DATA ANALYTICS BOOTCAMP

INSTRUCTOR:
Alex Sierra, CEO Sigma Ridge

A Leader in Education

Today's complex, global economy requires a skilled workforce that can leverage technology to fuel success. Since 2011, General Assembly has transformed careers and teams through pioneering, experiential education in today's most in-demand skills.

GA's robust suite of courses includes all the fundamental pillars of innovation to give individuals and teams options for growth and development. These skills — coding, data, design, digital marketing, and product management — foster innovation and drive the modern economy.

At a Glance

- Award-winning curriculum and expert instructors at **20** global campuses, online, and in-office.
- A thriving alumni community of **50,000+** full- and part-time graduates.
- Dedicated career coaching for full-time students, with **7,000+** hiring partners, including Capital One, IBM, and NBC.

Corporate training and hiring solutions with **350+** companies worldwide, including **39** of the Fortune 100.

More than **500,000** attendees at bootcamps, workshops, and events.

Alexander Sierra

Alexander Sierra has a diverse background. He has a bachelor's degree in Aerospace Engineering from the University of Florida, A master's in Finance from Harvard University, and an MBA from the University of Massachusetts. Over 23 years of experience as a leader in Consulting, Sales & Marketing. Alex has managed teams of over 130 direct reports in 13 different countries and Marketing budgets of over 25 million dollars a year. He has been able to achieve year-over-year growth for his clients from 25% to 120%.

His consulting practice **-Sigma Ridge-** was a spinoff from the Harvard University Consulting club where he works with companies like Cisco, Dell, and some of the largest fashion companies in the US.

Goals and Objectives

01

Introduce fundamental concepts for data analysis

- Define data
- Discuss tools used by data analysts
- Review the data analytics workflow

02

Transform data for descriptive analytics and business intelligence decision-making

- Use Google Sheets to:
    - Obtain the data
    - Understand the data
    - Prepare the data

03

Develop compelling visualizations that articulate the value of data assets

- Use Tableau Public to:
    - Analyze the data
    - Present the data

Meet and Greet

1- Name

2- Current role

3- What you are trying to get out of this bootcamp?

INTRODUCTION: DEFINING DATA

DEFINING DATA

➔We live in a world filled with data; it’s all around us. But, *what* is it?

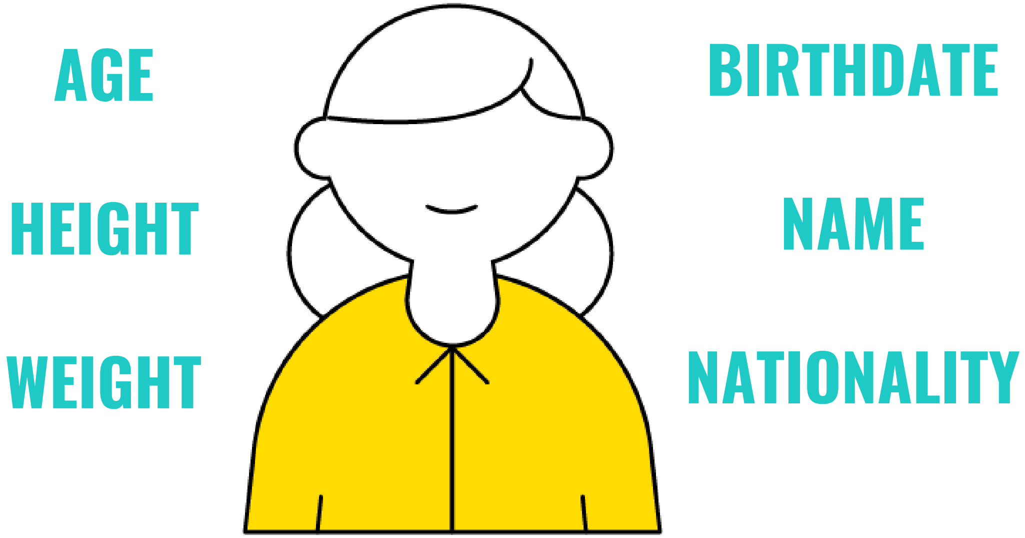
➔Data is *information* that exists in a variety of formats and sizes.

➔Data is an *object*, but it’s not necessarily seen at first glance.

➔Data leads us to decisions. Those decisions depend on the right data use

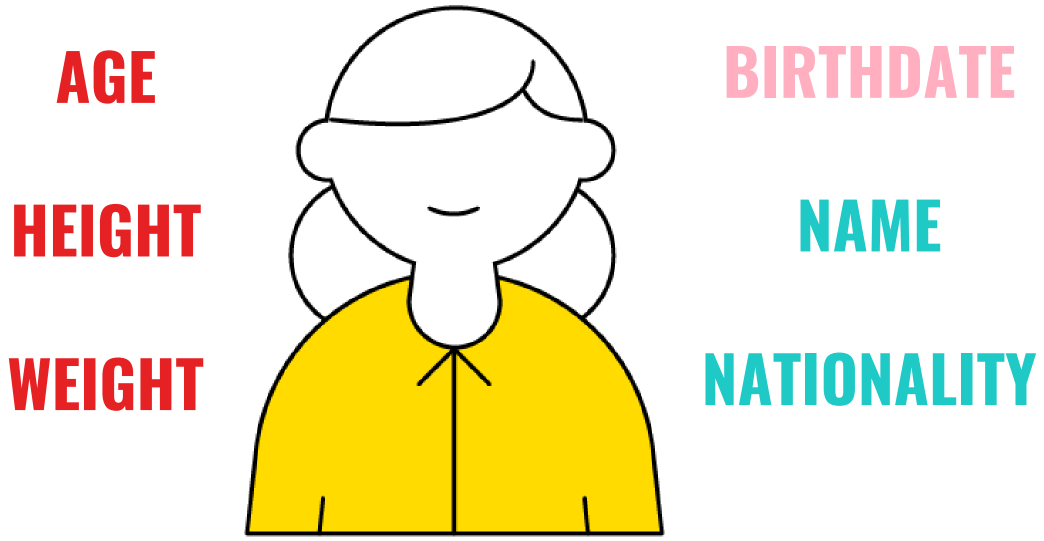
DEFINING DATA

- ➔ Even you have lots of data attached to you.
- ➔ You have *an age, a height, a weight, a birthdate, a name, and a nationality*.
- ➔ Each *piece* of information is considered data!

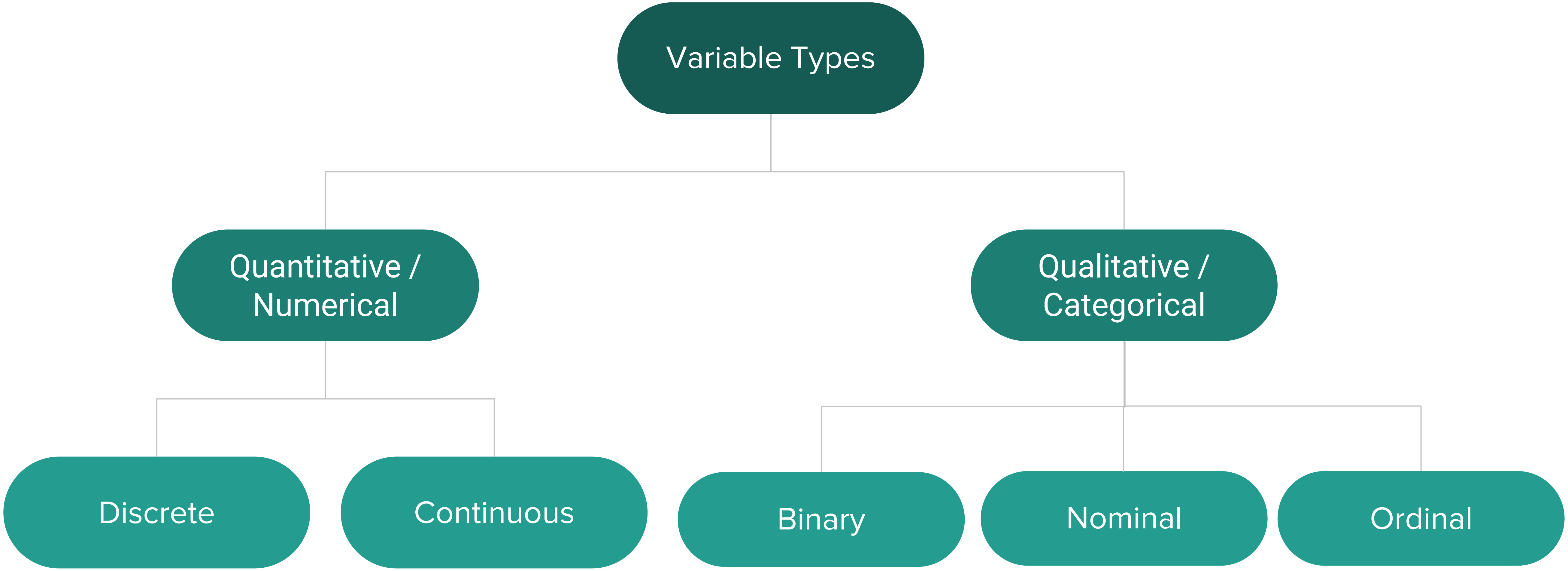


DEFINING DATA: TYPES OF DATA

- ➔This data, only about you, represents a variety of data types.
- ➔Your age, height, and weight are all **numbers**.
- ➔Your birthday is a **date**.
- ➔Your name and nationality are text (or **strings**, as we call them in data analytics).



DEFINING DATA: TYPES OF VARIABLES



Variable Types

Quantitative / Numerical

Qualitative / Categorical

Discrete

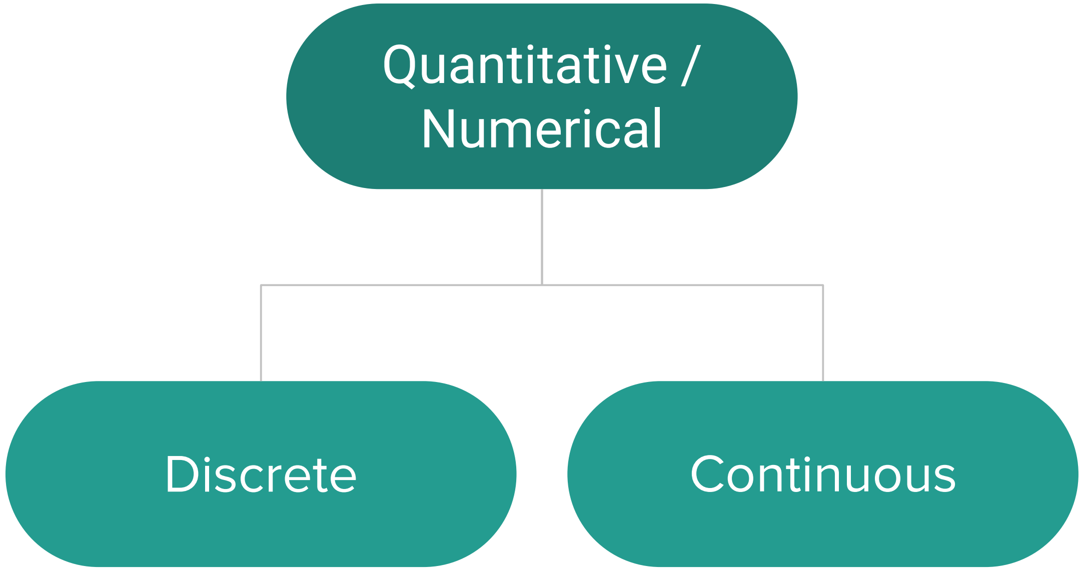
Continuous

Binary

Nominal

Ordinal

DEFINING DATA: TYPES OF VARIABLES

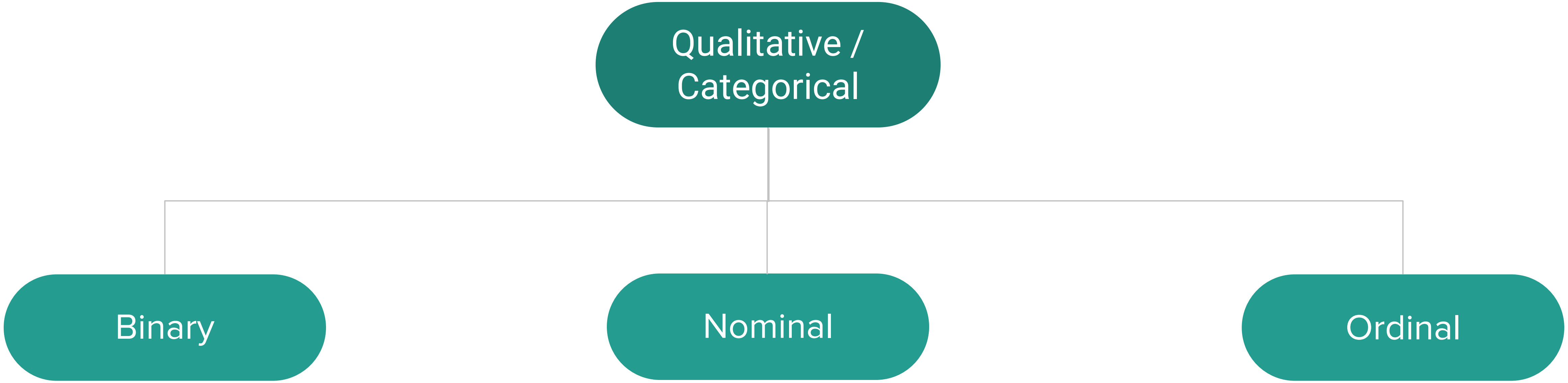


Quantitative / Numerical

Discrete

Continuous

DEFINING DATA: TYPES OF VARIABLES

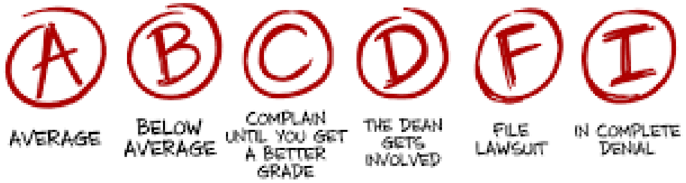


Qualitative / Categorical

Binary

Nominal

Ordinal



➔The rate at which data is being created is rapidly accelerating. The key for successful data analysis is taking data and forming actionable conclusions and insights.

➔Data Analysts do this by using a “workflow” to guide them through the process.

INTRODUCTION: TOOLS OF THE DATA ANALYST

TOOLS OF THE DATA ANALYST

- In order to proceed through the workflow, a data analyst uses a suite of tools to assist along the way.

