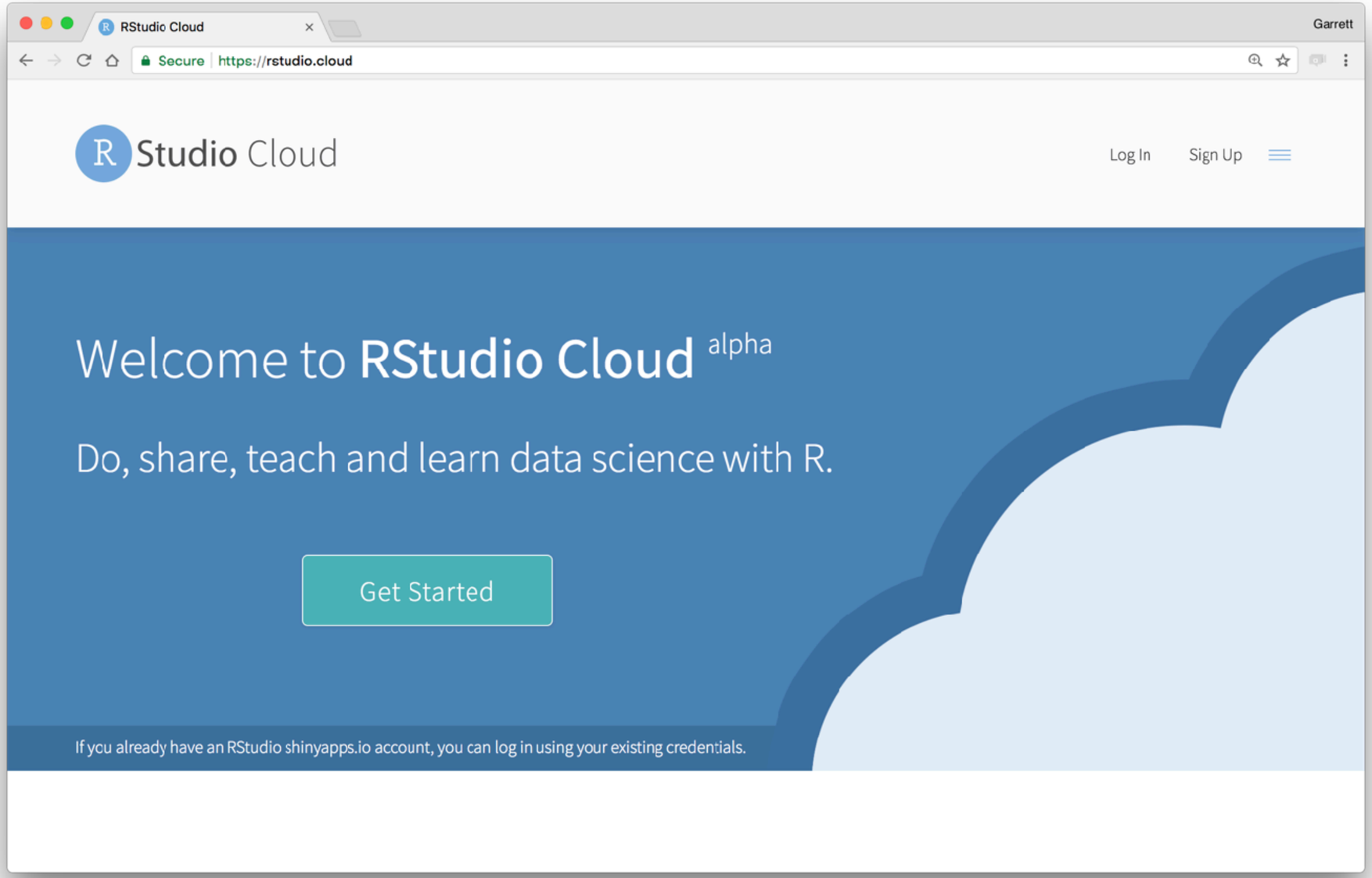
Pay strict attention to spelling, capitalization, and parentheses!

