Communicating & Data Visualizations

Nov 28, 2024

CPSC 330
(DSC 320!)

Motivation

Why should we care about effective communication?

- Most ML practitioners work in an organization with >1 people.
- There will very likely be stakeholders other than yourself.
- Some of them might not have any background in ML or computer science.
- If your ML model is going to automate some important decisions in the organization you need to be able to explain
    - What does a particular result mean?
    - When does the model work?
    - What are the risks? When does it fail?
    - Why the model made a certain prediction for a particular example?
    - What are the consequences of using your model?
- If you are able to convince your manager that using is model is beneficial, then only there are chances of your work going in production.
- That said, you want to be honest when discussing the aspects above. If you mis-communicate the performance of your model, people will find out when the deployed model does not quite give similar performance.

Main issues in ML-related communication

- Overstating one's results / unable to articulate the limitations
- Unable to explain the predictions
- Can we trust test error?
- Why did a particular model (e.g., CatBoost) make that prediction?
- What does it mean if `predict_proba` outputs 0.9?

These issues are there because these things are actually very hard to explain!

Is this misleading?

Home › News › ‘Machine learning could predict heart attack with over 90% accuracy’

News

‘Machine learning could predict heart attack with over 90% accuracy’

By **Elets News Network** - May 13, 2019

Like 46

Share

1. What additional information would you need to evaluate the validity of this claim?
2. How would you rephrase this headline to make it more accurate and less misleading?
3. What metrics would you use to evaluate the performance of a model like this?

Part 1: Importance of Communicating

Activity1 : Explaining Grid Search

Article 0

bit.ly/cpsc330_link0

Explanation 1

Machine learning algorithms, like an airplane's cockpit, typically involve a bunch of knobs and switches that need to be set.

For example, check out the documentation of the popular random forest algorithm here. Here's a list of the function arguments, along with their default values (from the documentation):

Article 2

bit.ly/cpsc330_link2

An introduction to Grid Search

Krishni · Follow
Published in DataDrivenInvestor · 2 min read · Jan 5, 2019

Activity1 : Explaining Grid Search

Article 0

bit.ly/cpsc330_link0

Article 2

bit.ly/cpsc330_link2

Discussion questions:

- What do you like about each explanation?
- What do you dislike about each explanation?
- What do you think is the intended audience for each explanation?
- Which explanation do you think is more effective overall for someone on Day 1 of CPSC 330?
- Each explanation has an image. Which one is more effective? What are the pros/cons?
- Each explanation has some sample code. Which one is more effective? What are the pros/cons?

Concepts *then* labels, not the other way around

The first explanation start with an analogy for the concept (and the label is left until the very end):

> Machine learning algorithms, like an airplane's cockpit, typically involve a bunch of knobs and switches that need to be set.

In the second explanation, the first sentence is wasted on anyone who doesn't already know what "hyperparameter tuning" means:

> Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model.

The effectiveness of these different statements depend on your audience.

See this video:

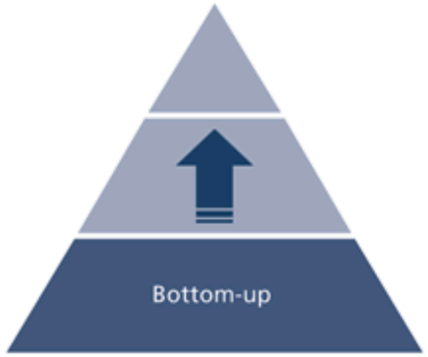
> I learned very early the difference between knowing the name of something and knowing something." - Richard Feynman.

Bottom-up explanations

The Curse of Knowledge leads to *top-down* explanations:

- When you know something well, you think about things in the context of all your knowledge.
- Those lacking the context, or frame of mind, cannot easily understand.

There is another way: *bottom-up* explanations:



When you're brand new to a concept, you benefit from analogies, concrete examples and familiar patterns.

In the previous examples, which one represented a bottom-up explanation and which one a top-down explanation?

New ideas in small chunks

The first explanation has a hidden conceptual skeleton:

1. The concept of setting a bunch of values.
2. Random forest example.
3. The problem / pain point.
4. The solution.
5. How it works - high level.
6. How it works - written example.
7. How it works - code example.
8. The name of what we were discussing all this time.