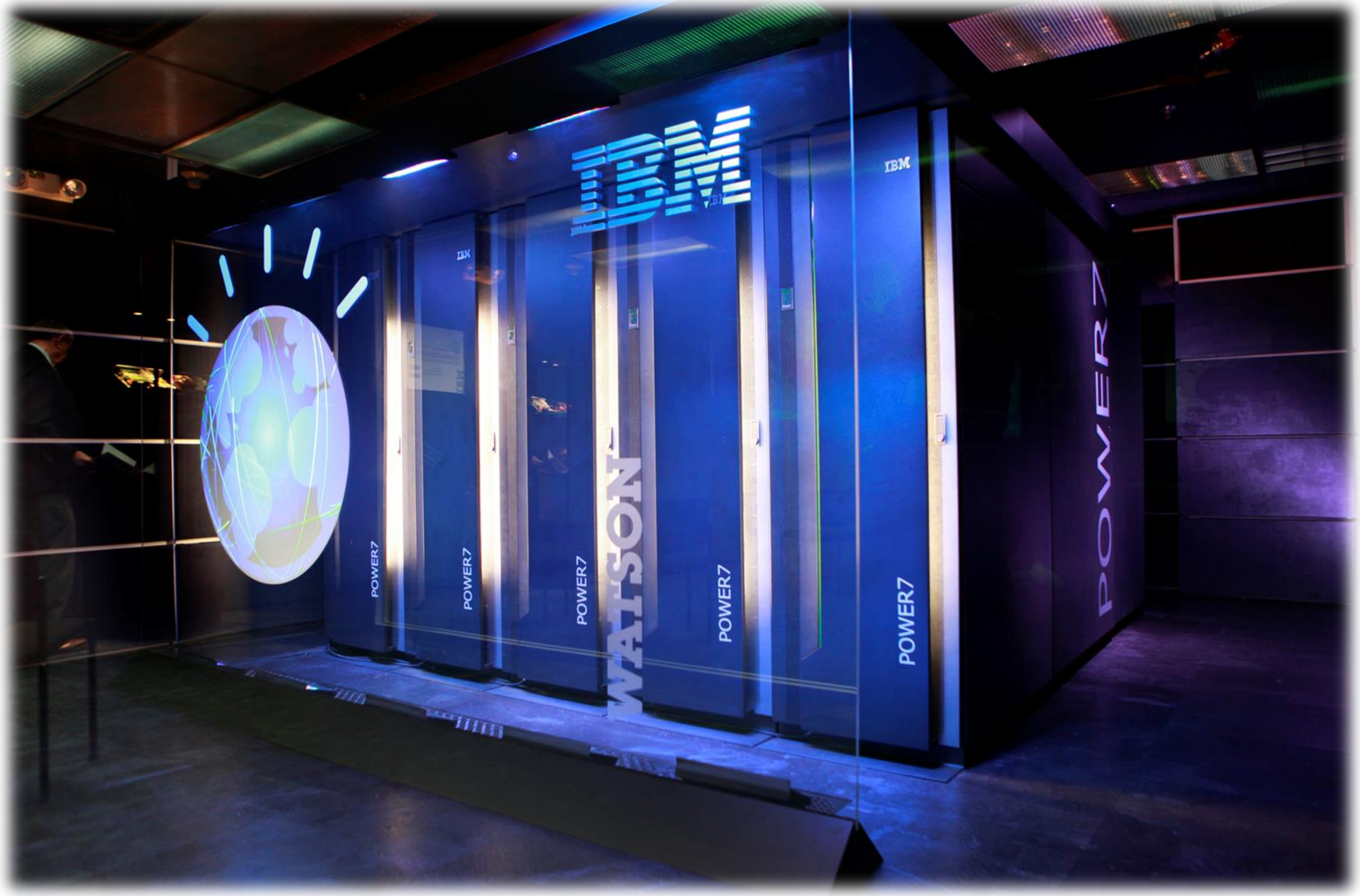
- *However, it's vital to remember:*
  - While data analysts can apply a specialized suite of tools, the analyst's *judgement and intuition* is the most important tool.

http://www.tylervigen.com/spurious-correlations

TOOLS OF THE DATA ANALYST

Is there a tool that will “Identify the Question” or “Understand the Data” for you?

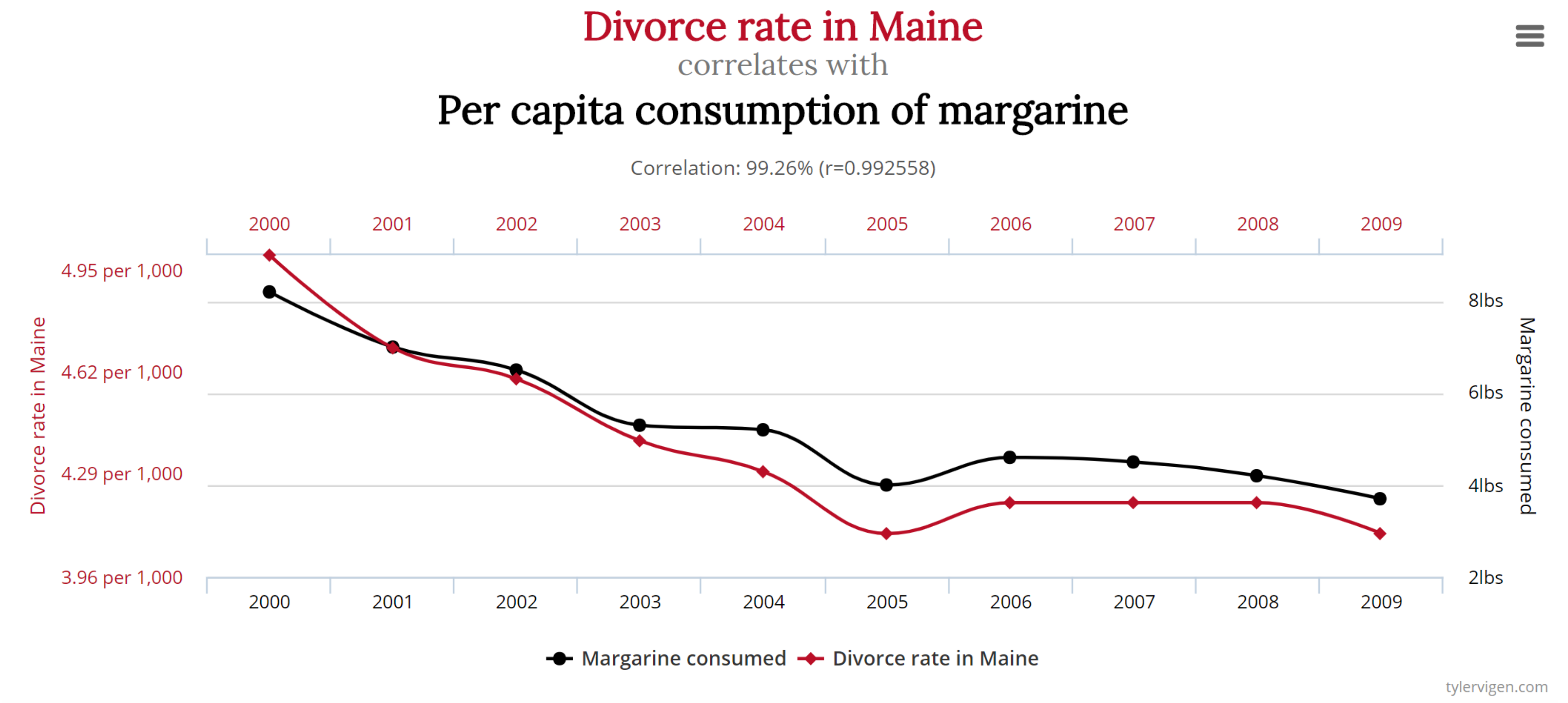
People are trying:
IBM Watson



TOOLS OF THE DATA ANALYST

- Is there a tool that will “Identify the Question” or “Understand the Data” for you?

Correlation is not causation



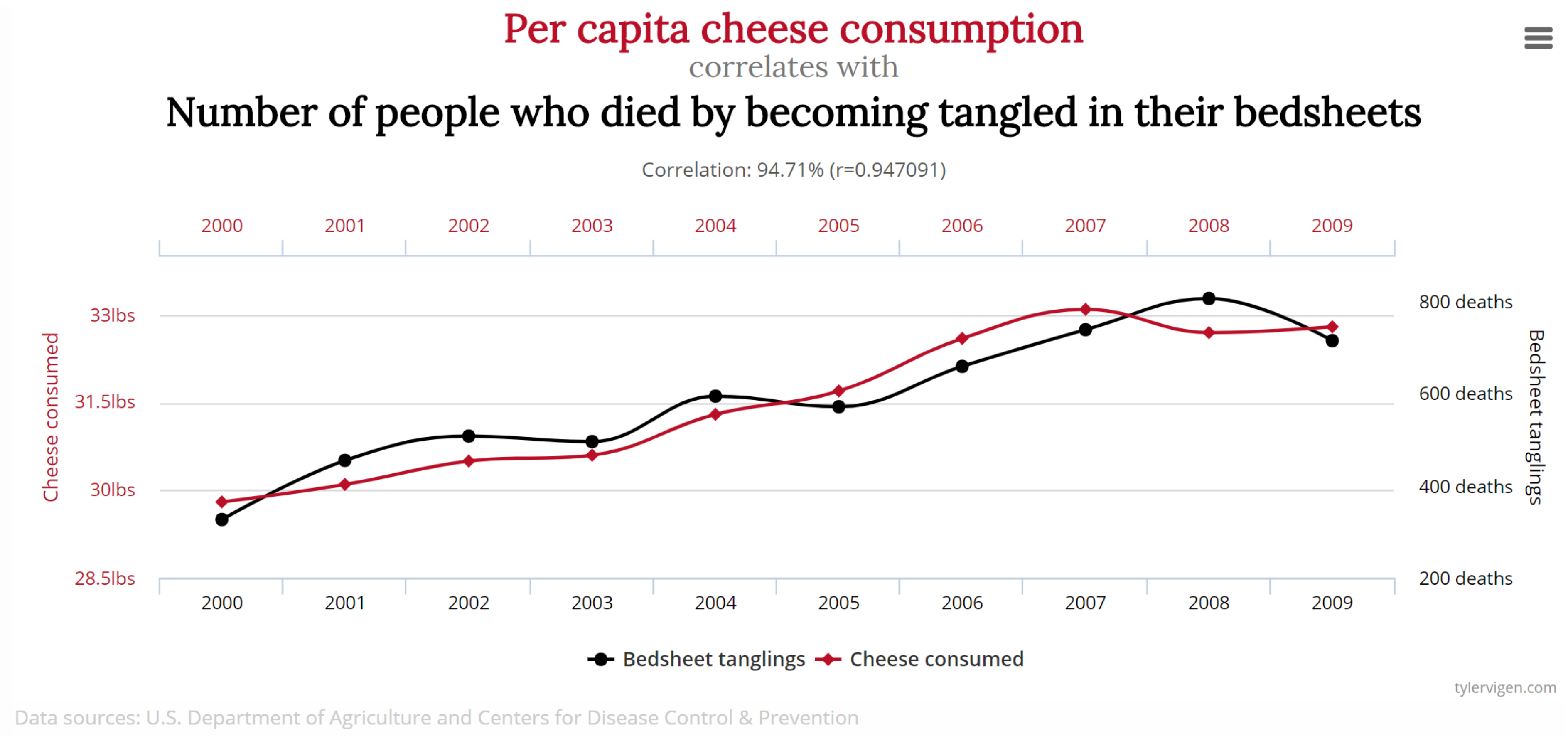
Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

http://www.tylervigen.com/spurious-correlations

TOOLS OF THE DATA ANALYST

- Is there a tool that will “Identify the Question” or “Understand the Data” for you?

Correlation is not causation



Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

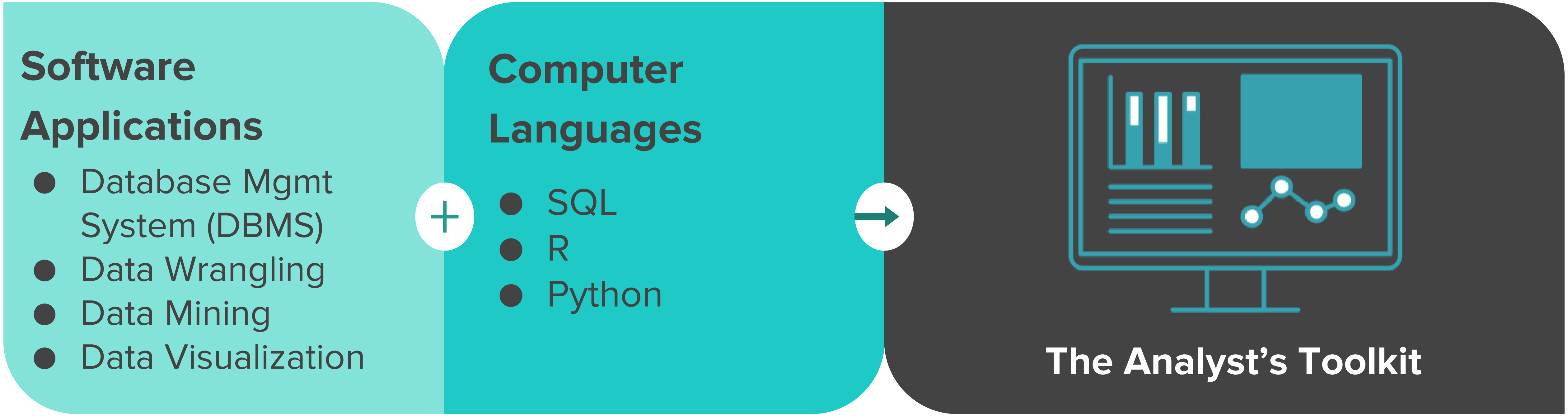
Correlation: 94.71% (r=0.947091)

Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

http://www.tylervigen.com/spurious-correlations

TOOLS OF THE DATA ANALYST

- No shortage of data tools to use for purposes that range from data cleaning to visualization.
- New tools and new versions of tools are constantly coming out.
- Some are highly specialized, some integrate with each other, and many are used in conjunction with each other.



Software Applications

- Database Mgmt System (DBMS)
- Data Wrangling
- Data Mining
- Data Visualization

+

Computer Languages

- SQL
- R
- Python

→

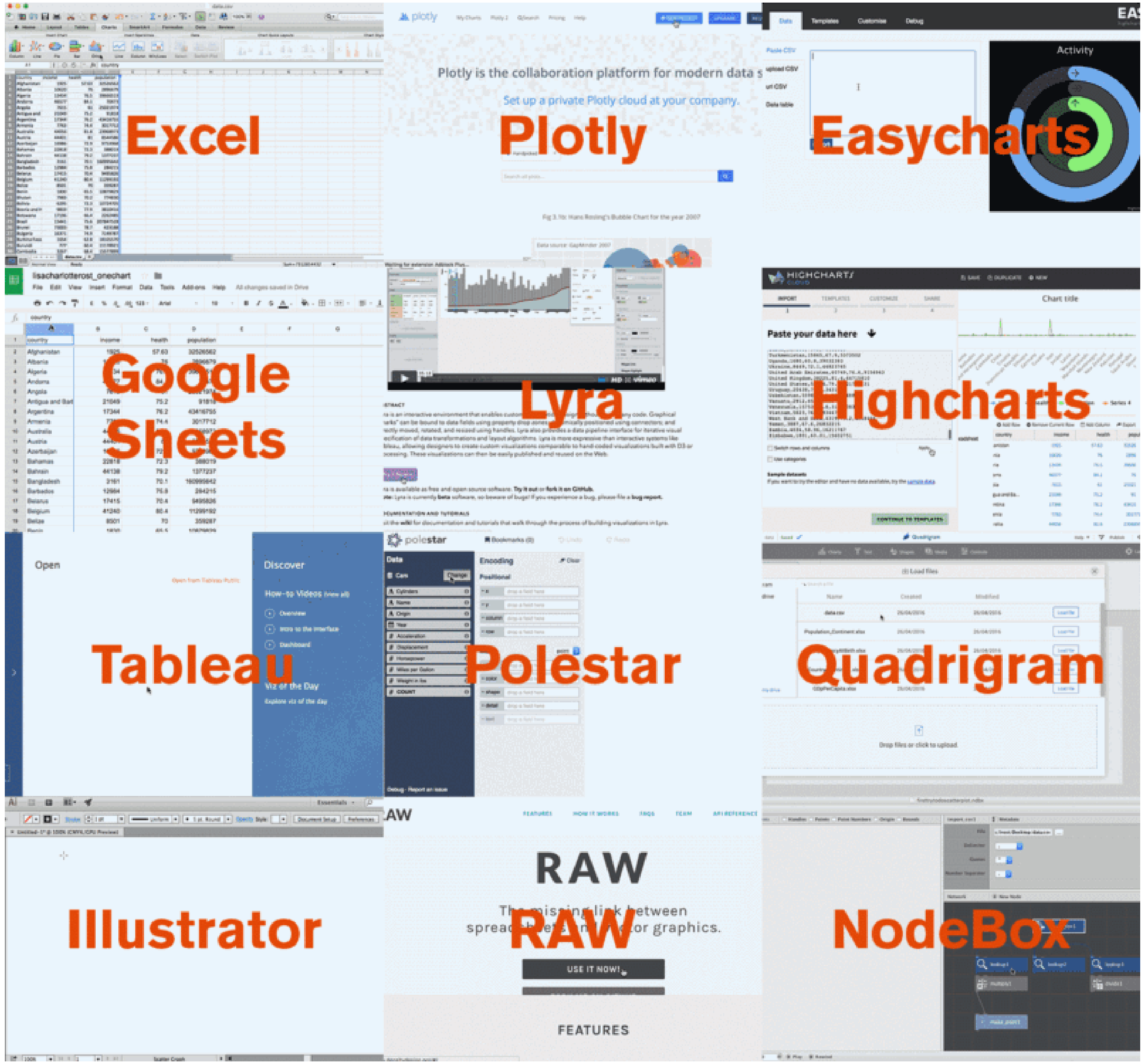
The Analyst's Toolkit

TOOLS OF THE DATA ANALYST

Tools	SQL	R	Python	Excel	Google Sheets	Tableau	Tableau Prep	Trifacta	Micrs. Power BI
Database Mgmt System (DBMS)	X								
Data Wrangling	X		X	X	X		X	X	
Data Mining		X	X	X	X	X			X
Data Visualization		X	X	X	X	X			X

TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

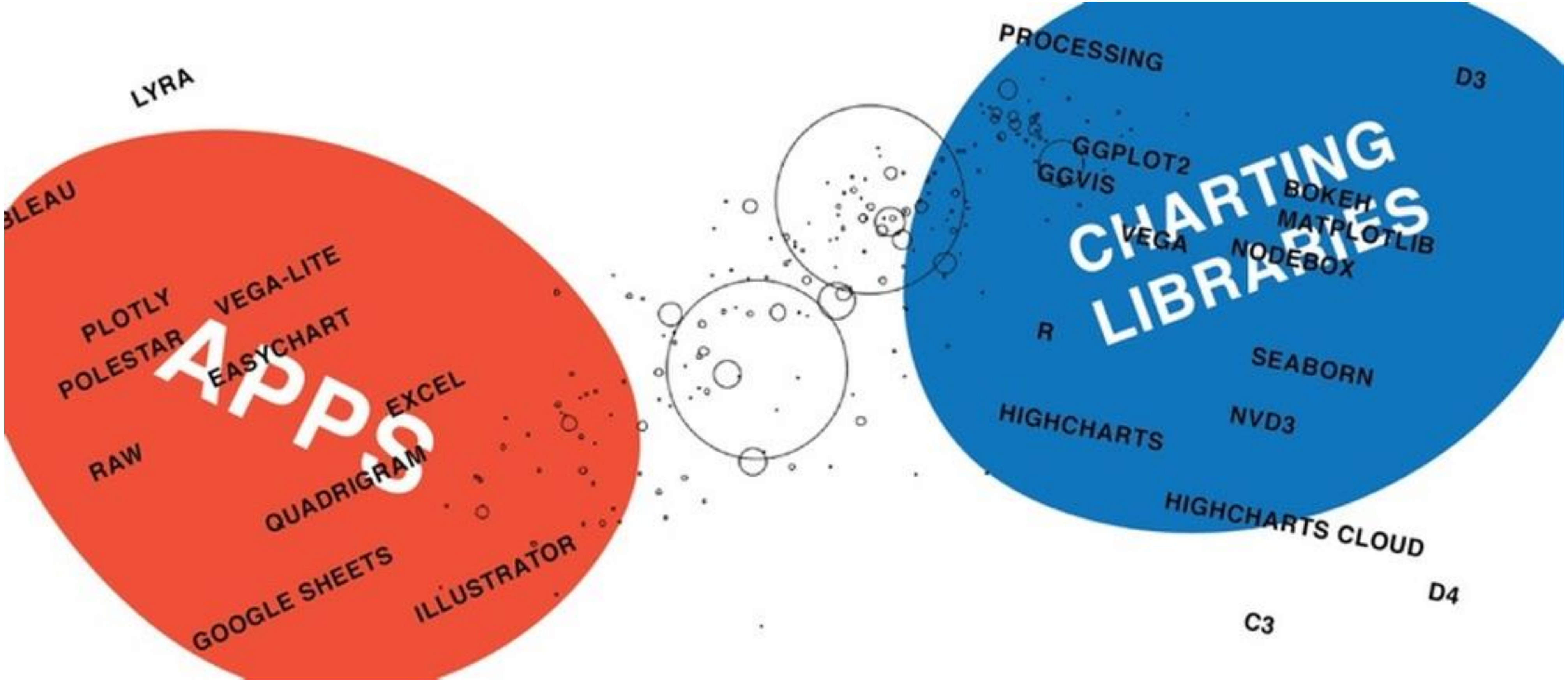
- Lisa Charlotte Rost, a fellow at National Public Radio, set out to make the same chart, using the same data, but by using multiple data tools.



https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

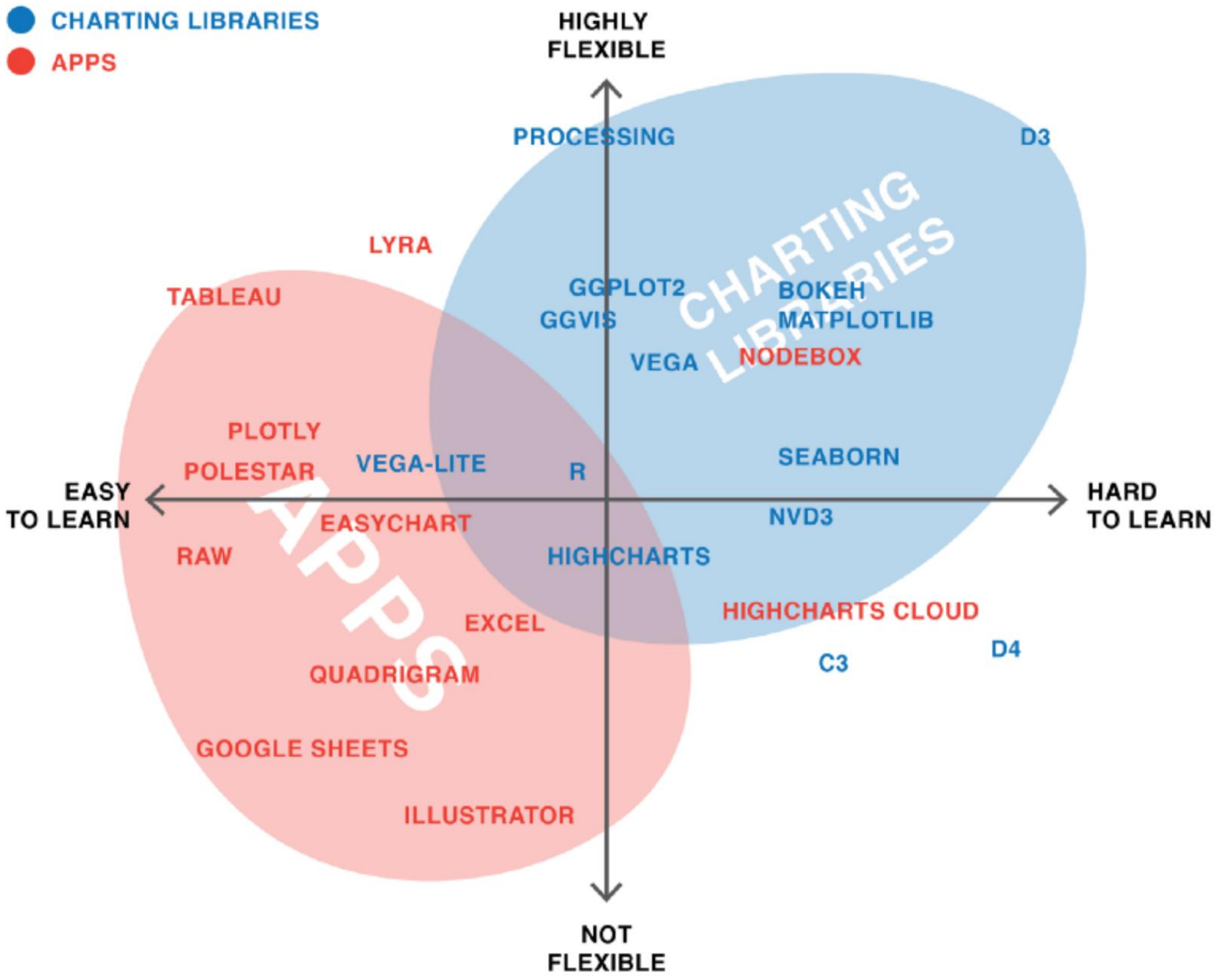
TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

Some comparisons



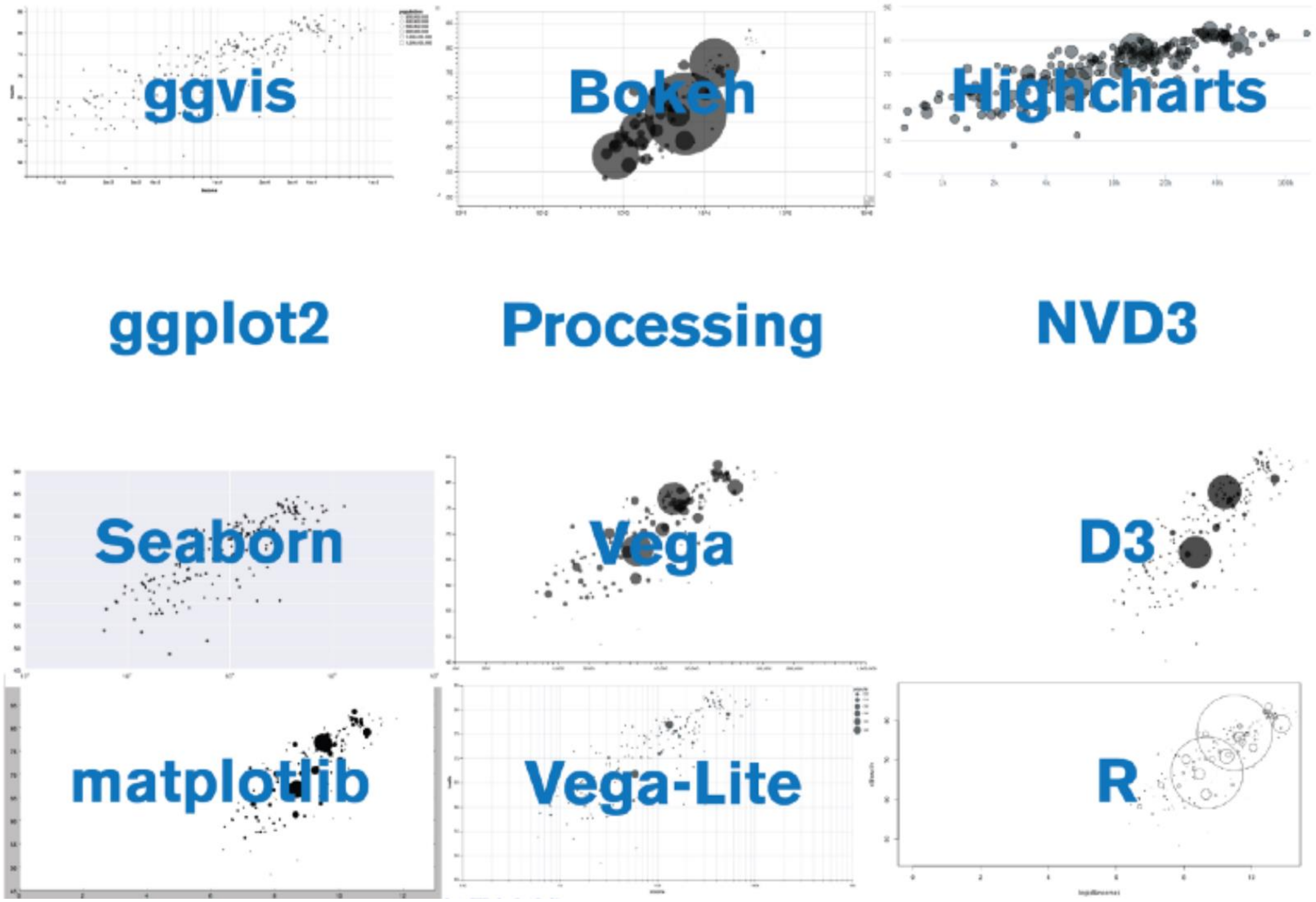
TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

Some comparisons (very subjective)



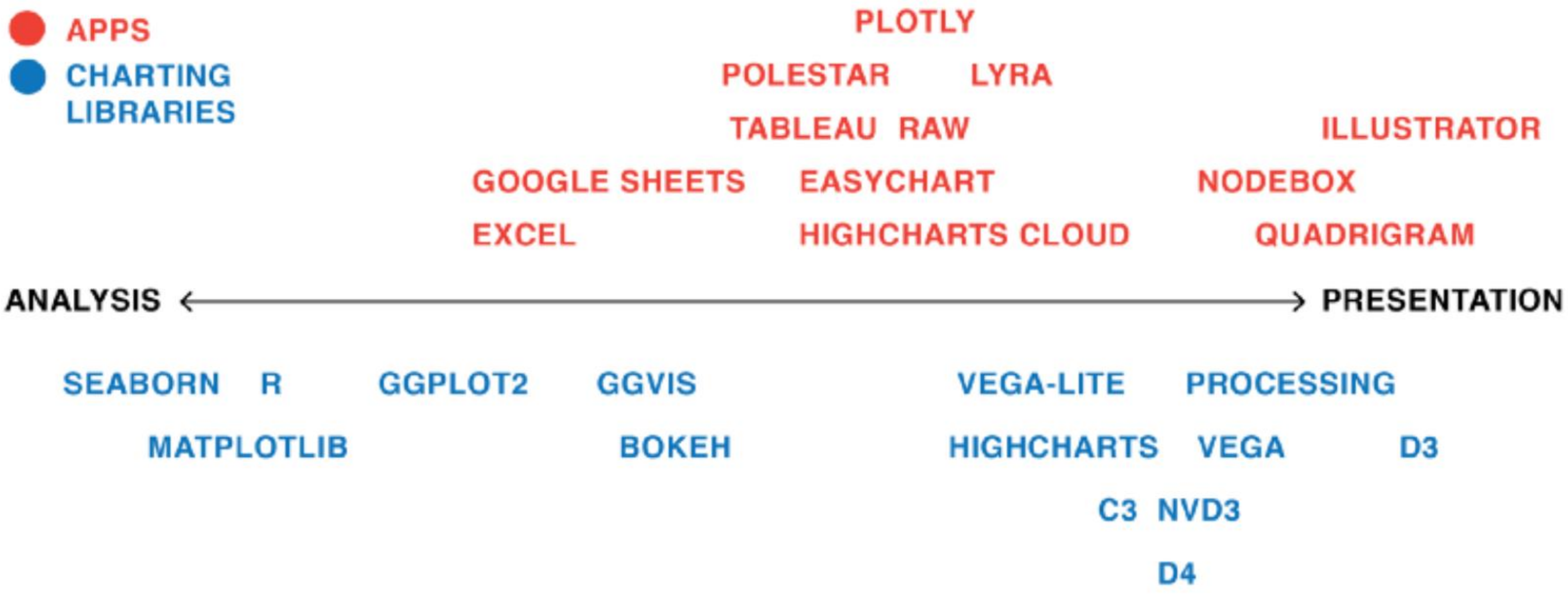
TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

Same chart, using the same data, but by using multiple data tools.



TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

Some comparisons (Analysis Vs. Presentation)



● APPS

● CHARTING LIBRARIES

PLOTLY

POLESTAR LYRA

TABLEAU RAW

ILLUSTRATOR

GOOGLE SHEETS EASYCHART

NODEBOX

EXCEL HIGHCHARTS CLOUD

QUADRIGRAM

ANALYSIS ⟷ PRESENTATION

SEABORN R GGPLOT2 GGVIS VEGA-LITE PROCESSING

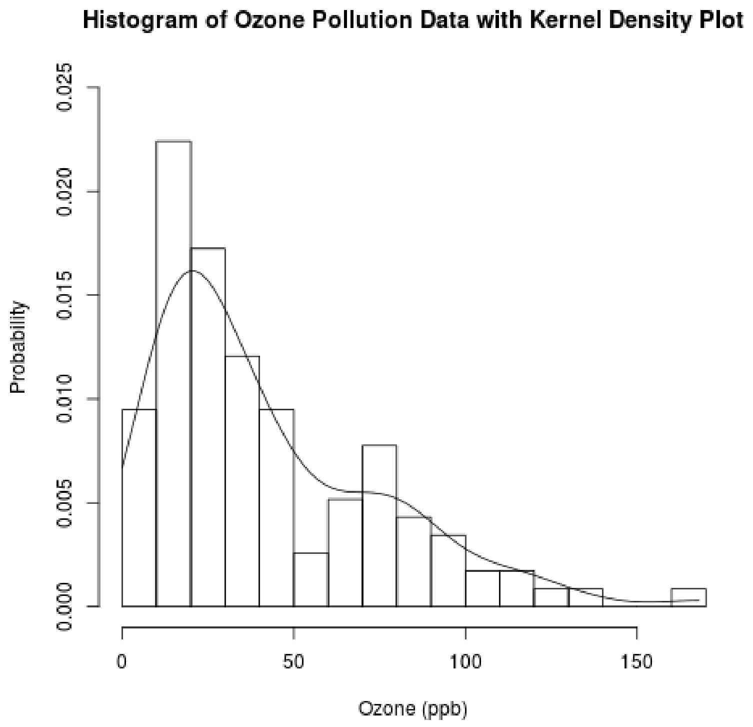
MATPLOTLIB BOKEH HIGHCHARTS VEGA D3

C3 NVD3

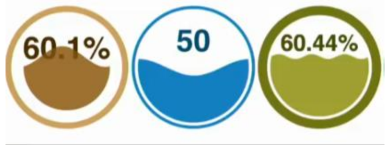
D4

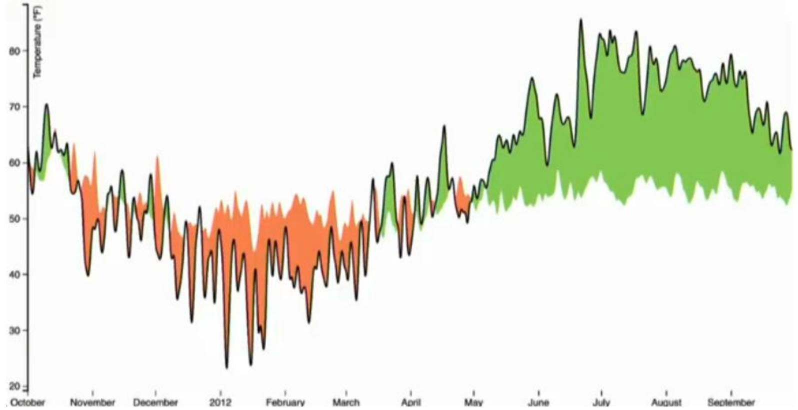
TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

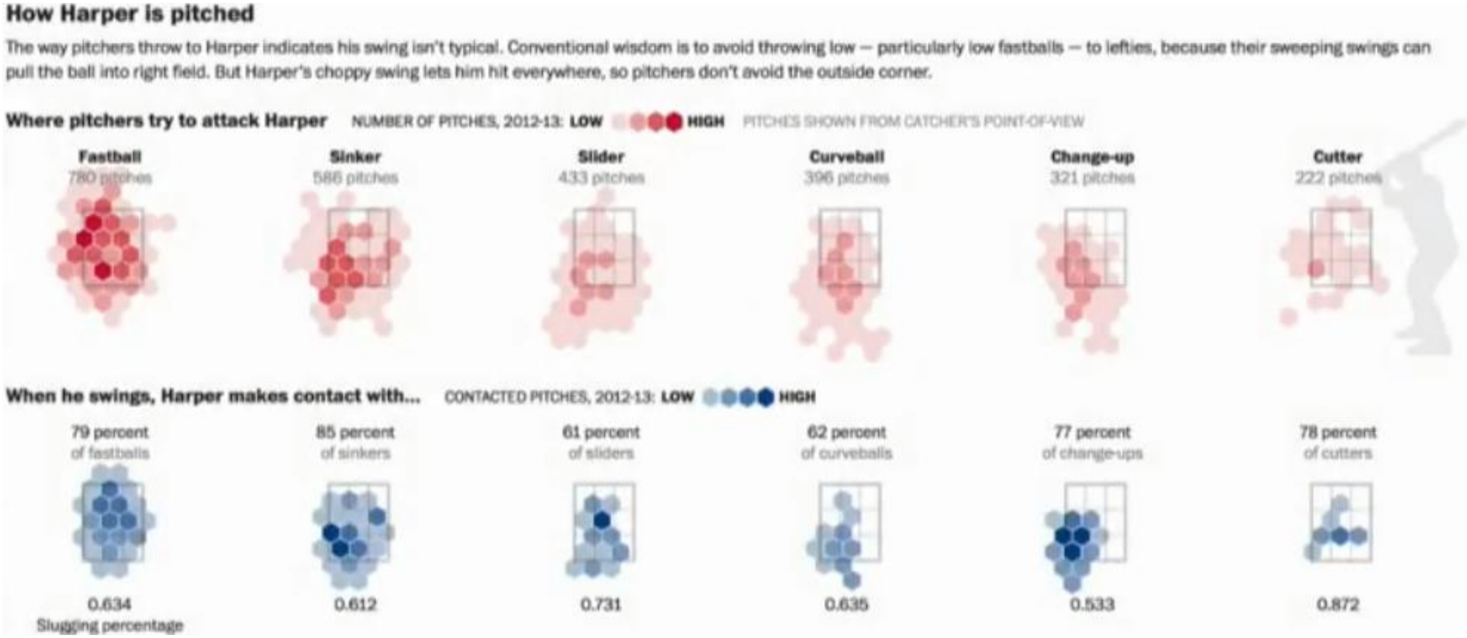
Some comparisons (Flexibility to manipulate charts R Vs. D3)



Vs.





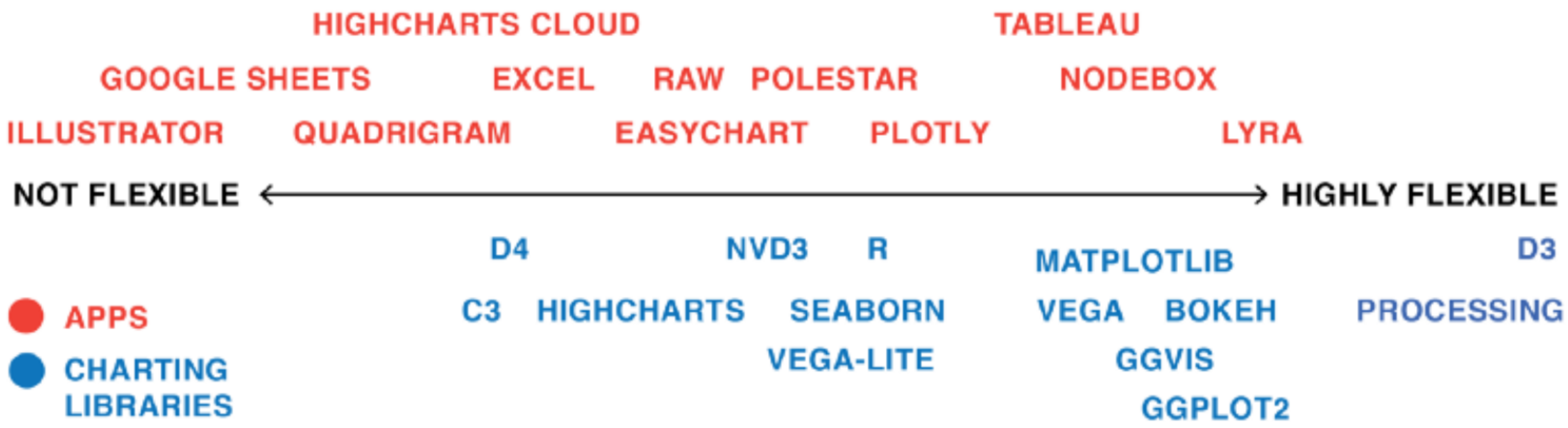


How Harper is pitched

The way pitchers throw to Harper indicates his swing isn't typical. Conventional wisdom is to avoid throwing low — particularly low fastballs — to lefties, because their sweeping swings can pull the ball into right field. But Harper's choppy swing lets him hit everywhere, so pitchers don't avoid the outside corner.

TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

Some comparisons (Flexibility to manipulate charts)



HIGHCHARTS CLOUD

TABLEAU

GOOGLE SHEETS

EXCEL

RAW

POLESTAR

NODEBOX

ILLUSTRATOR

QUADRIGRAM

EASYCHART

PLOTLY

LYRA

NOT FLEXIBLE ⟷ HIGHLY FLEXIBLE

D4

NVD3

R

MATPLOTLIB

D3

C3

HIGHCHARTS

SEABORN

VEGA

BOKEH

PROCESSING

VEGA-LITE

GGVIS

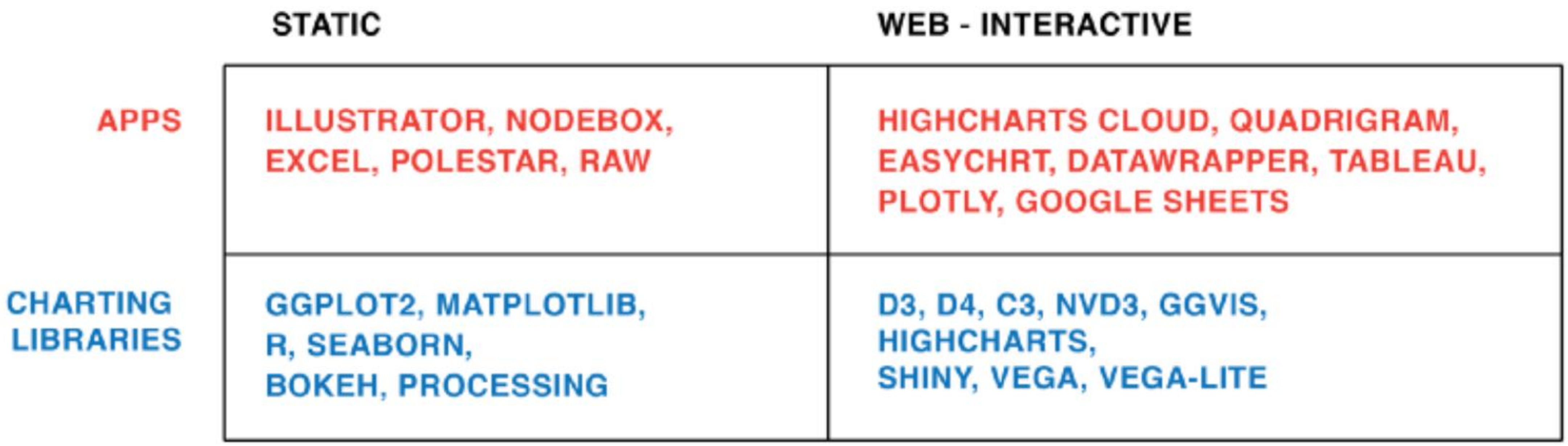
GGPLOT2

● APPS

● CHARTING LIBRARIES

TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

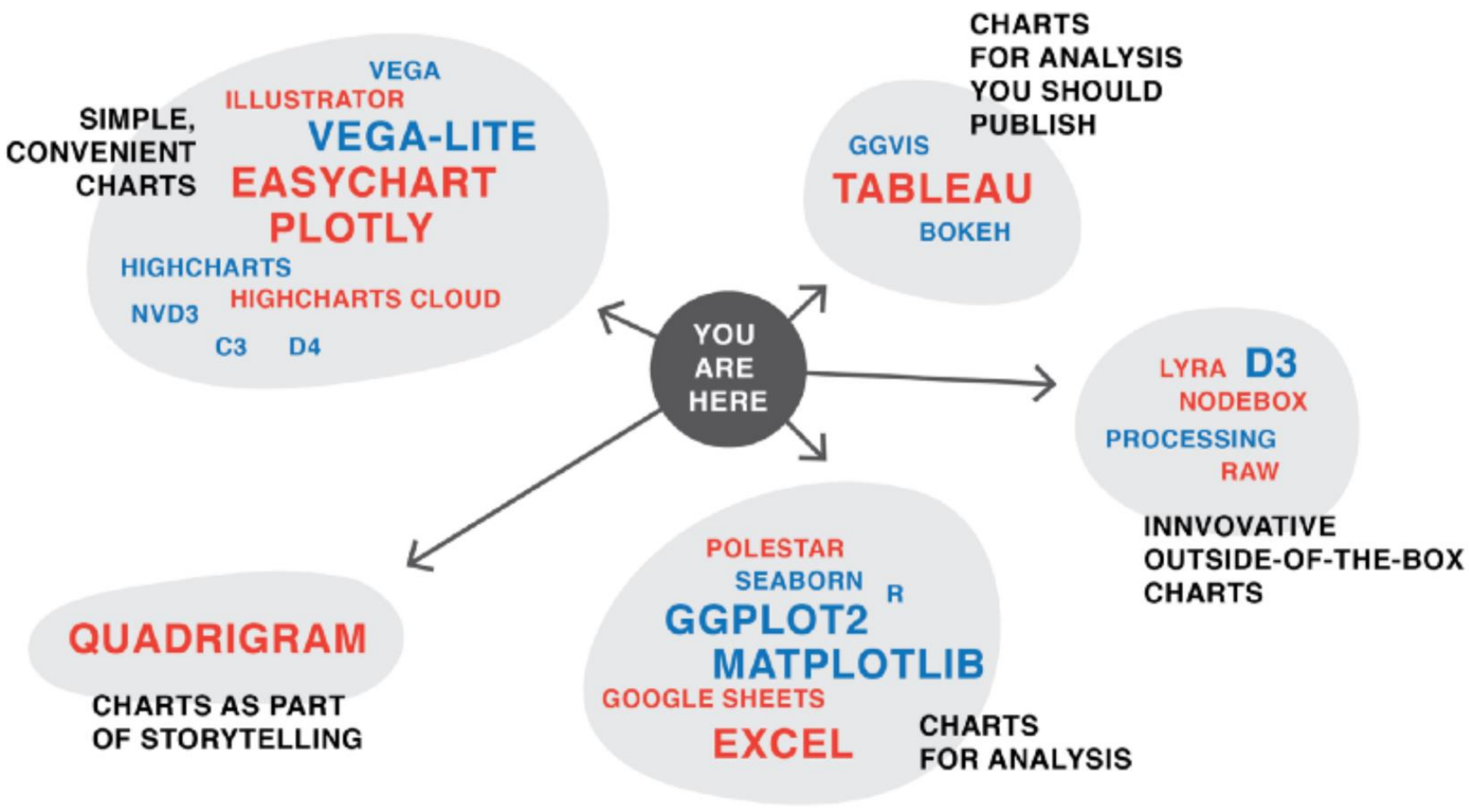
Some comparisons (Static or Interactive)



	STATIC	WEB - INTERACTIVE
APPS	ILLUSTRATOR, NODEBOX, EXCEL, POLESTAR, RAW	HIGHCHARTS CLOUD, QUADRIGRAM, EASYCHRT, DATAWRAPPER, TABLEAU, PLOTLY, GOOGLE SHEETS
CHARTING LIBRARIES	GGPLOT2, MATPLOTLIB, R, SEABORN, BOKEH, PROCESSING	D3, D4, C3, NVD3, GGVIS, HIGHCHARTS, SHINY, VEGA, VEGA-LITE

TOOLS OF THE DATA ANALYST: DATA VISUALIZATION

Some comparisons (Subjective results)



THE TOOL: GOOGLE SHEETS

➔Let's talk through some important vocabulary related to the Google Sheets environment:

- **Workbook:** A collection of worksheets.
- **Worksheet:** The area where data is arranged and calculations are performed.
- **Column:** A vertical collection of cells.
- **Row:** A horizontal collection of cells.
- **Cell:** The intersection of a column and row on a worksheet.
- **Array:** A collection of cells in a row, column, or across rows or columns.
- **Formula:** Starts with "=". Calculates the value of a single cell.
    - **Function**: A predefined formula (like SUM).

