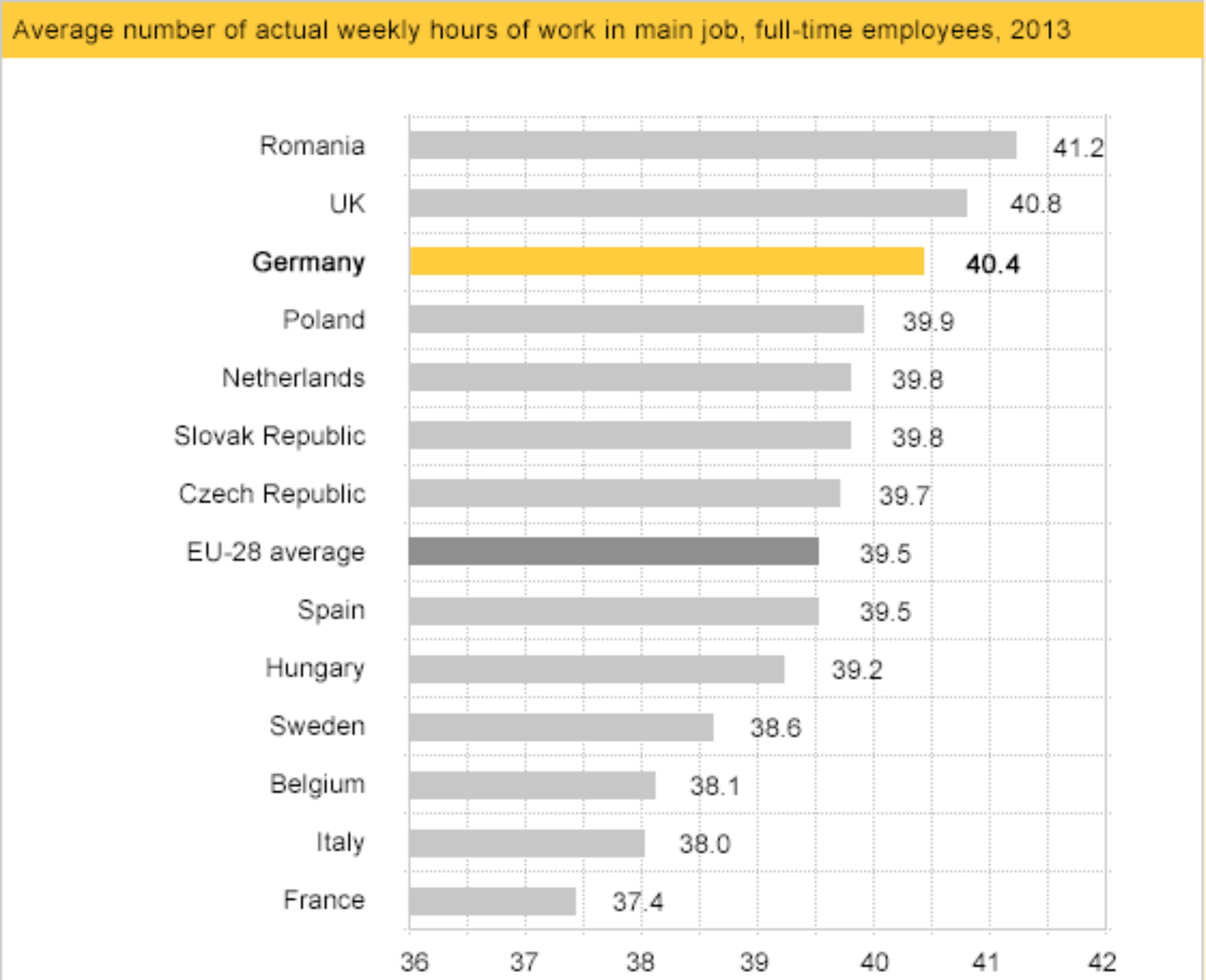
Introduction to Data Visualization

July 27, 2021

DATA 301

Part 1: Importance of data visualization

Credit to: callingbull.org



Weekly hours for full-time employees

Claim: "German workers are more motivated and work more hours than workers in other EU nations."

Q1: How confident are you in the claim based on the visualization?