```
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

Photo by Rafaela Biazi on Unsplash

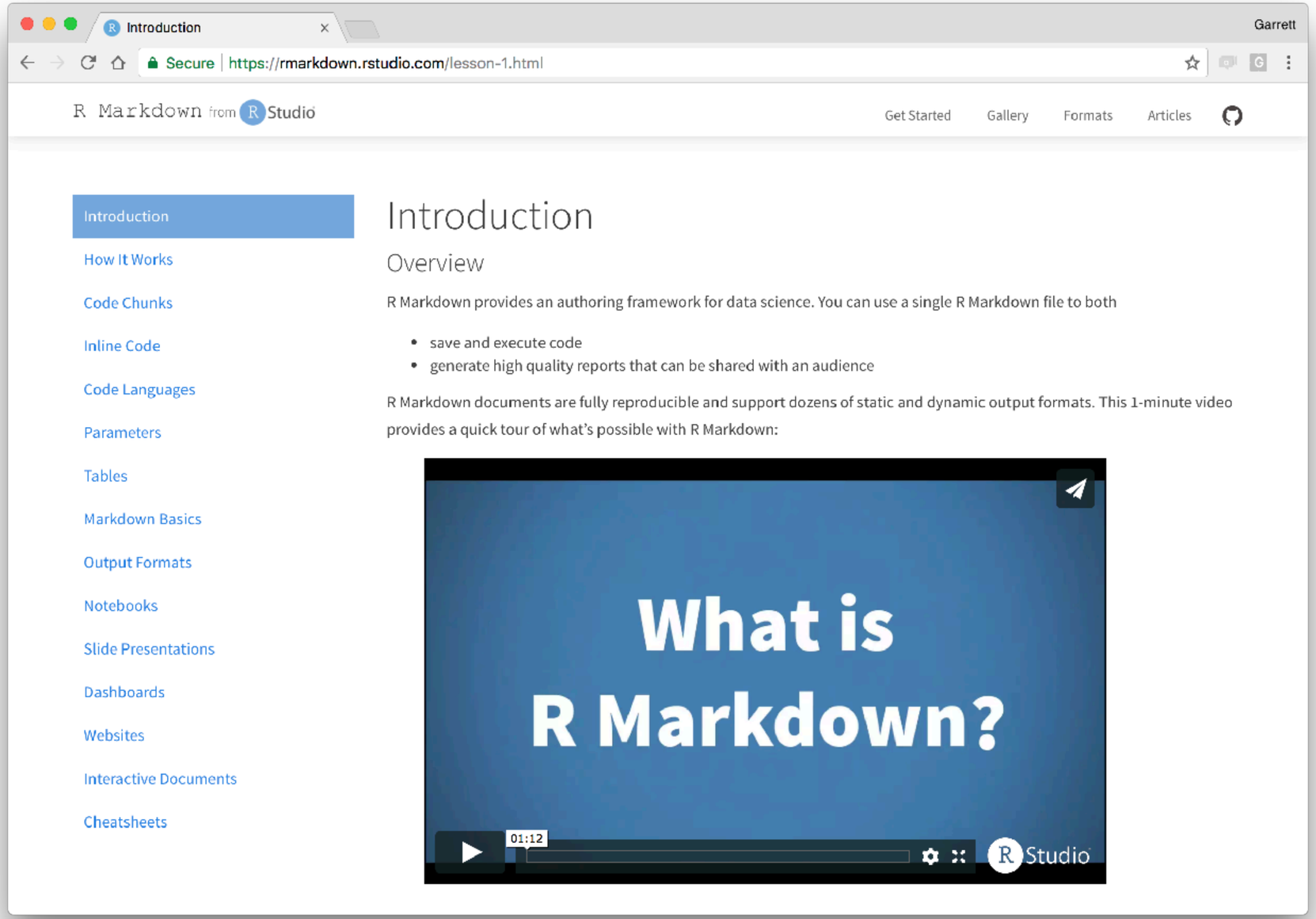
Reproducibility

Photo by Keagan Henman on Unsplash

www.bit.ly/reprodu

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

rmarkdown.rstudio.com/lesson-1.html



Collaboration

Photo by Max Liao on Unsplash

Github

Photo by Owen Massey McKnight on flickr

Github

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

Photo by Owen Massey McKnight on flickr

Photo by SpaceX on Unsplash