Reuse your running examples

Effective explanations often use the same example throughout the text and code. This helps readers follow the line of reasoning.

Approach from all angles

When we're trying to draw mental boundaries around a concept, it's helpful to see examples on all sides of those boundaries. If we were writing a longer explanation, it might have been better to show more, e.g.

- Performance with and without hyperparameter tuning.
- Other types of hyperparameter tuning (e.g. `RandomizedSearchCV`).

When experimenting, show the results asap

The first explanation shows the output of the code, whereas the second does not. This is easy to do and makes a big difference.

Interesting to you != useful to the reader (aka it's not about you)

Here is something which was deleted from the explanation:

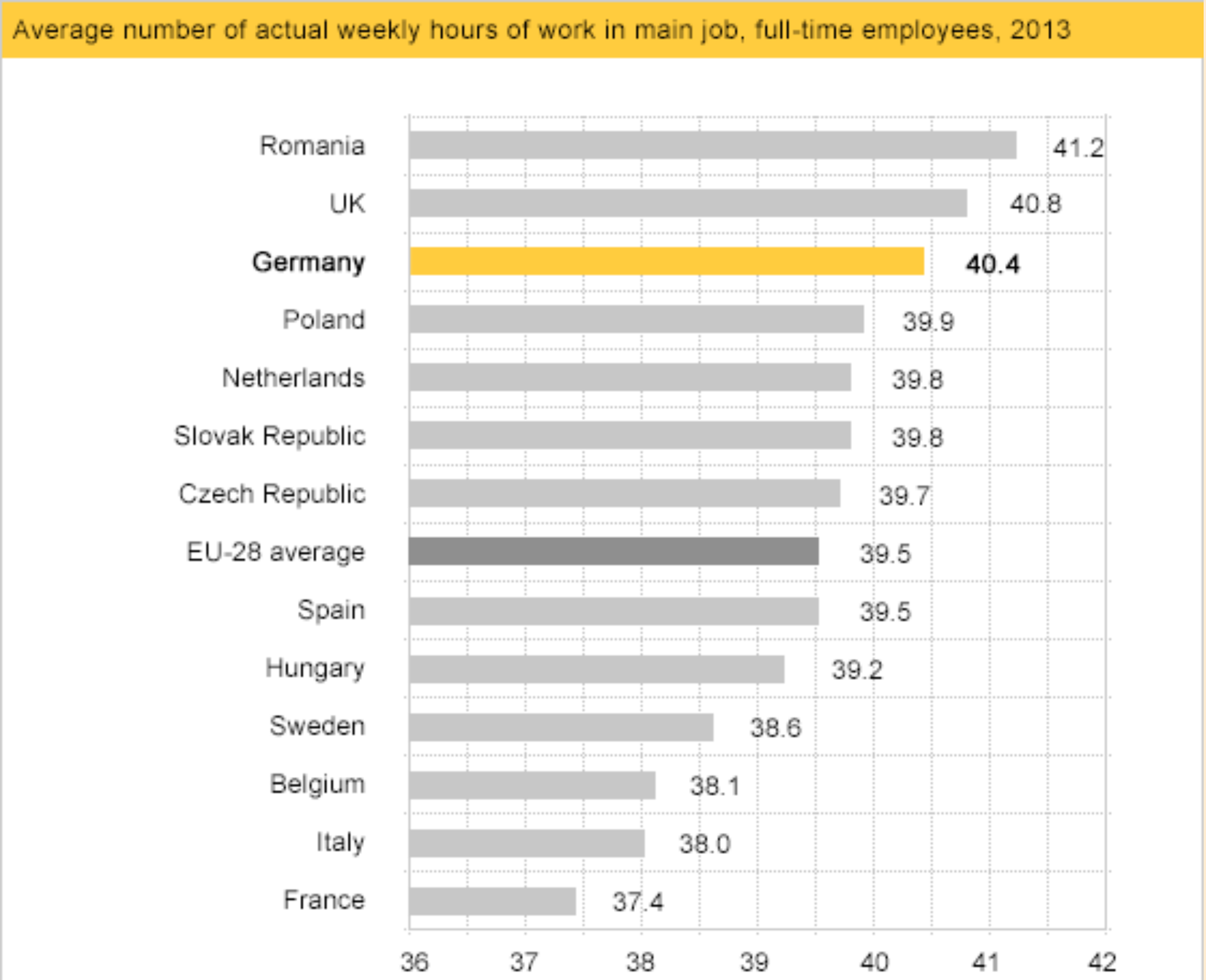
> Some hyperparameters, like `n_estimators` are numeric. Numeric hyperparameters are like the knobs in the cockpit: you can tune them continuously. `n_estimators` is numeric. Categorical hyperparameters are like the switches in the cockpit: they can take on (two or more) distinct values. `criterion` is categorical.