A. Very confident

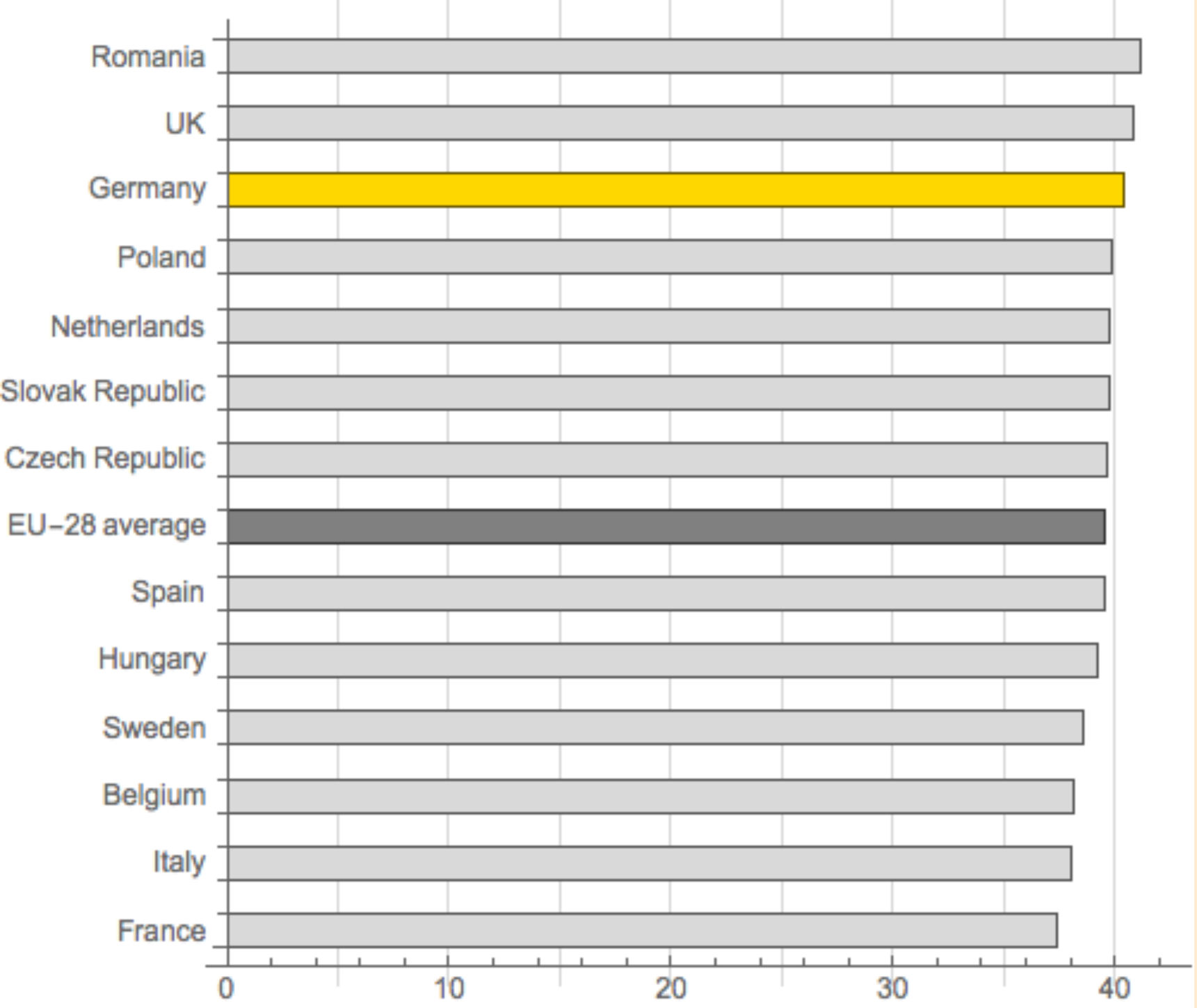
B. Confident

C. Not Confident

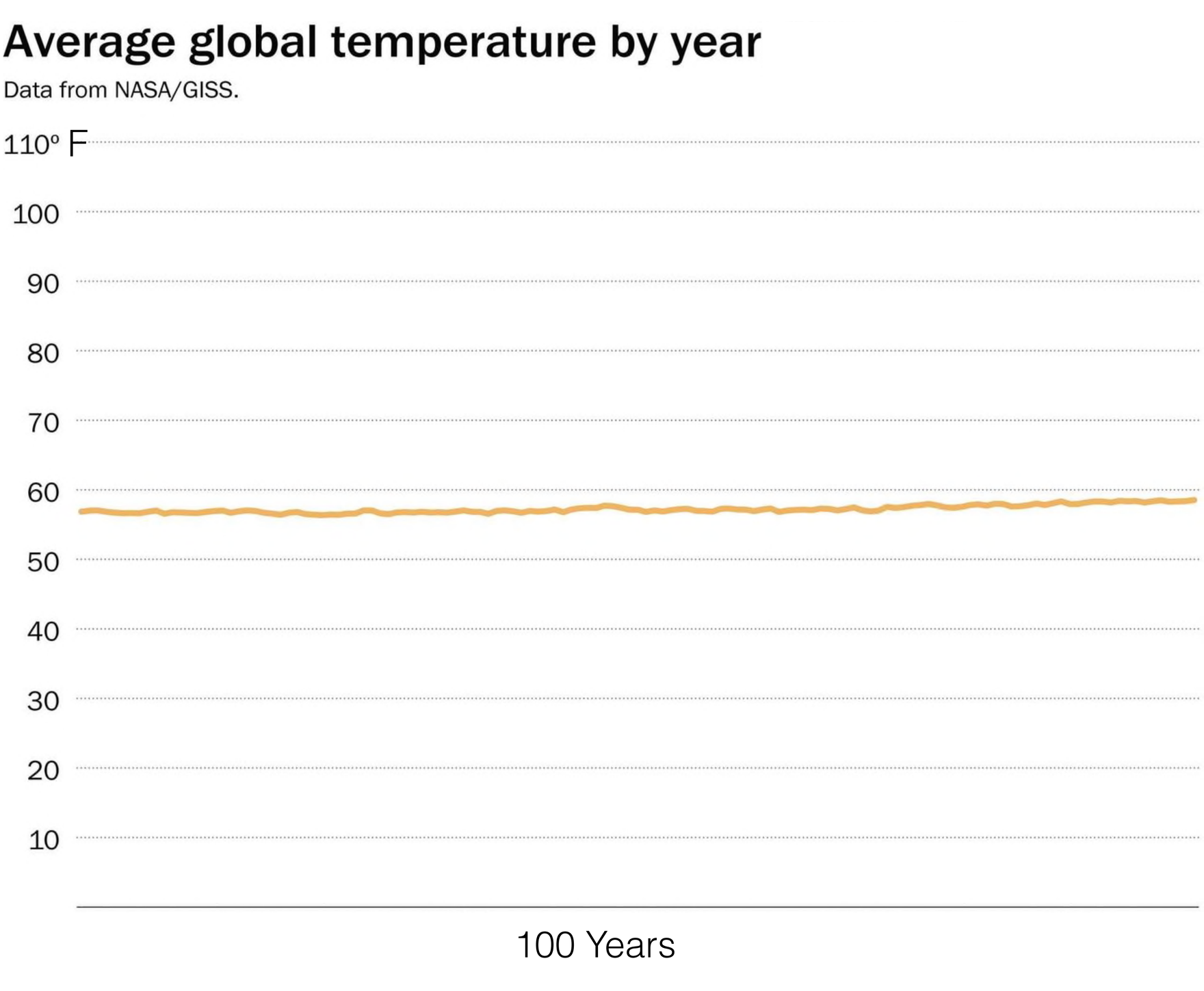
D. Claim is wrong

Data Source: Eurofound 2014
Example Source: Callingbull.org

Weekly hours for full-time employees



Example Source: Callingbull.org



Average Global Temperatures by Year

Claim: “Over **these 100 years**, there is a negligible change in global temperature.”

Q2: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

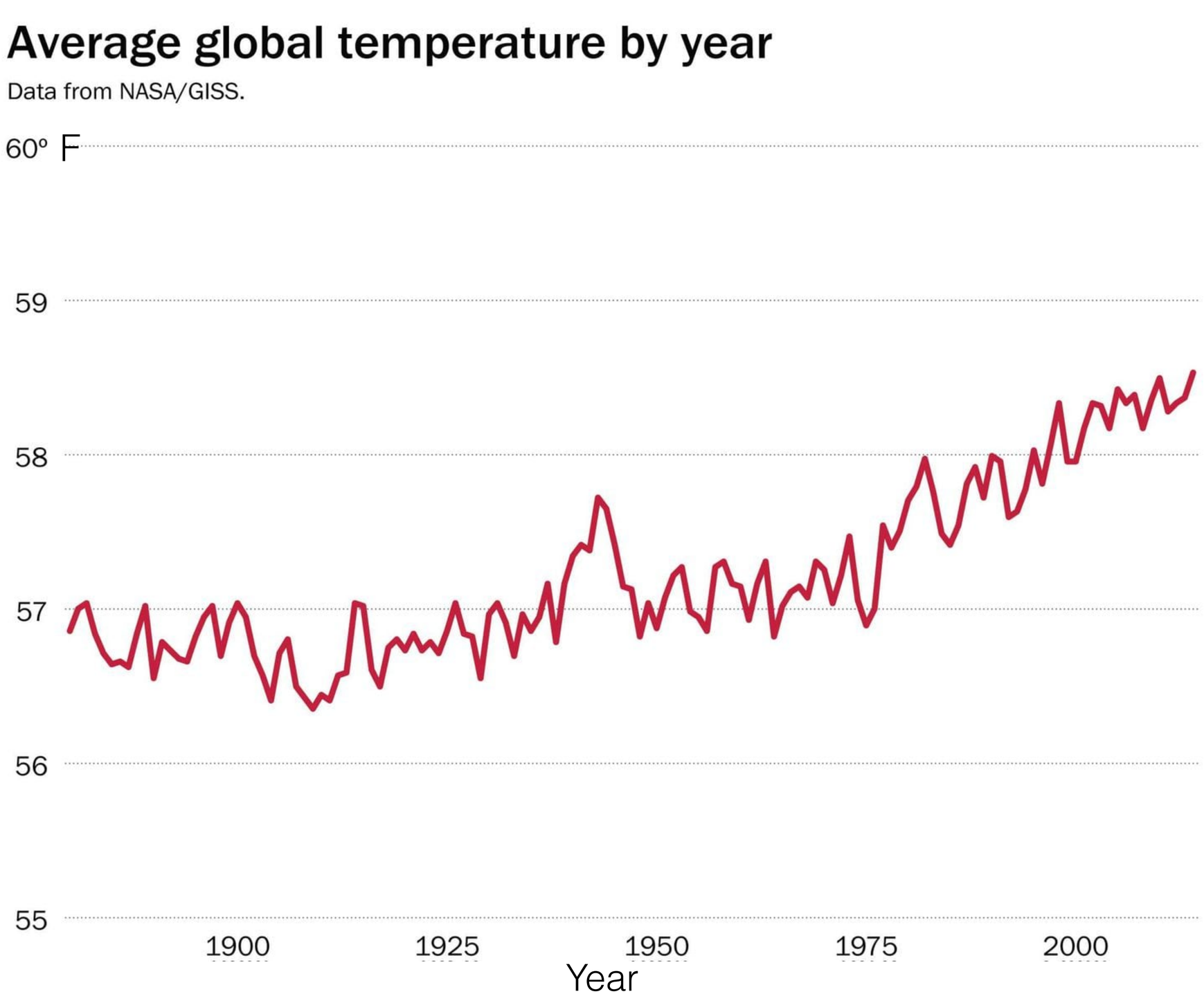
C. Not Confident

D. Claim is wrong

Data Source

Example Source

Average Global Temperatures by Year



Data Source

Example Source

nature
International journal of science

Brief Communication | Published: 29 September 2004

Athletics

Momentous sprint at the 2156 Olympics?

Andrew J. Tatem ✉, Carlos A. Guerra, Peter M. Atkinson & Simon I. Hay

Nature **431**, 525 (2004) | Download Citation ↓

1743 Accesses | **46** Citations | **78** Altmetric | Metrics »

Women sprinters are closing the gap on men and may one day overtake them.