A COMPARISON: GOOGLE SHEETS and EXCEL

➔**DEMO:**

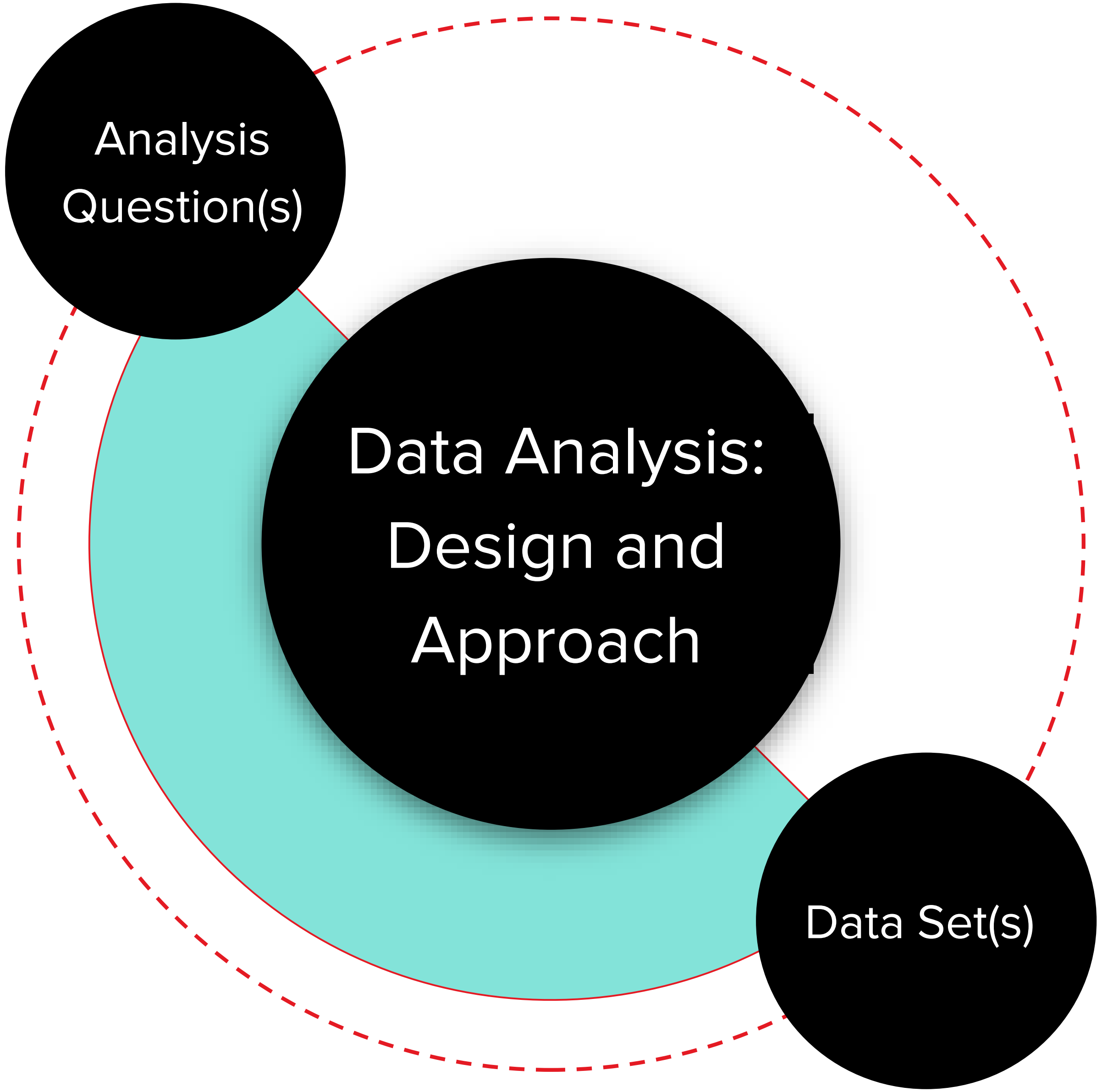
- Conditional formatting
  - Custom Formula
  - Color Scale
  - Duplicates
- Filtering by color
- Creating and running a macros

How do you choose which tools to learn and use?

- ➔The best tool depends on several factors:
  - Stage of the analytical workflow
  - Analysis goals
  - Budget
  - End Users / Learning Curve
- ➔Refer to Gartner's Magic Quadrants:
  - https://www.gartner.com/en/research/magic-quadrant
  - https://cdn2.hubspot.net/hubfs/2172371/Q1%202017%20Gartner.pdf?t=14962606

INTRODUCTION: HOW DATA ANALYSTS THINK ABOUT DATA

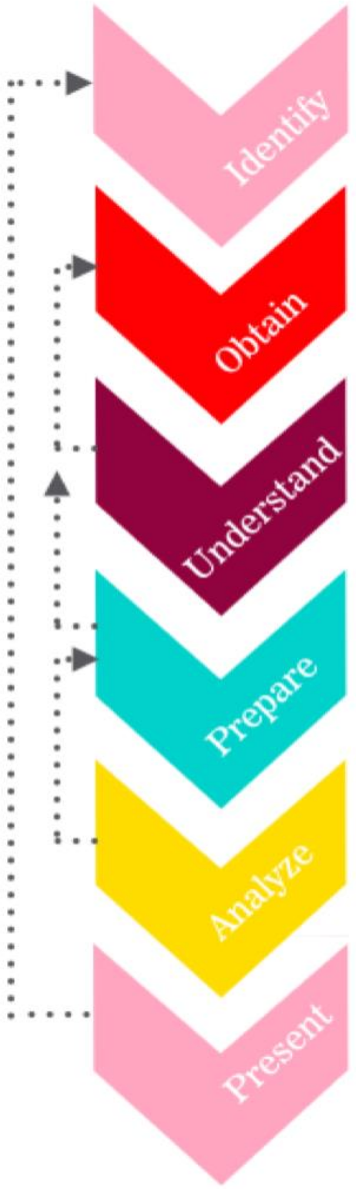
HOW DATA ANALYSTS THINK ABOUT DATA



Your analysis design will be based on:

- **the questions you are trying to answer**
- **the actual data you have available**

DATA ANALYSIS WORKFLOW



IDENTIFY THE PROBLEM

OBTAIN THE DATA

UNDERSTAND THE DATA

PREPARE THE DATA

ANALYZE THE DATA

PRESENT THE RESULTS

DATA ANALYSIS WORKFLOW

These general steps are necessary for each and every data analysis you do. ***However***, each time will be a little different, as well.

Project A

Sometimes the question is clearly defined already, but obtaining the data you need is difficult or even impossible.

Project B

Other times, the data is readily available but in such a difficult format that you will spend the majority of your time cleaning it before your analysis.

Project C

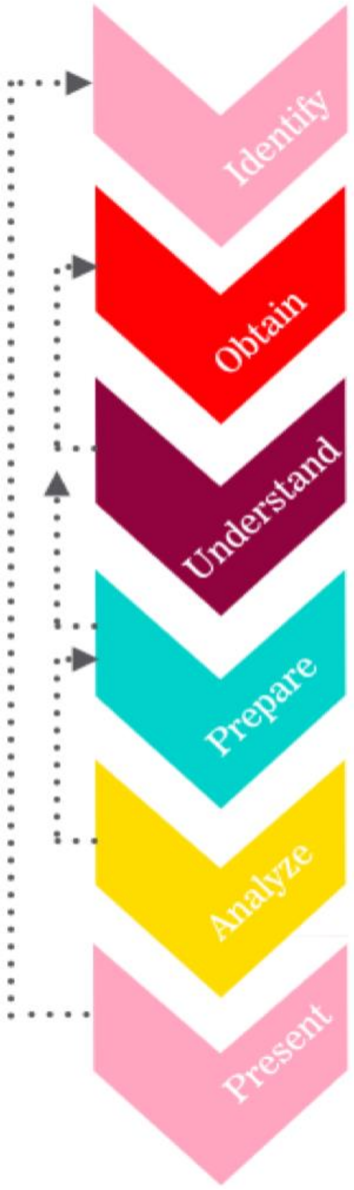
How you share your results will depend largely on your audience. How data savvy are they? How much time do they have to understand the data?

GUIDED PRACTICE: DATA ANALYTICS WORKFLOW

DATA ANALYSIS WORKFLOW

➔**Identify the Problem**

- Before you begin working with any data, you must understand the problem that you're trying to answer:
  - Specific Questions
  - Measures
  - Goals / Objectives



IDENTIFY THE PROBLEM

OBTAIN THE DATA

UNDERSTAND THE DATA

PREPARE THE DATA

ANALYZE THE DATA

PRESENT THE RESULTS

USING DATA TO INFORM BUSINESS DECISIONS

- It’s time to get our hands dirty working with real data.
- Business Context:
  - We are a consulting company -- *GA Consultants*
  - The Client: WineMag.com
    - Client has a dataset with wine reviews from their Wine Enthusiast Magazine.
    - Observed that a significant amount of their web traffic are readers who follow specific wine reviewers.
    - Client wants to utilize data set to summarize and report information on their editors (the wine reviewers): https://www.winemag.com/editors/
  - Sample of web-scraped data
  - Source: https://www.kaggle.com/zynicide/wine-reviews

IDENTIFY THE PROBLEM

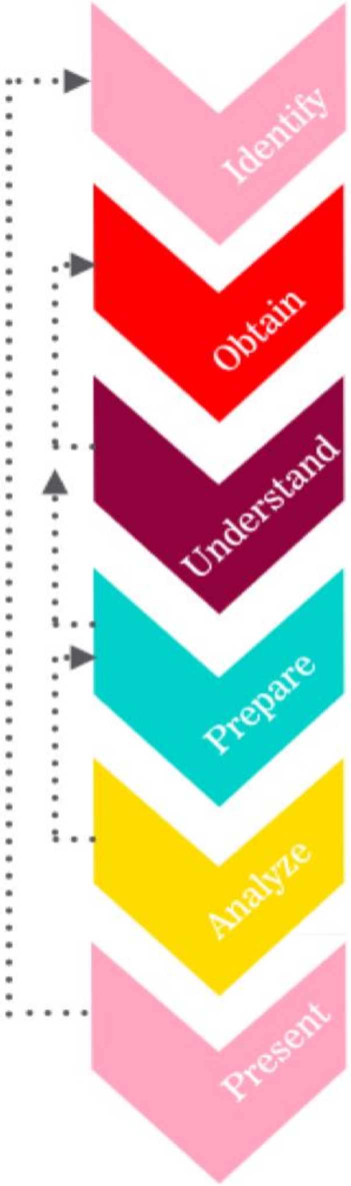
- Business Goals:
    - Summarize and report information on their contributing editors (the wine reviewers)
    - Cater content to the readers who follow specific wine experts (in order to build this readership base)

ACTIVITY

- We have a clear direction on which “business” questions to focus our analysis on, but we will want to provide insights based on “data” questions:
    - What fields does the data set contain?
    - Which information is useful for readers who follow the wine reviewers?
    - Can this information also be useful for non-followers?

DATA ANALYSIS WORKFLOW

- **Obtain the Data**
  - To work with the data, you first have to find it or collect it, and it has to be the **right data** to help you answer the question.



IDENTIFY THE PROBLEM

OBTAIN THE DATA

UNDERSTAND THE DATA

PREPARE THE DATA

ANALYZE THE DATA

PRESENT THE RESULTS

OBTAIN THE DATA

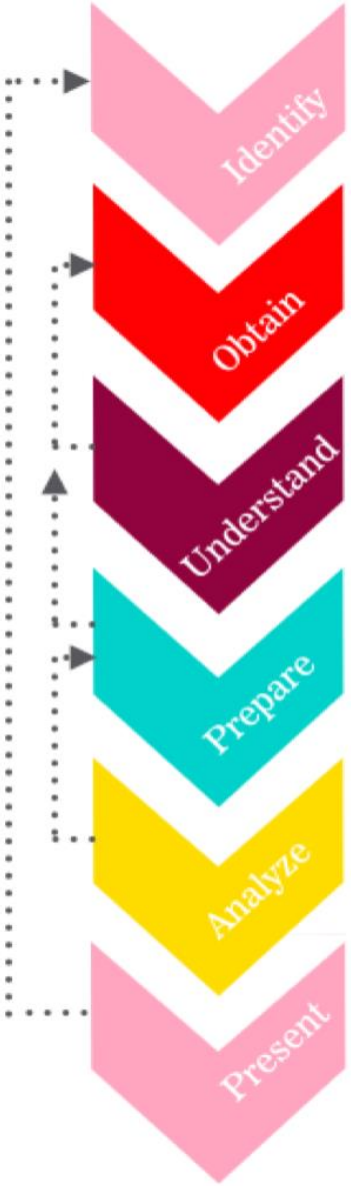
➔Let’s access the Google Sheet we will be using for the remainder of today.

- You must be logged into a Google account.
- Visit this link: goo.gl/wLfv6g (case sensitive!) This has the frame for Data Dictionary. Use emailed link.
- Next, you’ll need to click on “File” and “Make a copy…” to create a version of the file on your own Google Drive.
  - This Google Sheet is in “**View only**” mode, so you **can’t edit** it.
- When you **copy the file to** your Google Drive, you’ll **have full edit access** to the copied version.

GUIDED PRACTICE: UNDERSTAND THE DATA

DATA ANALYSIS WORKFLOW

- **Understand the Data**
  - Confirm what can and can’t be measured and which questions can be answered.
  - Then you need to ensure you can correctly interpret the results and trust the data.
    - Data Types
    - Missing Values
    - *Suspicious* Data



IDENTIFY THE PROBLEM

OBTAIN THE DATA

UNDERSTAND THE DATA

PREPARE THE DATA

ANALYZE THE DATA

PRESENT THE RESULTS

UNDERSTAND THE DATA

➔First, what do we know about this data?

- What are the data fields?
- What is the unit of observation *(the unit described by the data that one analyzes)*?
- What types of data and variables do we have: Numerical? Categorical? Strings?
- Is there any missing data?
- What data ranges do we have?

Let’s create a Data Dictionary!

UNDERSTAND THE DATA

ACTIVITY

➔Create a Data Dictionary with the following information:

- Field names *(columns)*
- Description of information in column
- Data Type
- Total # of Rows
- # of Unique Values
- Any missing data?
- What data ranges do we have?

REVISITING: IDENTIFY THE PROBLEM

➔Business Goals:

- Summarize and report information on their contributing editors (the wine reviewers)
- Cater content to the readers who follow specific wine experts (in order to build this readership base)

ACTIVITY

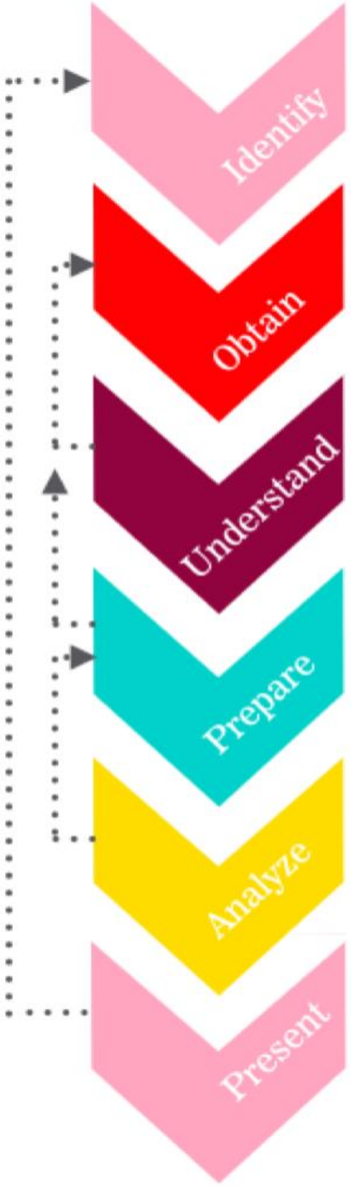
- **Now that we know what data we have, let’s get specific:**
    - What are the distributions of reviews by score and price?
    - Is there a relationship between wine score and bottle price?
    - Who are the top contributing wine experts?
        - Overall
        - By wine variety
        - By country (of wine)
        - By Price Categories
        - By Score Categories

GUIDED PRACTICE: PREPARE THE DATA

DATA ANALYSIS WORKFLOW

➔**Prepare the Data**

- ◆You should make sure the data doesn’t contain incorrect or missing values.
- ◆Structure and content of the data table(s).



IDENTIFY THE PROBLEM

OBTAIN THE DATA

UNDERSTAND THE DATA

PREPARE THE DATA

ANALYZE THE DATA

PRESENT THE RESULTS

PREPARE THE DATA

➔Empty cells in the dataset may represent missing information or actually indicate ‘zero’ or ‘none’.

➔How do we address missing data?

- **Step 1:** Research the reason for the omission of data. Is there actually no data available? Was it a data entry error?
- **Step 2:** Make a decision on how to correct the omission: Leave it, delete it, or change it (edit to correct value, use imputation, etc.)

➔Unit of Analysis: Wine Experts

- Remove rows where data is missing for field *taster_name*

Let’s address missing values!