Abstract

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

Gender parity in the Olympics (100m race)

Claim: "Women sprinters are closing the gap on men and may one day overtake them."

Q3: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

C. Not Confident

D. Claim is wrong

Gender parity in the Olympics (100m race)

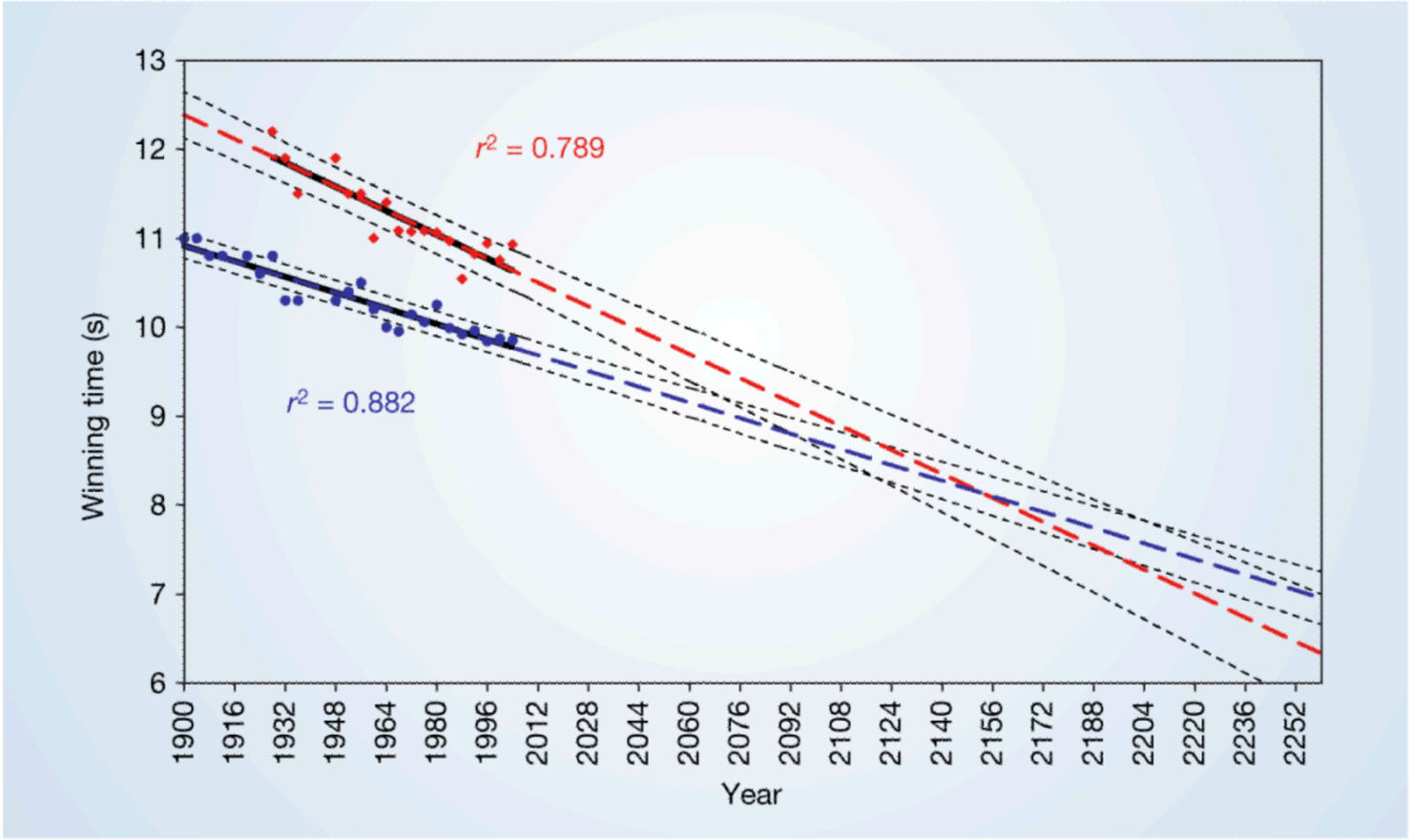
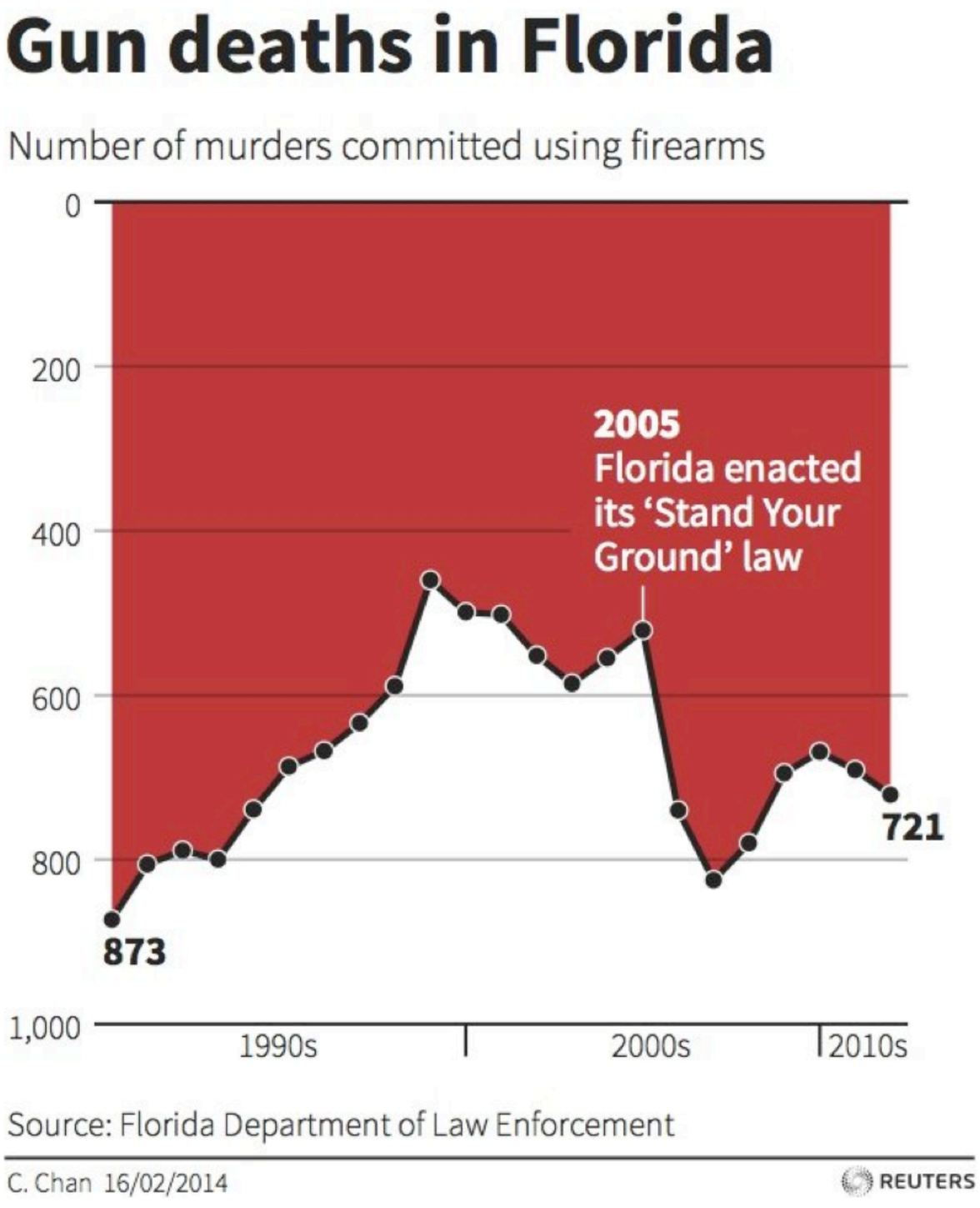


Figure 1.
The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.



Gun deaths in Florida after legislation

Claim: "After enacting new gun legislation in Florida, gun deaths sharply declined."

Q4: How confident are you in the claim based on the visualization?

A. Very confident

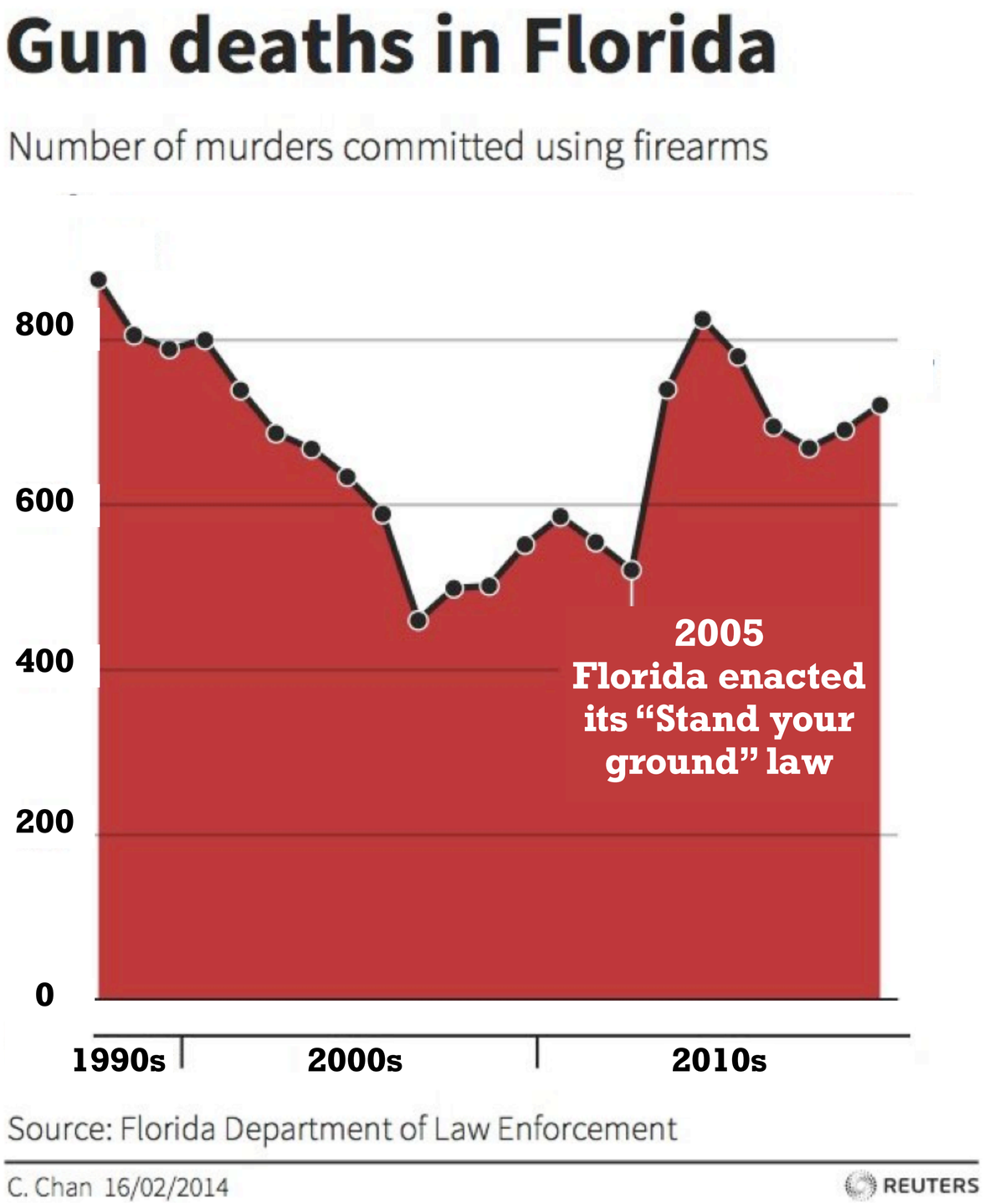
B. Confident

C. Not Confident

D. Claim is wrong

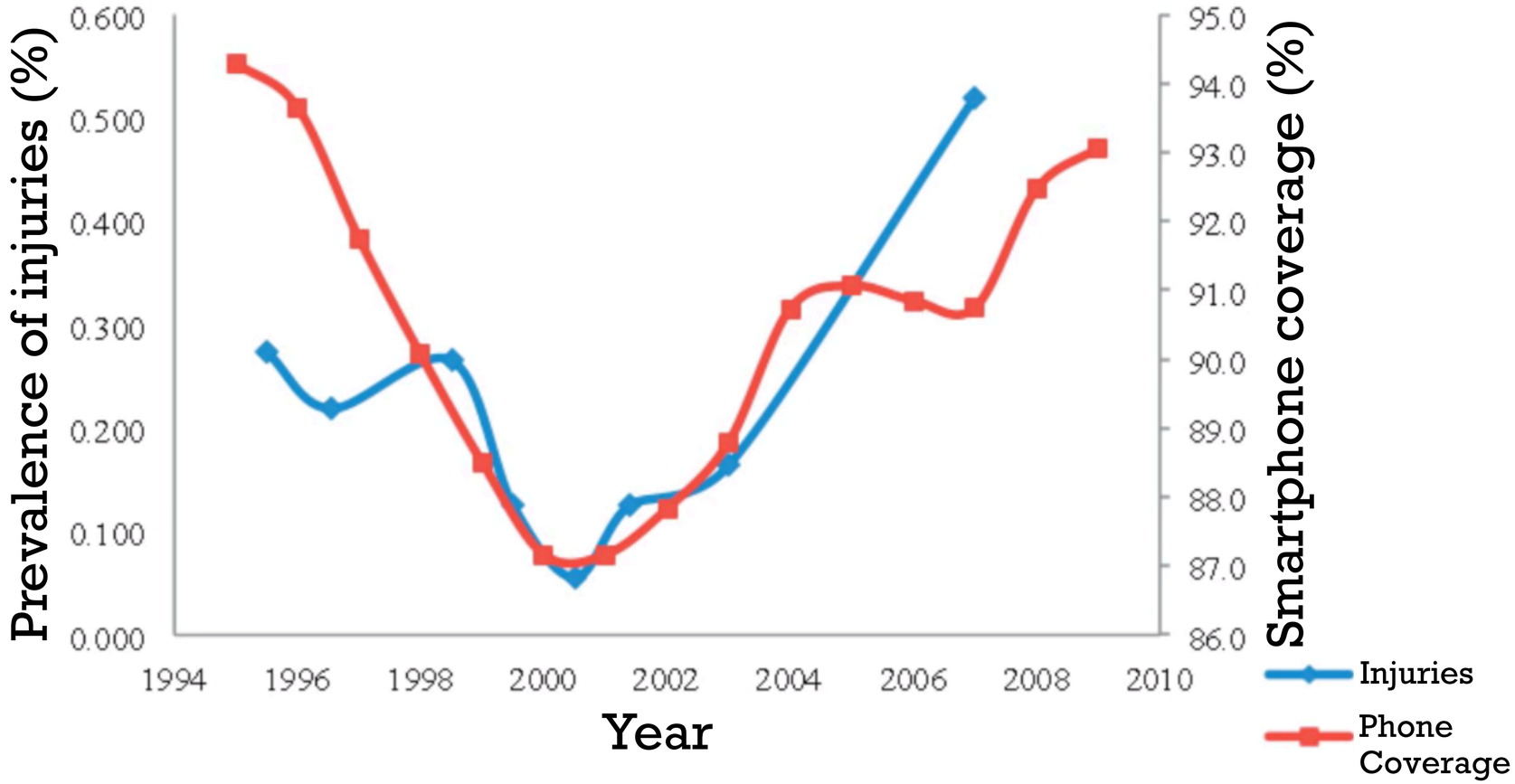
Data Source: Florida Department of Law Enforcement
Example Source: Callingbull.org & Reuters

Gun deaths in Florida after legislation



Data Source: Florida Department of Law Enforcement
Example Source: Callingbull.org & Reuters

Prevalence of wrist/thumb injuries in population and smartphone coverage in Bristol, UK



Injuries after smartphone coverage

Claim: "The rise of smartphones in the population have dramatically increased prevalence of wrist/thumb injuries."

Q5: How confident are you in the claim based on the visualization?