PREPARE THE DATA

➔Remove columns not needed for analysis.

➔Create a column for score_category and price_category using the following functions in Google Sheets:

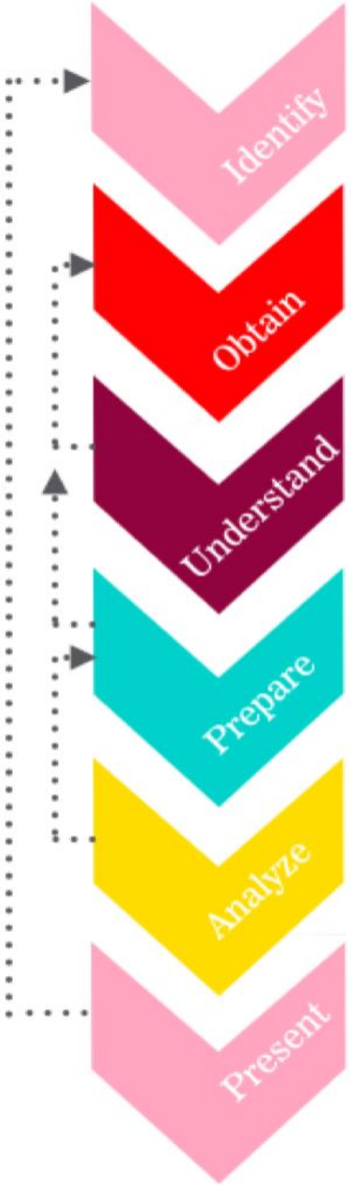
- IF
- ISBLANK
- Conditional Formatting

Let’s create calculated fields!

GUIDED PRACTICE: ANALYZE THE DATA

DATA ANALYSIS WORKFLOW

- ➔**Analyze the Data**
  - ◆Now, you are ready to uncover the answer to your questions, assuming you haven't ended up at a prior step due to missing data or a poorly understood question.
  - ◆Insights from the data to help answer your analysis/business questions or support recommendations.



IDENTIFY THE PROBLEM

OBTAIN THE DATA

UNDERSTAND THE DATA

PREPARE THE DATA

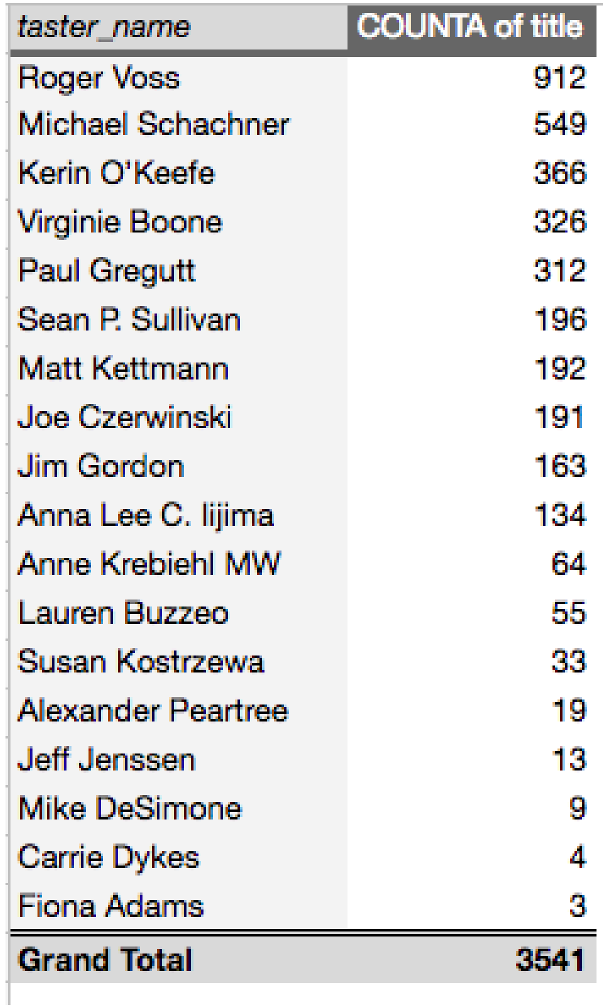
ANALYZE THE DATA

PRESENT THE RESULTS

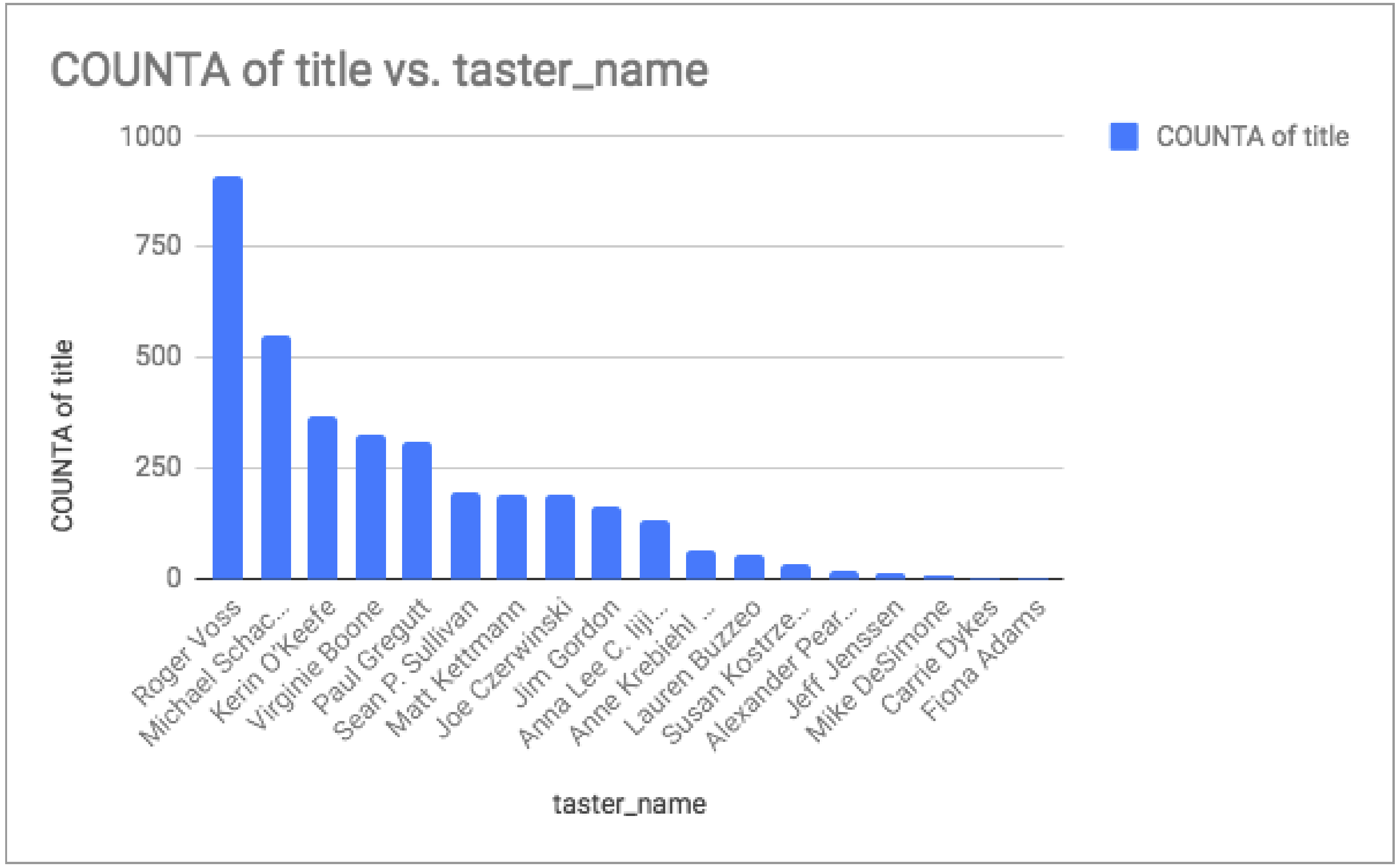
For each wine expert, how many total number of reviews?

- ➔ Pivot Table
- ➔ Bar Chart
- ➔ Stacked Bar Chart

For each wine expert, how many total number of reviews?

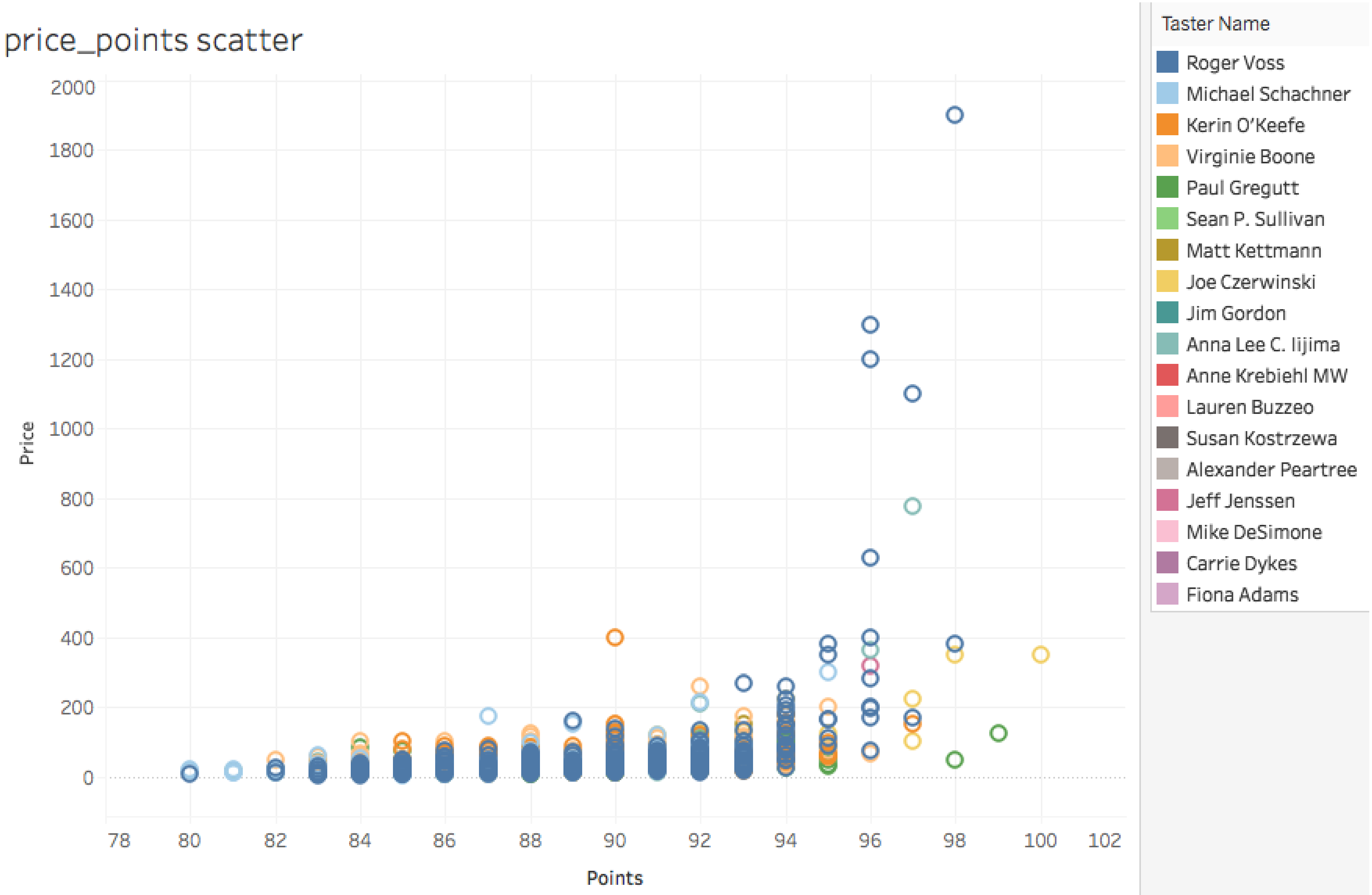


taster_name	COUNTA of title
Roger Voss	912
Michael Schachner	549
Kerin O'Keefe	366
Virginie Boone	326
Paul Gregutt	312
Sean P. Sullivan	196
Matt Kettmann	192
Joe Czerwinski	191
Jim Gordon	163
Anna Lee C. Iijima	134
Anne Krebiehl MW	64
Lauren Buzzeo	55
Susan Kostrzewa	33
Alexander Peartree	19
Jeff Jenssen	13
Mike DeSimone	9
Carrie Dykes	4
Fiona Adams	3
Grand Total	**3541**