A. Very confident

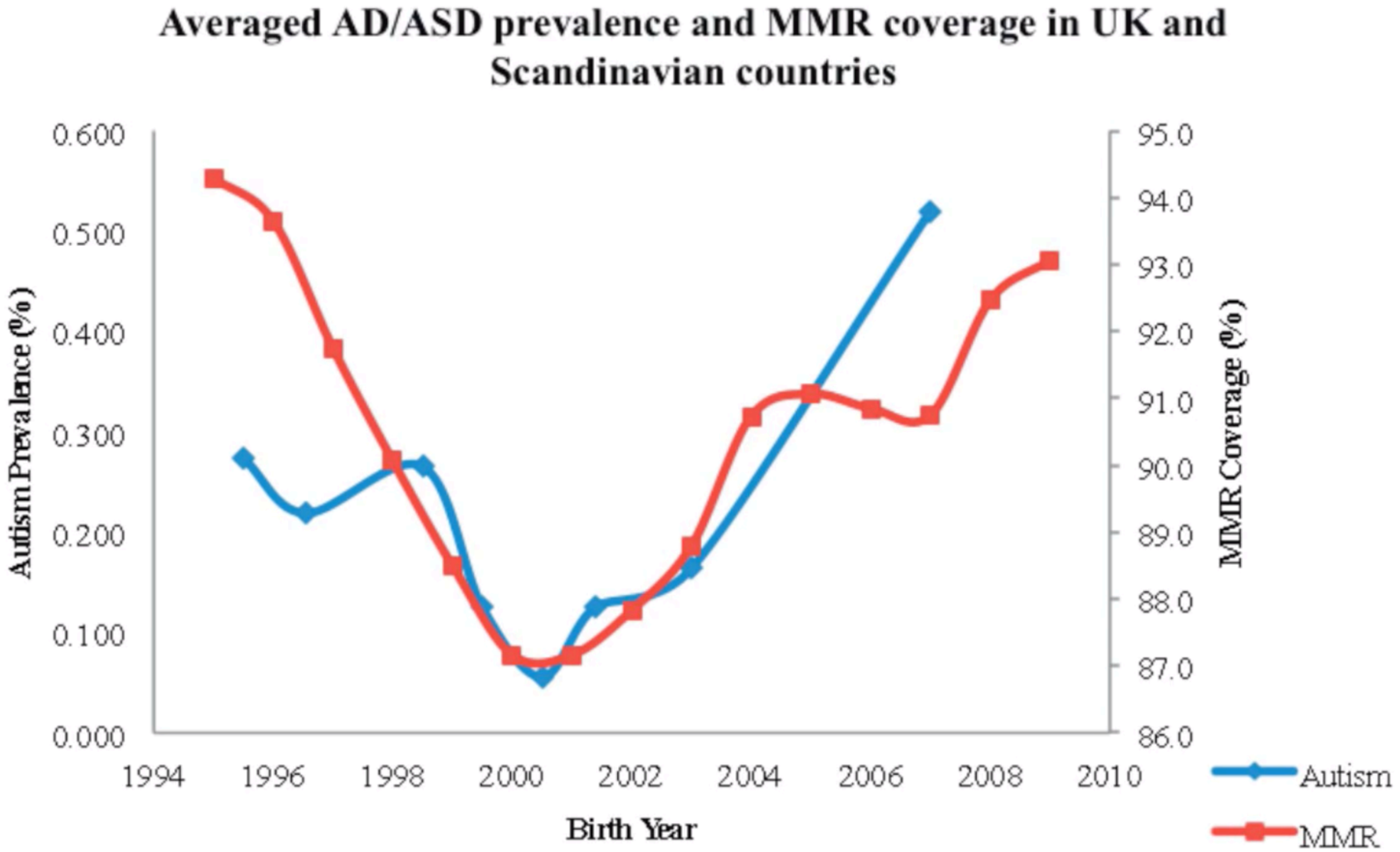
B. Confident

C. Not Confident

D. Claim is wrong

Adapted from: callingbull.org

Autism and MMR coverage



Claim: "The rise of smartphones in the population have dramatically increased prevalence of wrist/thumb injuries."

Q5: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

C. Not Confident

D. Claim is wrong

Adapted from: callingbull.org

Autism and MMR coverage

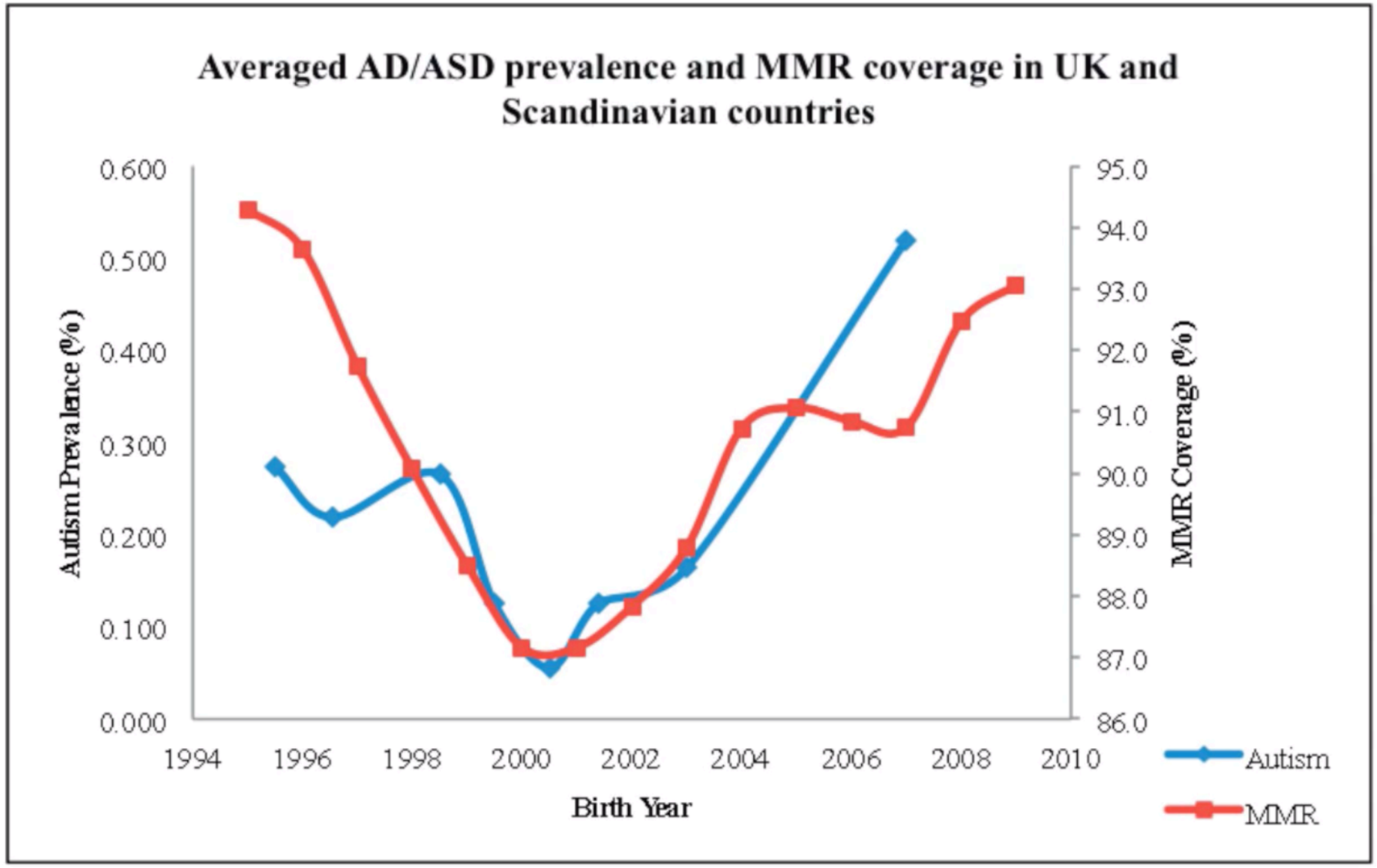


Figure 1-Averaged AD/ASD prevalence and MMR coverage in UK, Norway and Sweden. Both MMR and AD/ASD data are normalized to the maximum coverage/prevalence during the time period of this analysis.

Claim: "The rise of MMR coverage in the population have dramatically increased prevalence of Autism Spectrum Disorder."