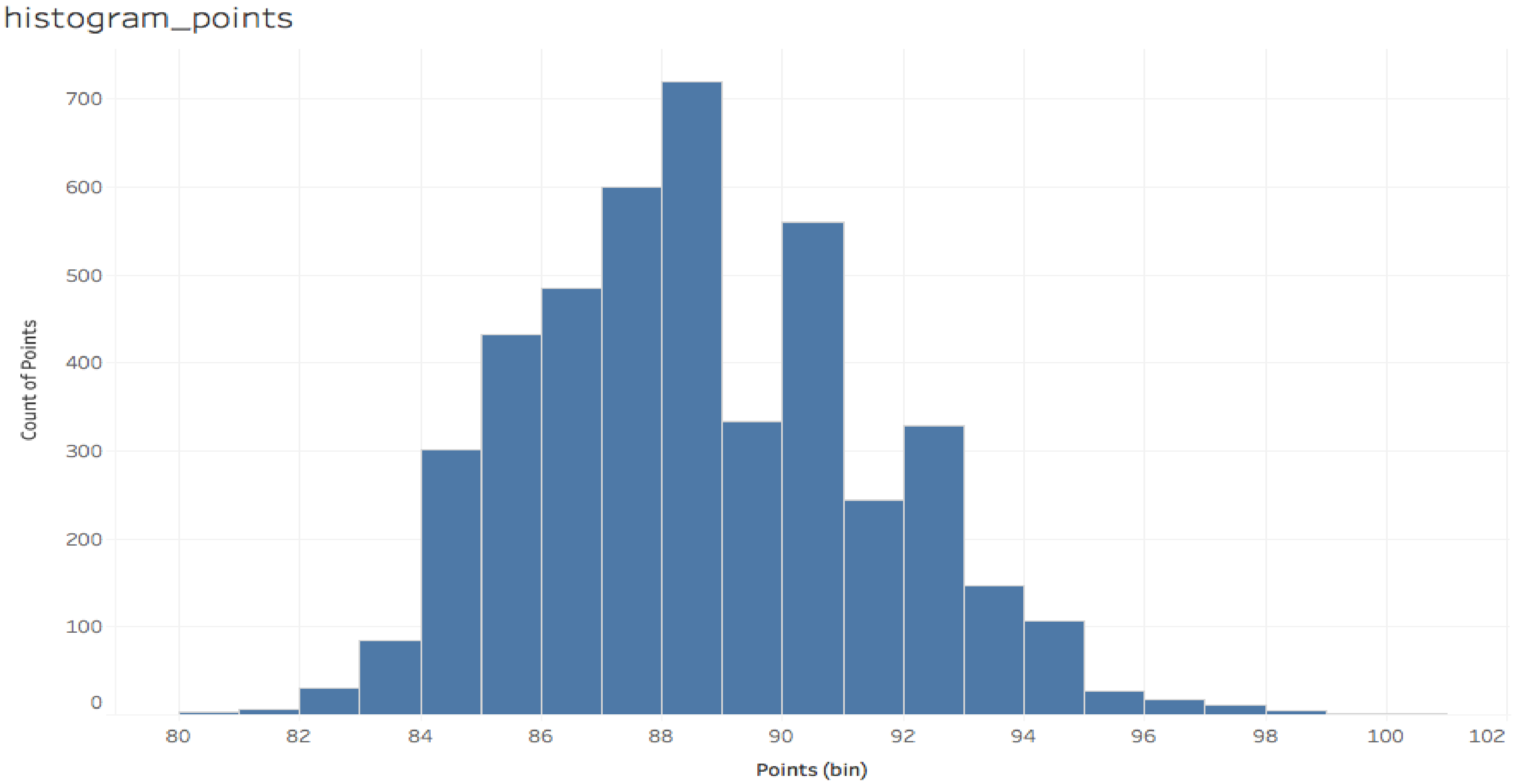


INTERPRET YOUR FINDINGS

- → Insight
- → Follow-up Questions

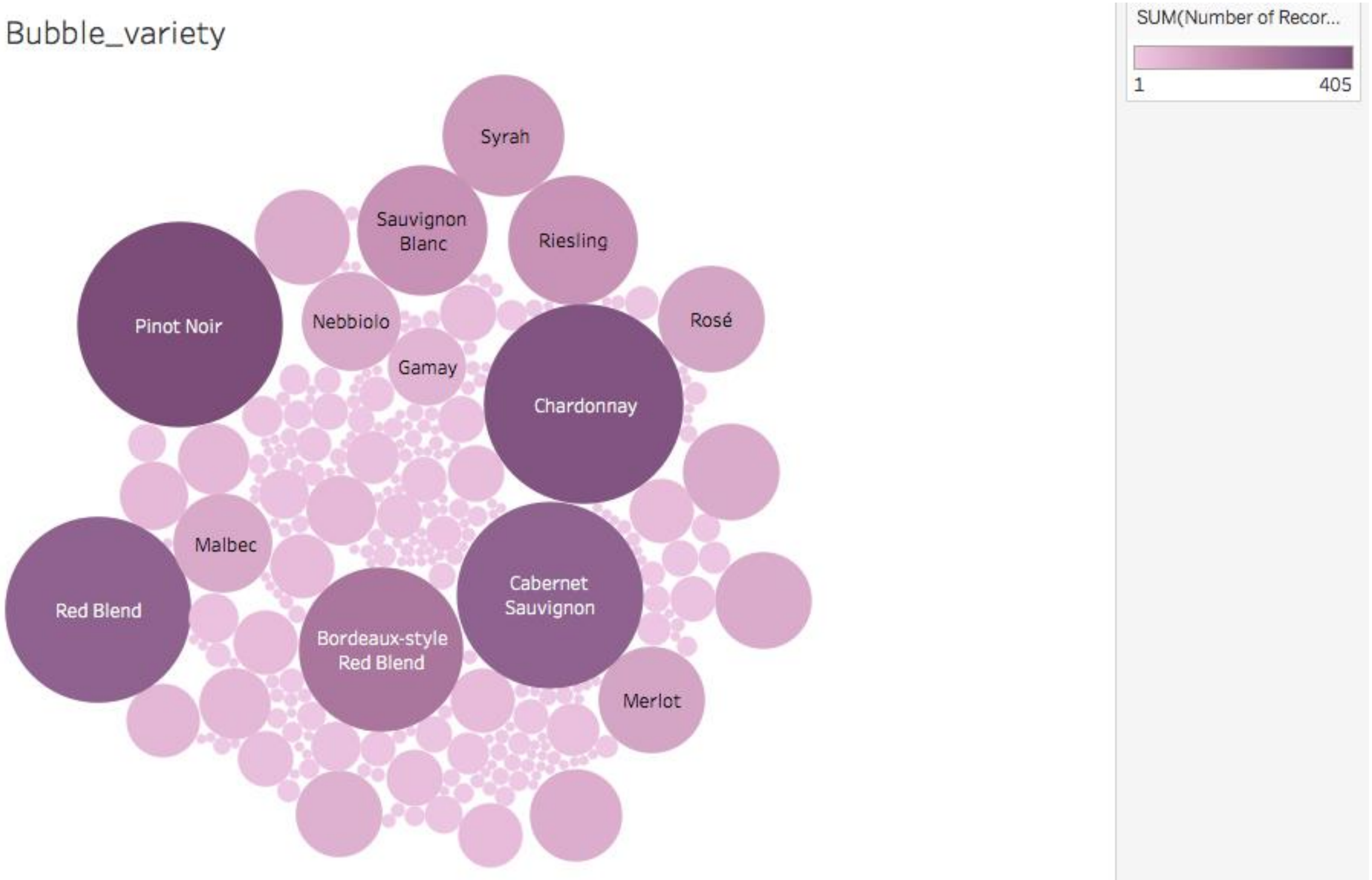


ANALYZING WINE REVIEWS

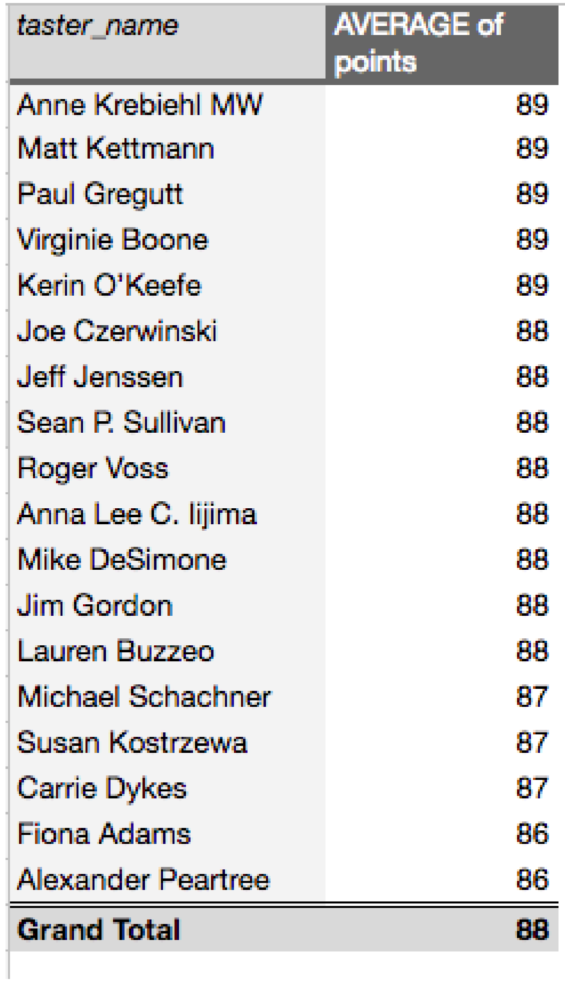


ANALYZING WINE REVIEWS

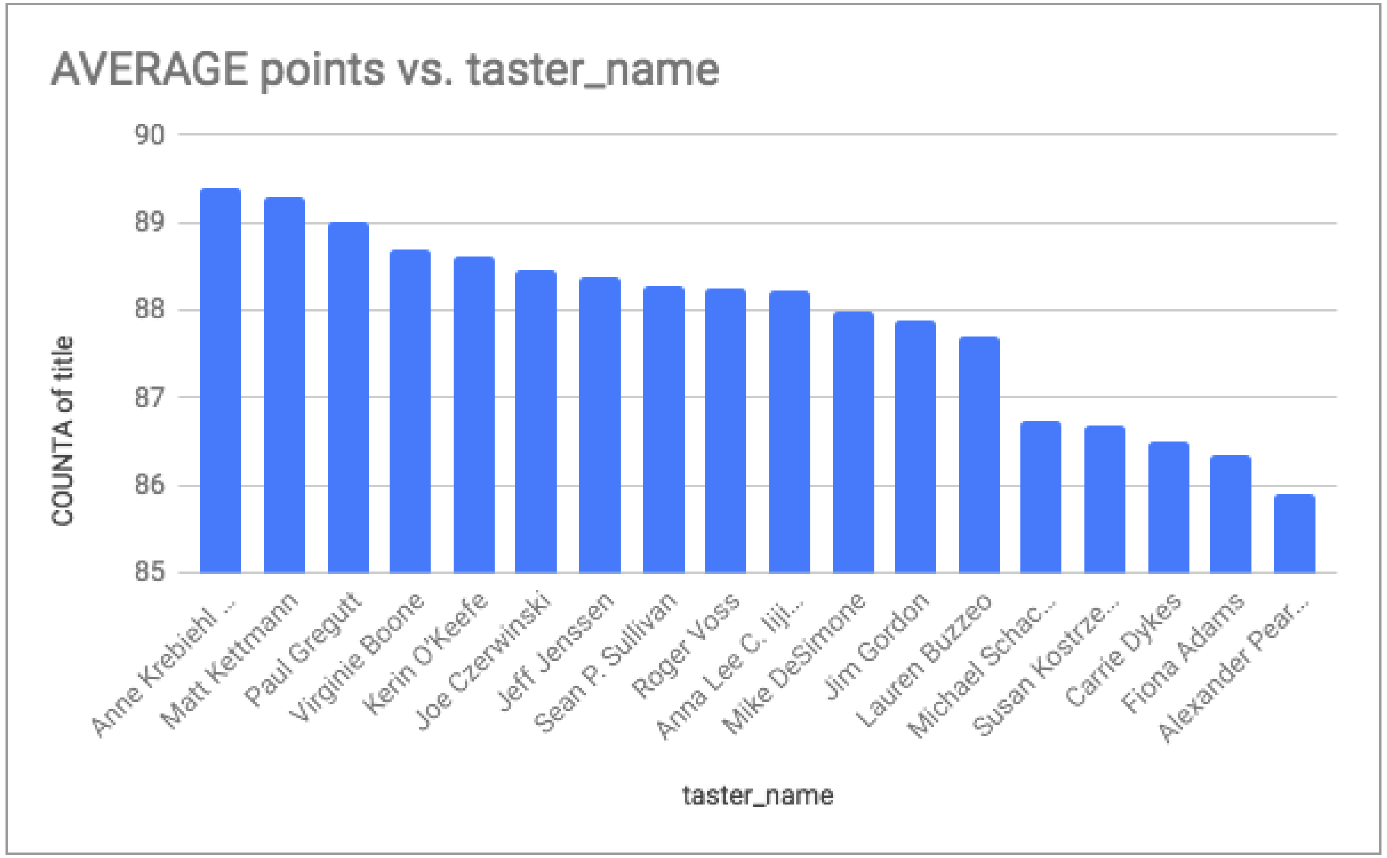
- ➔ Insight
- ➔ Follow-up questions



What was the average score given by each wine expert for all the bottles they reviewed?

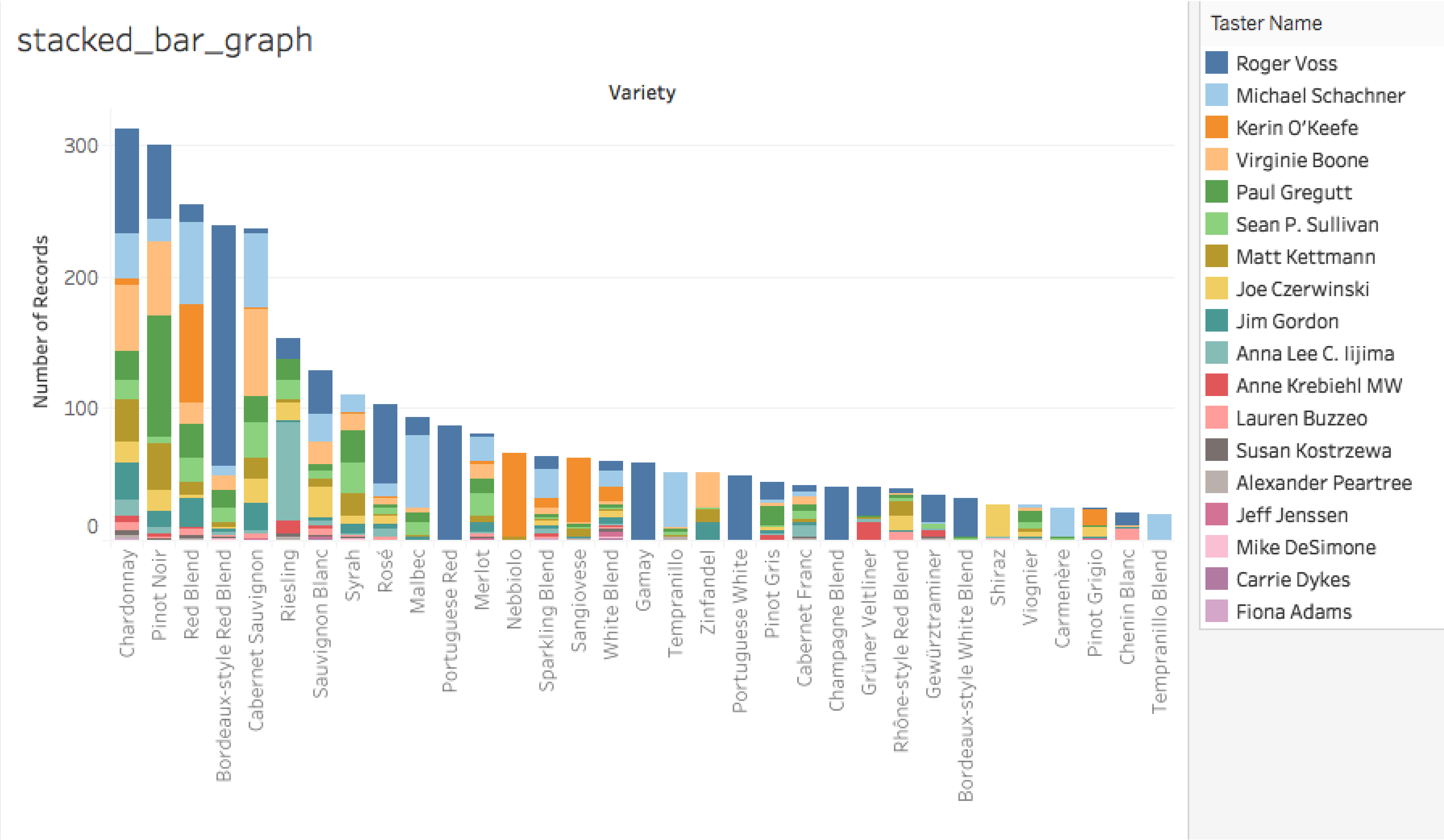


taster_name	AVERAGE of points
Anne Krebiehl MW	89
Matt Kettmann	89
Paul Gregutt	89
Virginie Boone	89
Kerin O'Keefe	89
Joe Czerwinski	88
Jeff Jenssen	88
Sean P. Sullivan	88
Roger Voss	88
Anna Lee C. Iijima	88
Mike DeSimone	88
Jim Gordon	88
Lauren Buzzeo	88
Michael Schachner	87
Susan Kostrzewa	87
Carrie Dykes	87
Fiona Adams	86
Alexander Peartree	86
Grand Total	**88**



ANALYZING WINE REVIEWS

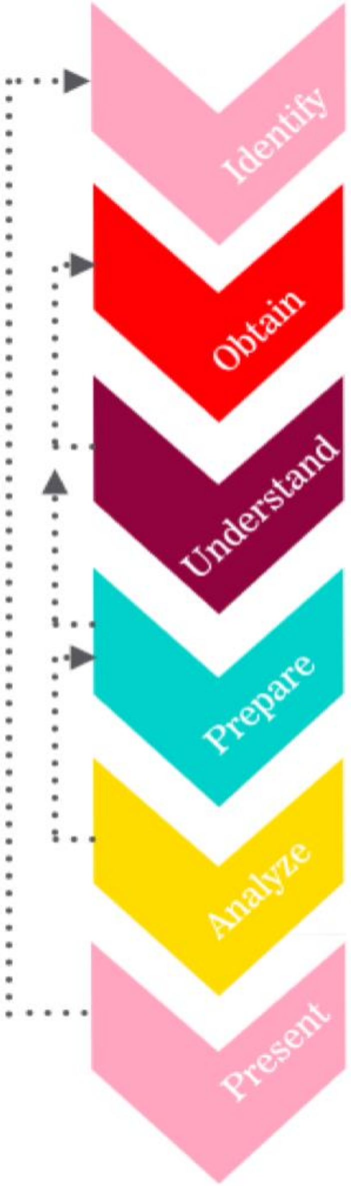
- → Insight
- → Follow-up questions



DATA ANALYSIS WORKFLOW

➔**Present the Results**

- ◆You need to determine the best way to share your results with others.
- ◆How you share your results will depend largely on your audience.
  - ●How data savvy are they?
  - ●How much time do they have to understand the data?
  - ●What context do they need to review your insights gleaned from the data?



IDENTIFY THE PROBLEM

OBTAIN THE DATA

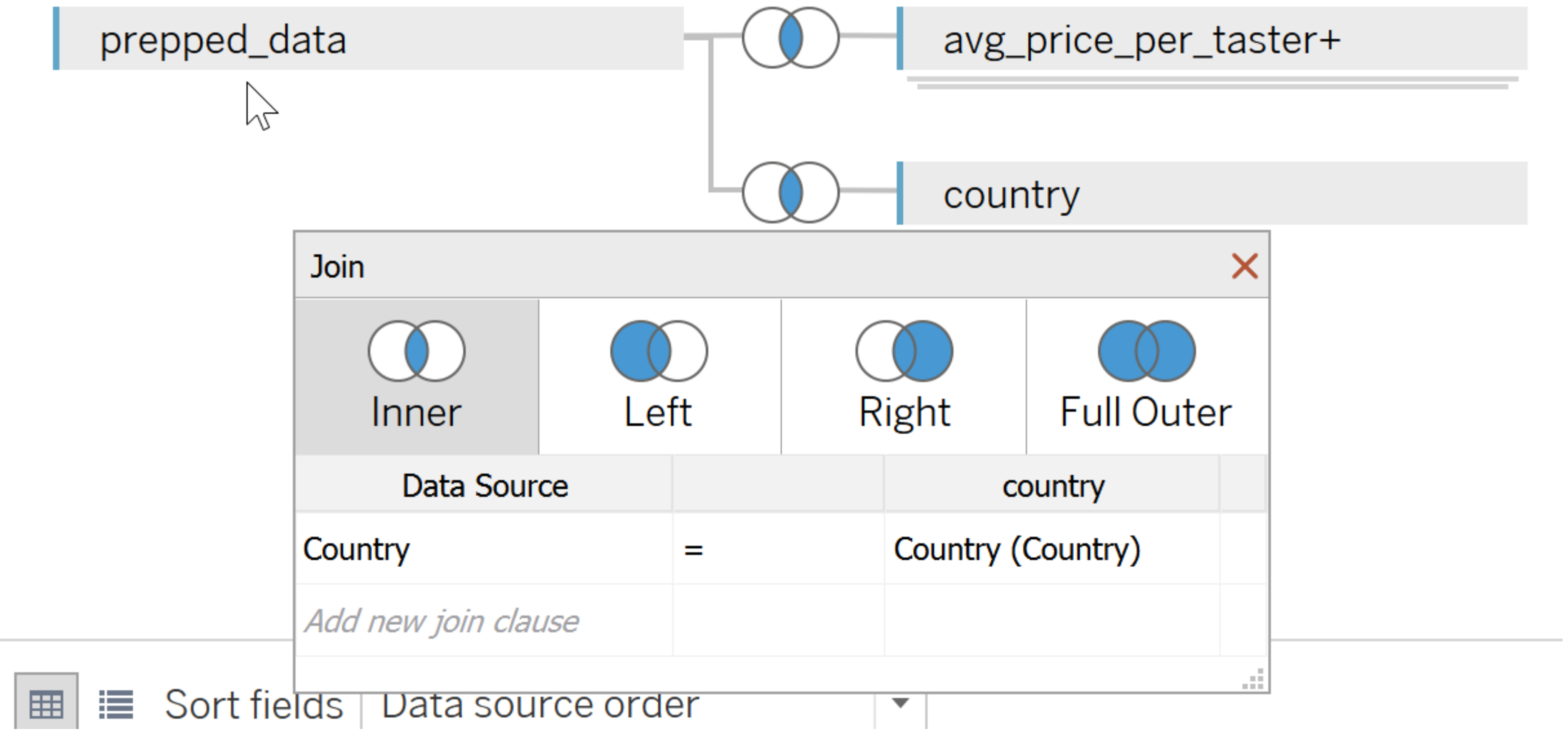
UNDERSTAND THE DATA

PREPARE THE DATA

ANALYZE THE DATA

PRESENT THE RESULTS

DATA ANALYSIS WORKFLOW



Data Source		country
Country	=	Country (Country)
Add new join clause		

- **Watch video:**
- https://www.youtube.com/watch?v=jEgVto5QME8
- https://www.youtube.com/watch?v=mnWm-oIZAoo

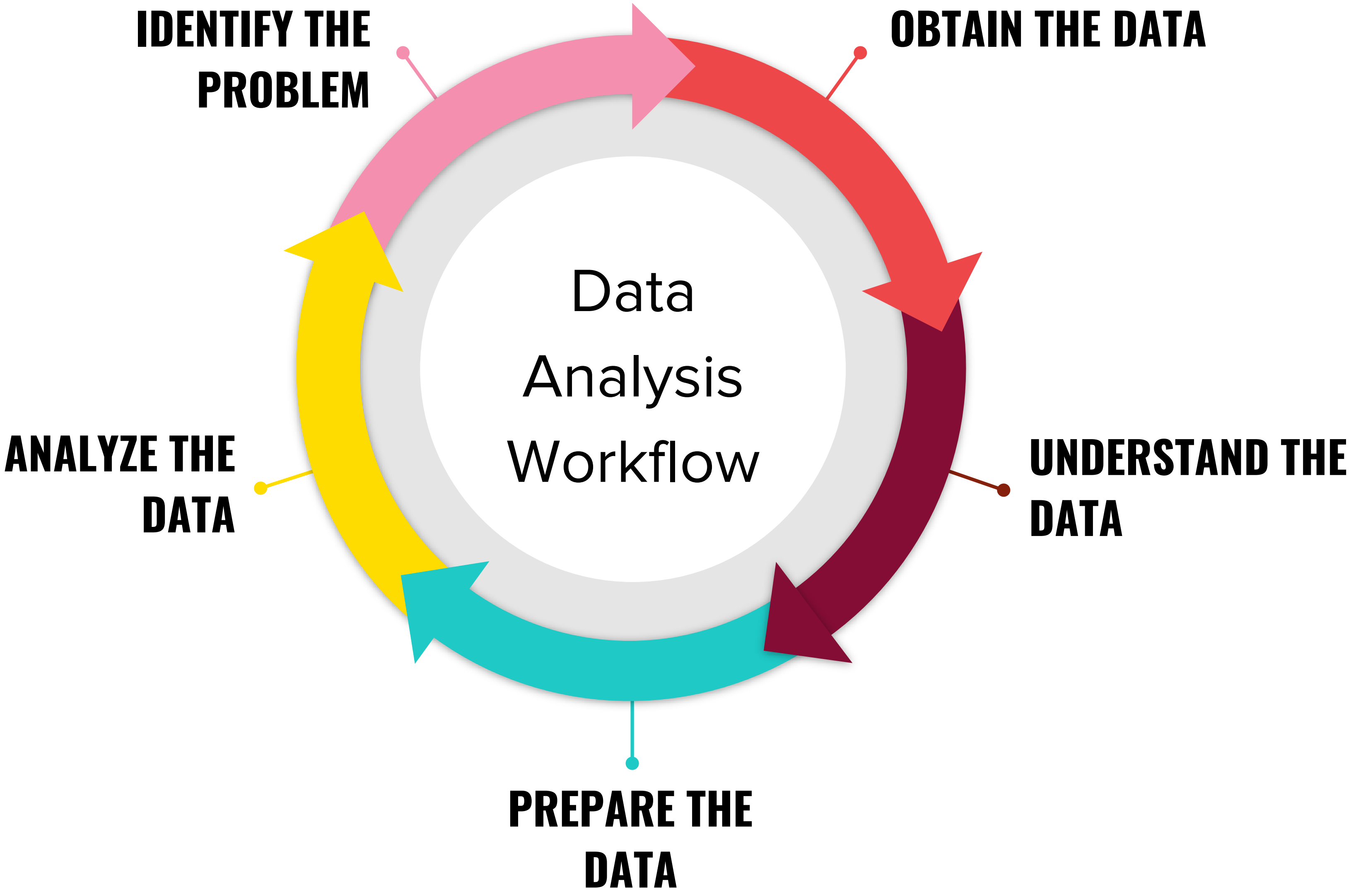
THE DATA ANALYSIS WORKFLOW IS AN ITERATIVE PROCESS

- ➔One of the most important things you'll notice is that the **workflow is not strictly linear**. Even though you begin at the top and end at the bottom, you will revisit various steps along the way as needed.

- ➔Instead of thinking of this as step-by-step instructions for doing data analysis, you should think about these as stages and as the guiding principles for analyzing data *(your analysis design)*.

- ➔Experience in navigating these competing factors is what separates a good data analyst from a great one.

THE DATA ANALYSIS WORKFLOW IS AN ITERATIVE PROCESS



Data Analysis Workflow

IDENTIFY THE PROBLEM

OBTAIN THE DATA

UNDERSTAND THE DATA

PREPARE THE DATA

ANALYZE THE DATA

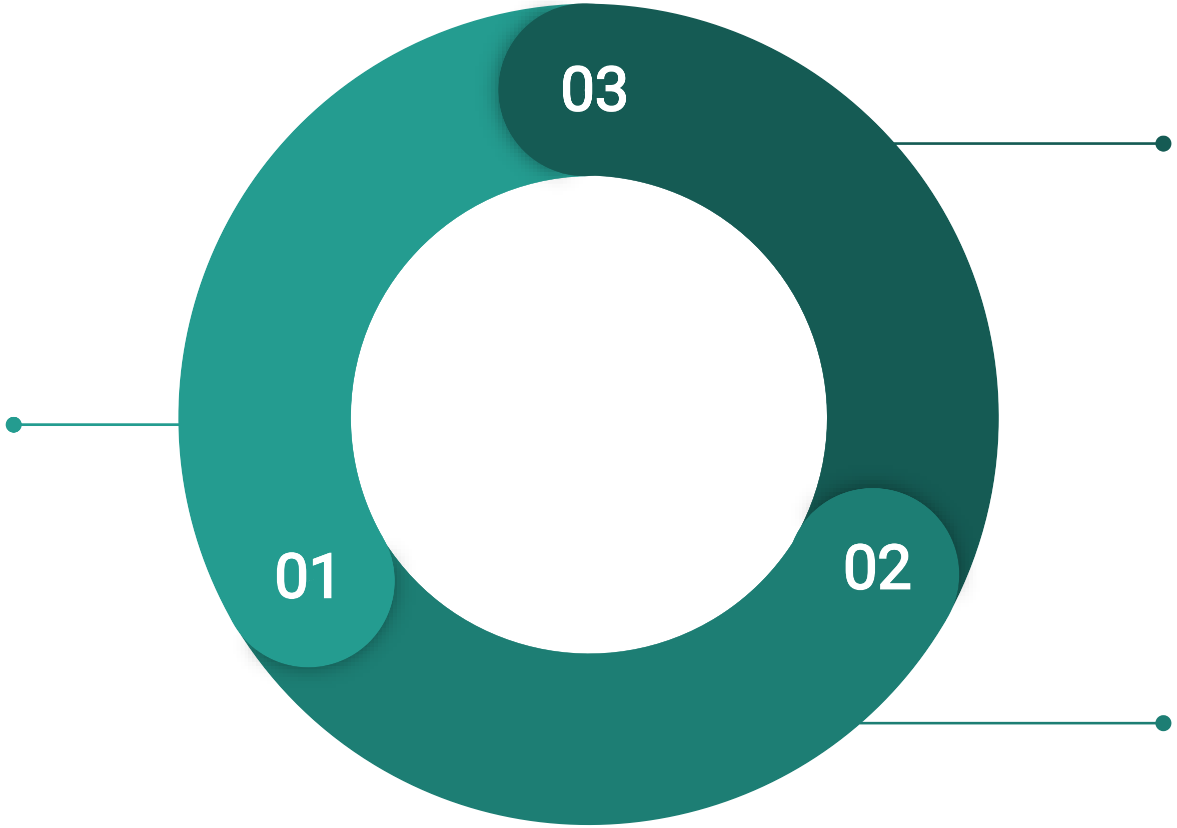
CONCLUSION

FULL CIRCLE

RECAP: LEARNING OBJECTIVES

- Explore common tools and workflows that support data analysis.
- Evaluate the quality and structure of a dataset.
- Use Google Sheets to perform descriptive and exploratory analysis on datasets.
- Use data analytics to inform business recommendations.

3 Types of Data Scientists



Data Analyst

Descriptive Stats and EDA
Business Intelligence

Data Scientist

Statistician and
ML Model Developer

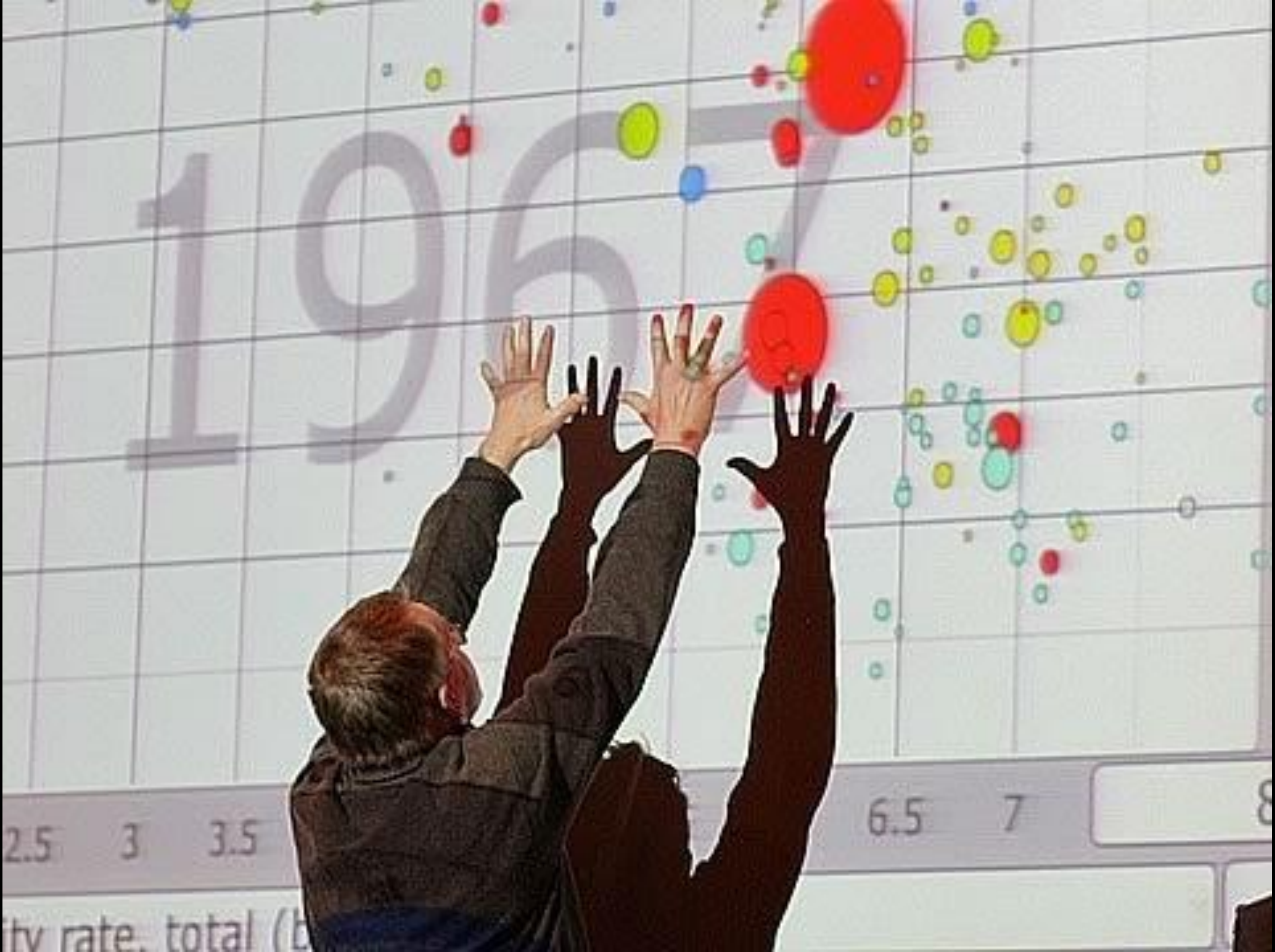
Data Engineer

Data Pipelines and
Infrastructure

CONCLUSION

- Want more help?
  - Google Sheets Help Documentation by Google
  - Google Sheets Functions List
  - Tableau: https://www.youtube.com/watch?v=jEgVto5QME8

- Want to continue your data journey at GA?
  - Google Analytics Bootcamp
  - SQL Bootcamp
  - Data Analytics
  - Python for Data Science


1967
2.5
3
3.5
6.5
7
8
ity rate, total (b

THANK YOU FOR COMING!

Contact Information:

Email: marketing@sigmaridge.com

Twitter: @sigmaridge