Q5: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

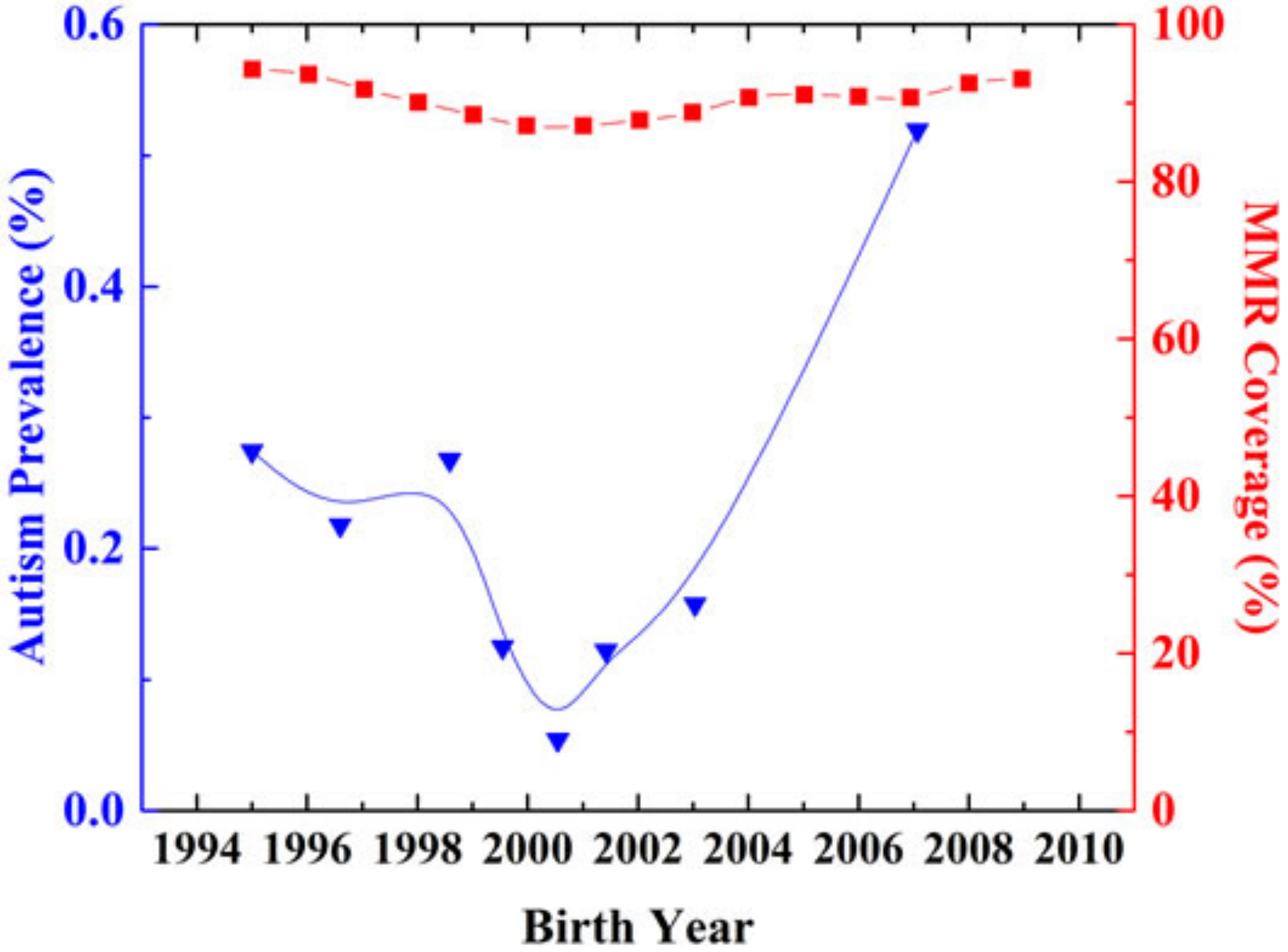
C. Not Confident

D. Claim is wrong

Diesher *et al.* 2015 *Issues in Law and Medicine*

Adapted from: callingbull.org

Autism and MMR coverage



Revised graph courtesy of Matt Carey

Full analysis/critique of paper

Adapted from: callingbull.org

20 minutes

Part 2:
Exploratory Data Analysis

7 Exploratory Data Analysis

7.1 Introduction

This chapter will show you how to use visualisation and transformation to explore your data in a systematic way, a task that statisticians call exploratory data analysis, or EDA for short. EDA is an iterative cycle. You:

1. Generate questions about your data.

2. Search for answers by visualising, transforming, and modelling your data.

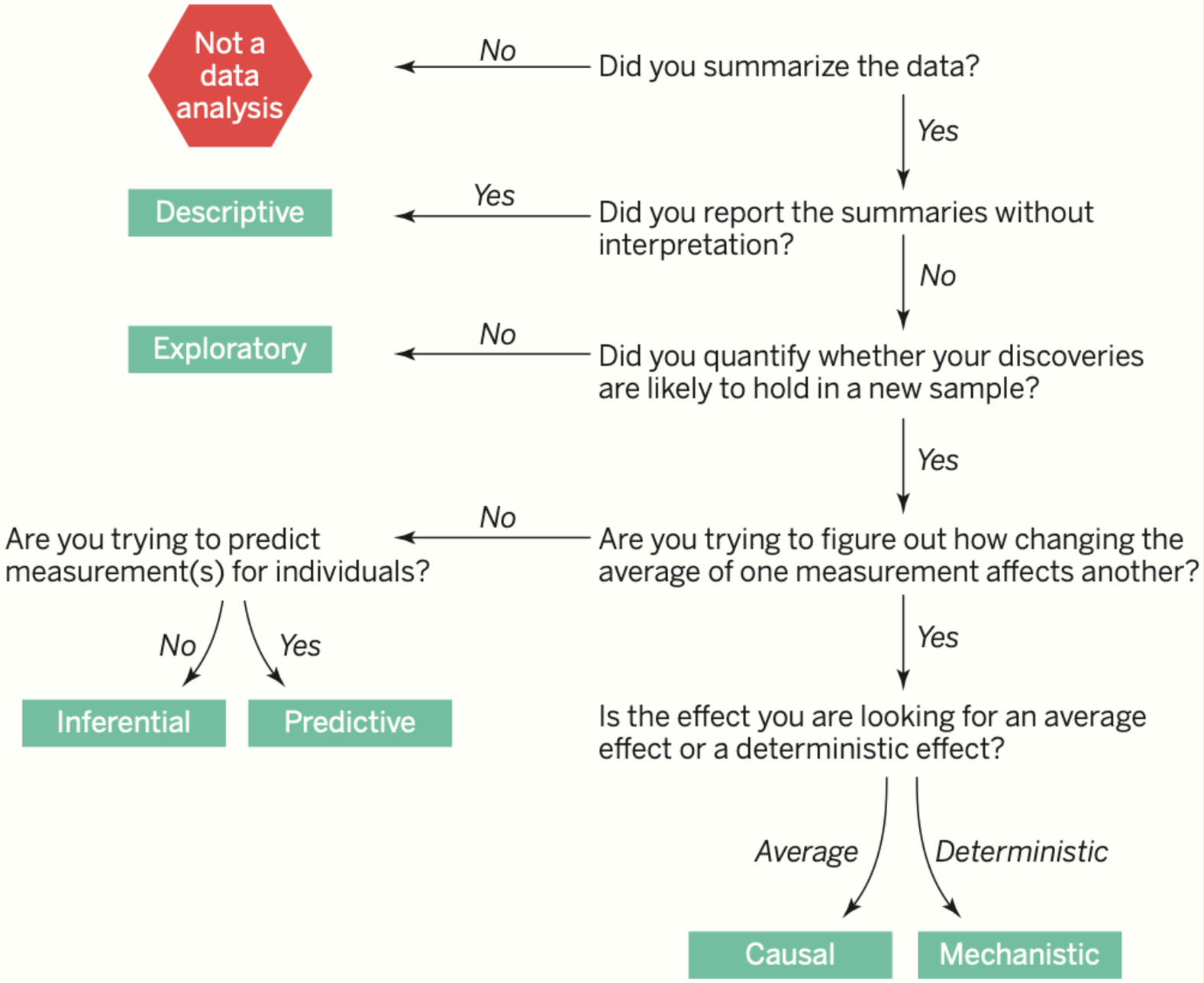
3. Use what you learn to refine your questions and/or generate new questions.

EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive areas that you'll eventually write up and communicate to others.

Types of Research Questions

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Data analysis flowchart



Source: What is the Question?
By Jeffery Leek and Roger Peng

Research Questions

1. Descriptive

- one that seeks to summarize a characteristic of a set of data
- no interpretation of the result itself as the result is a fact, an attribute of the data set you are working with

- e.g., What is the frequency of viral illnesses in a set of data collected from a group of individuals?
- e.g., How many people live in each US state?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

2. Exploratory

- one in which you analyze the data to see if there are patterns, trends, or relationships between variables
- looking for patterns that would support proposing a hypothesis to test in a future study

- e.g., Do diets rich in certain foods have differing frequencies of viral illnesses in a set of data collected from a group of individuals?
- e.g., Does air pollution correlate with life expectancy in a set of data collected from groups of individuals from several regions in the United States?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

3. Inferential

- **one in which you analyze the data to see if there are patterns, trends, or relationships between variables in a representative sample**
- **want to quantify how much the patterns, trends, or relationships between variables is applicable to all individuals units in the population**

- **e.g., Is eating at least 5 servings a day of fresh fruit and vegetables is associated with fewer viral illnesses per year?**
- **e.g., Does air pollution correlate with life expectancy in the United States?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

4. Predictive

- **one where you are trying to predict measurements or labels for individuals (people or things)**
- **less interested in what causes the predicted outcome, just what predicts it**

- **e.g., How many viral illnesses will someone have next year?**
- **e.g., What political party will someone vote for in the next US election?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

5. Causal

- asks about whether changing one factor will change another factor, on average, in a population.
- Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal (e.g., randomized experiment or trial)

- e.g., Does eating at least 5 servings a day of fresh fruit and vegetables cause fewer viral illnesses per year?
- e.g., Does smoking cause cancer?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

6. Mechanistic

- **one that tries to explain the underlying mechanism of the observed patterns, trends, or relationship (how does it happen?)**

- **e.g., How do changes in diet lead to a reduction in the number of viral illnesses?**
- **e.g., How does airplane wing design changes air flow over a wing, leading to decreased drag?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Types of Research Questions

Exploratory Data Analysis

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Steps of EDA

1. Describe your dataset

Task: Describe your dataset. Consider the following questions to guide you in your exploration

- Who: Which company/agency/organization provided this data?
- What: What is in your data?
- When: When was your data collected (for example, for which years)?
- Why: What is the purpose of your dataset? Is it for transparency/accountability, public interest, fun, learning, etc...
- How: How was your data collected? Was it a human collecting the data? Historical records digitized? Server logs?

Steps of EDA

2. Load the dataset

Task: Load your dataset from a file, or URL. This needs to be a pandas dataframe so you can use it with Seaborn. Remember that others may be running your jupyter notebook so it's important that the data is accessible to them. If your dataset isn't accessible as a URL, make sure to commit it into your repo.

Steps of EDA

3. Explore your dataset

Task: Explore the columns in your dataset. Which ones are interesting/relevant? You can use df.profile_report()

To install pandas-profiling:

conda install -c conda-forge pandas-profiling

Source: EDA with Pandas Profiling

Steps of EDA

4. Initial thoughts

Task: Use this a place to record any observations you come up with, anything jump out at you as surprising or particularly interesting? Where do you think you'll go with exploring this dataset? Feel free to take notes in this section and use it as a scratch pad. Any content in this area will not be marked.

Steps of EDA

5. Wrangling

Task: You can do any wrangling you need to do here. Describe what you're doing (or did) using comments within your code.

Steps of EDA

6. Research questions

Task: come up with at least two research questions about your dataset that will require data visualizations to help answer. Recall that for this purpose, you should only aim for "Descriptive" or "Exploratory" research questions.

Steps of EDA

7. Data Analysis & Visualizations

Task: Create data visualizations (and justify your choices) using Seaborn that will help you answer your research questions.

Steps of EDA

8. Summary and conclusions

Task: Summarize your findings and describe any conclusions and insight you were able to draw from your visualizations.

20 minutes

Part 3:
Motivating the need for EDA

Recap: Exploratory Data Analysis (EDA)

B1. Describe your dataset (2 marks)

Consider the following questions to guide you in your exploration:

- Who: Which company/agency/organization provided this data?
- What: What is in your data?
- When: When was your data collected (for example, for which years)?
- Why: What is the purpose of your dataset? Is it for transparency/accountability, public interest, fun, learning, etc...
- How: How was your data collected? Was it a human collecting the data? Historical records digitized? Server logs?

B2. Load the dataset from a file, or URL (1 mark)

This needs to be a pandas dataframe. Remember that others may be running your jupyter notebook so it's important that the data is accessible to them. If your dataset isn't accessible as a URL, make sure to commit it into your repo. If your dataset is too large to commit (>100 MB), and it's not possible to get a URL to it, you should contact your instructor for advice.

B3. Explore your dataset (3 marks)

Which of your columns are interesting/relevant? Remember to take some notes on your observations, you'll need them for the next EDA step (initial thoughts).

B4. Initial Thoughts (2 marks)

Does anything jump out at you as surprising or particularly interesting?

Where do you think you'll go with exploring this dataset? Feel free to take notes in this section and use it as a scratch pad.

B5. Wrangling (5 marks)

The next step is to wrangle your data based on your initial explorations. Normally, by this point, you have some idea of what your research question will be, and that will help you narrow and focus your dataset.

B6. Research questions (2 marks)

B7. Data Analysis and Visualizations

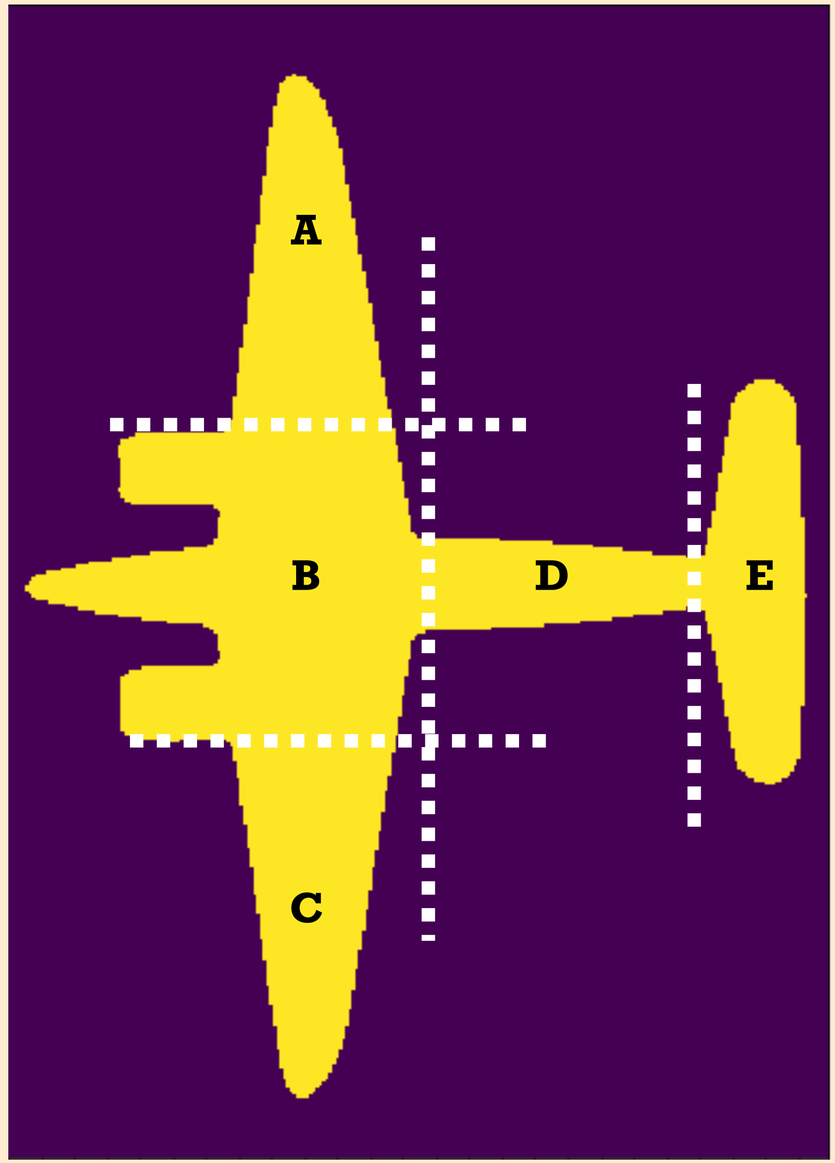
Case Study: Planes in WW2

You have been given a dataset and tasked with trying to solve a problem. In WW2, **expensive fighter planes were going down quite frequently due to bullet fire.** The military decided to conduct an analysis and surveyed all the surviving planes in an effort to catalogue which regions of the plane should be reinforced.

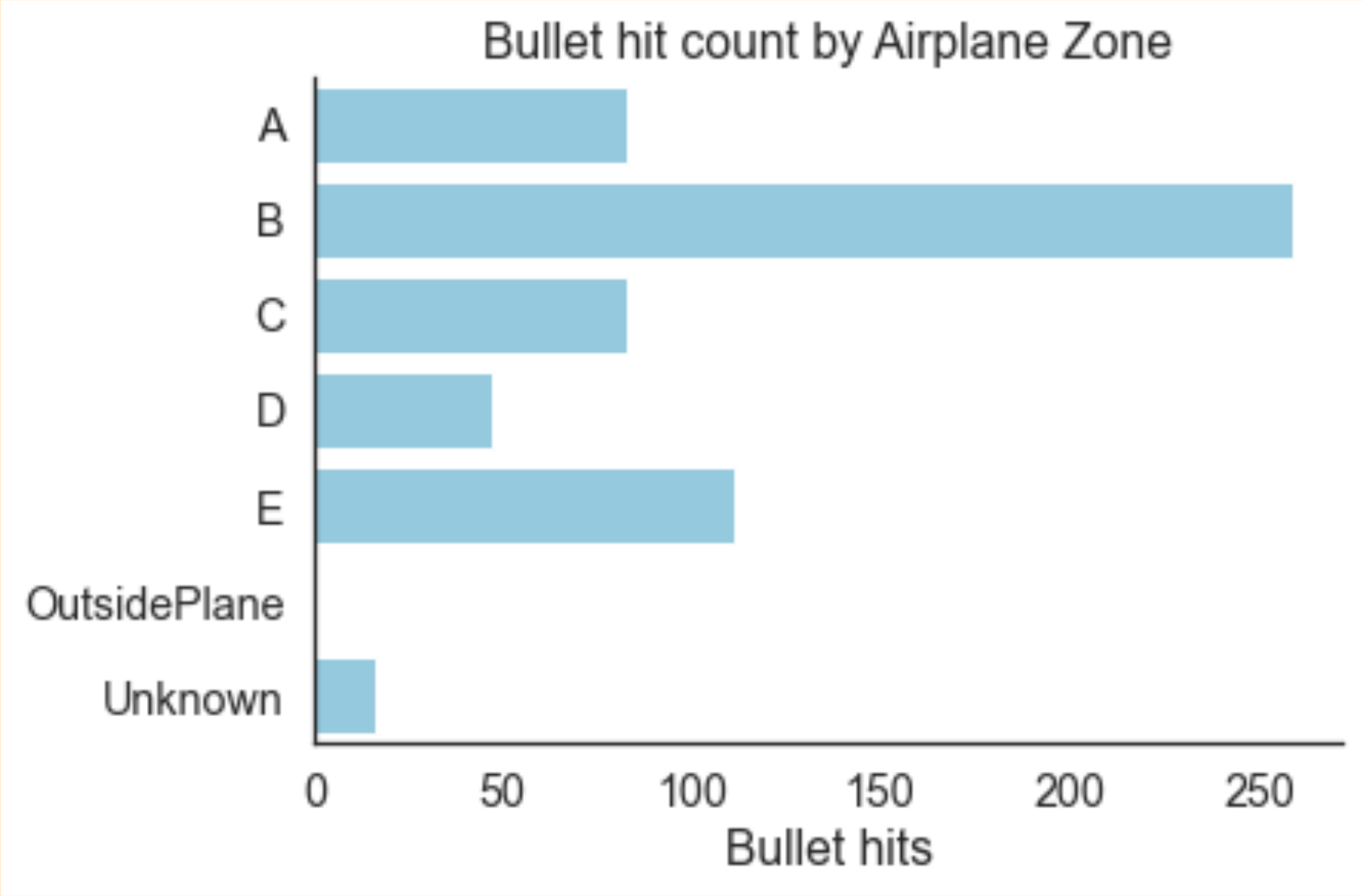
With limited resources, the military could only reinforce a maximum of two zones. Your task is to look at the bullet data for the planes and help determine which areas of the plane should be reinforced.

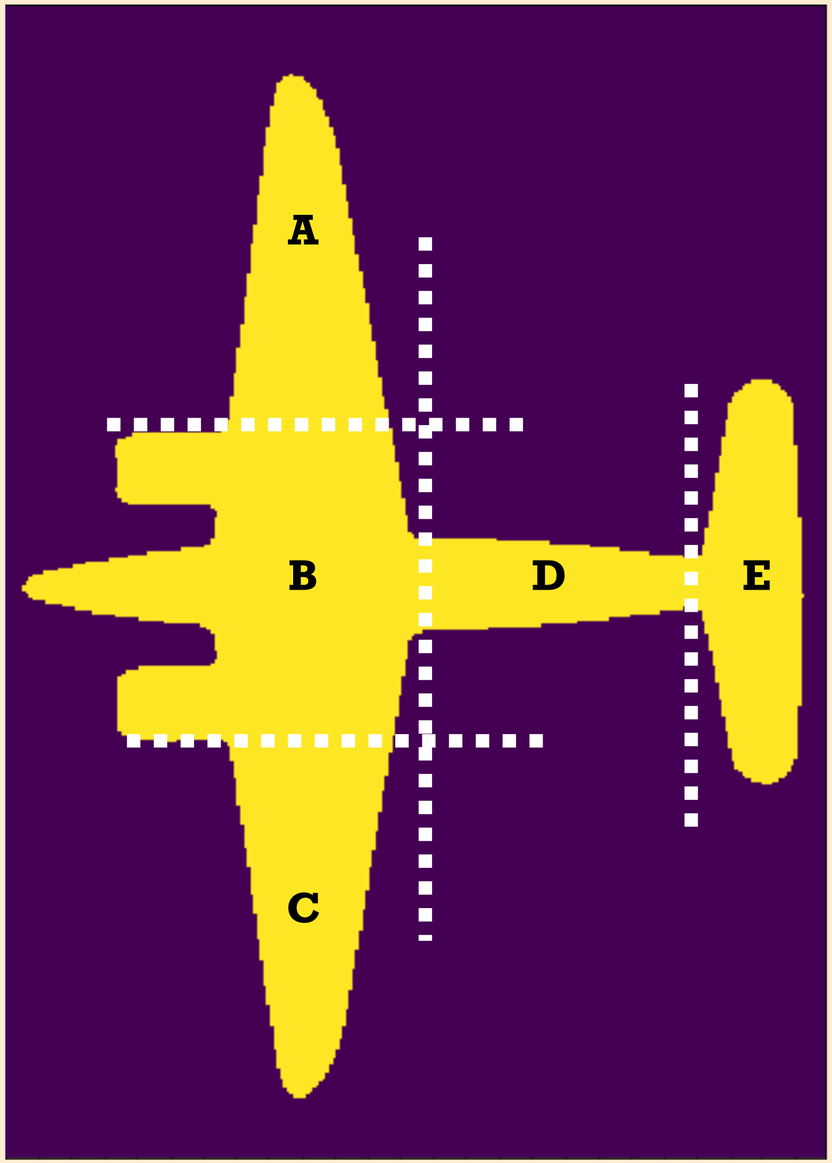
You're given a schematic of the plane, and told that the workers added a grid to the schematic, divided it up into regions A,B,C,D,E and recorded a value of 1 wherever there was a bullet hole across all the planes that returned. Areas without bullet holes are marked as 0.

They gave you a csv file with this information called `bullet_data.csv`. Yes, these WW2 workers are very sophisticated and had access to a computer :-).



Case Study: Planes in WW2





Debrief - EDA is important!

- Look at your data.
- Talk to someone about your data.
- Look at your data another way.
- Think about your data and what it means!