It's a very elegant analogy! But is it helpful?

And furthermore, what is my hidden motivation for wanting to include it? Elegance, art, and the pursuit of higher beauty? Or *making myself look smart*? So maybe another name for this principle could be **It's not about you.**

Part 2: 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

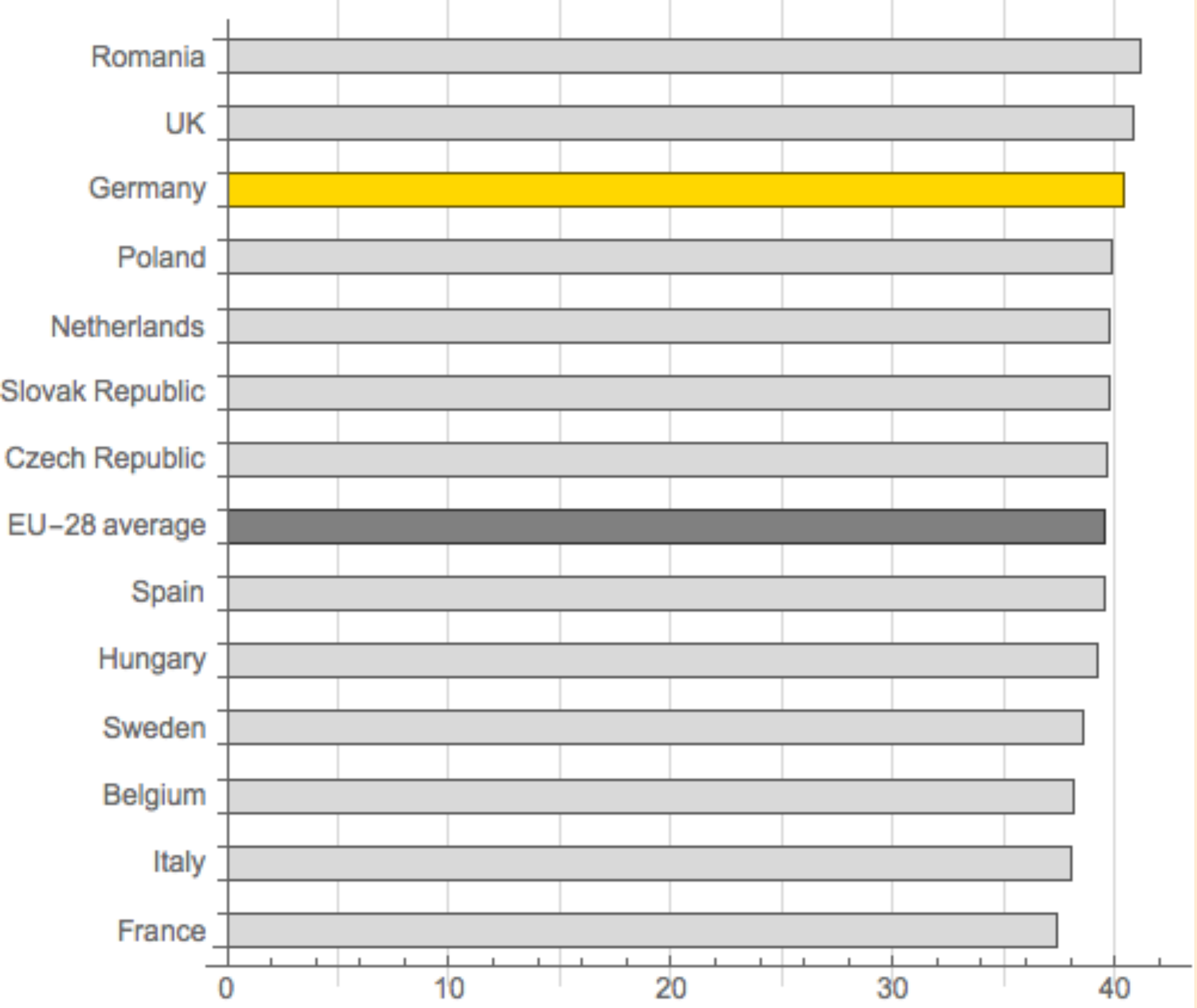
B. Somewhat Confident

C. Not Confident

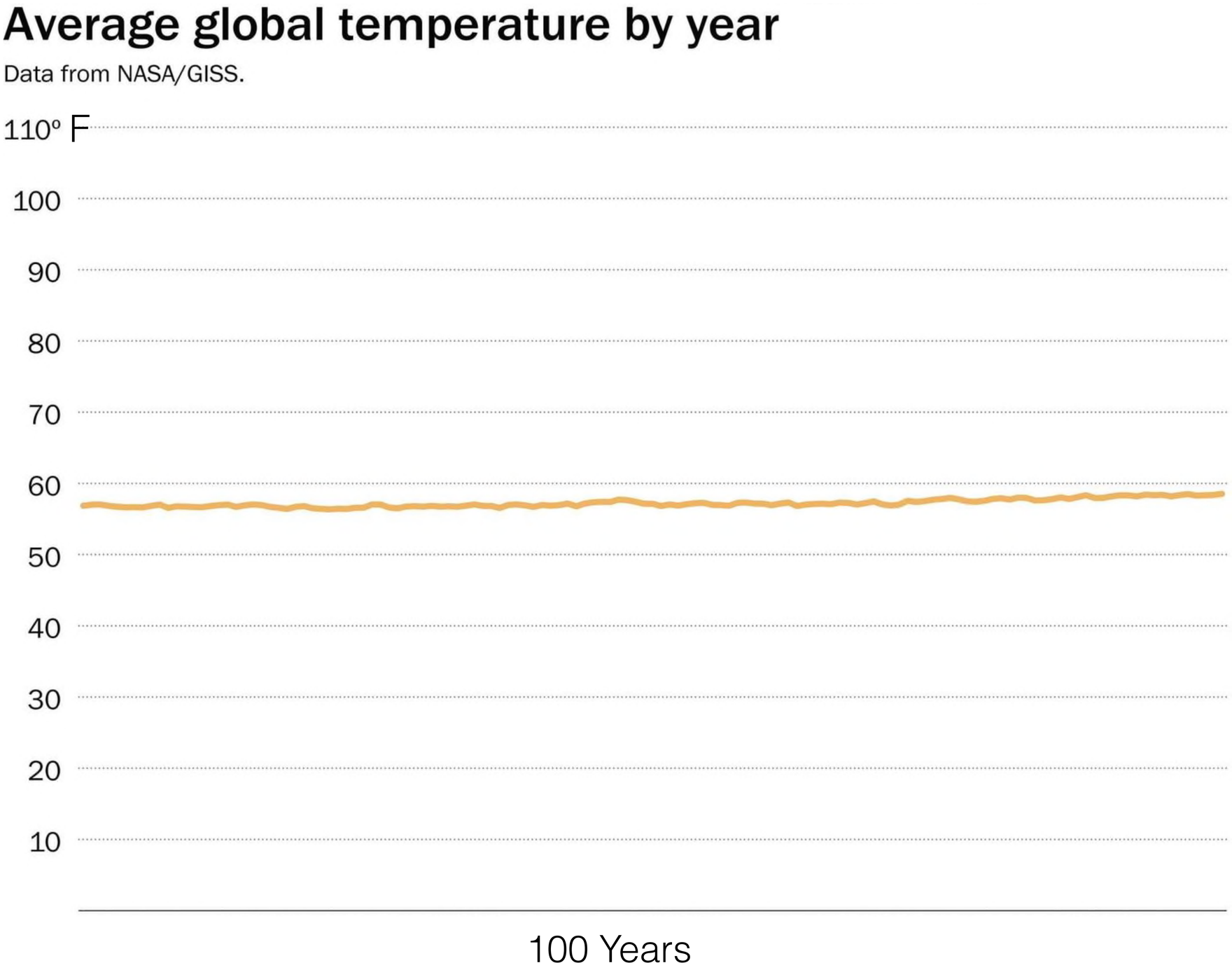
D. I need to do Machine Learning first.

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. Somewhat Confident

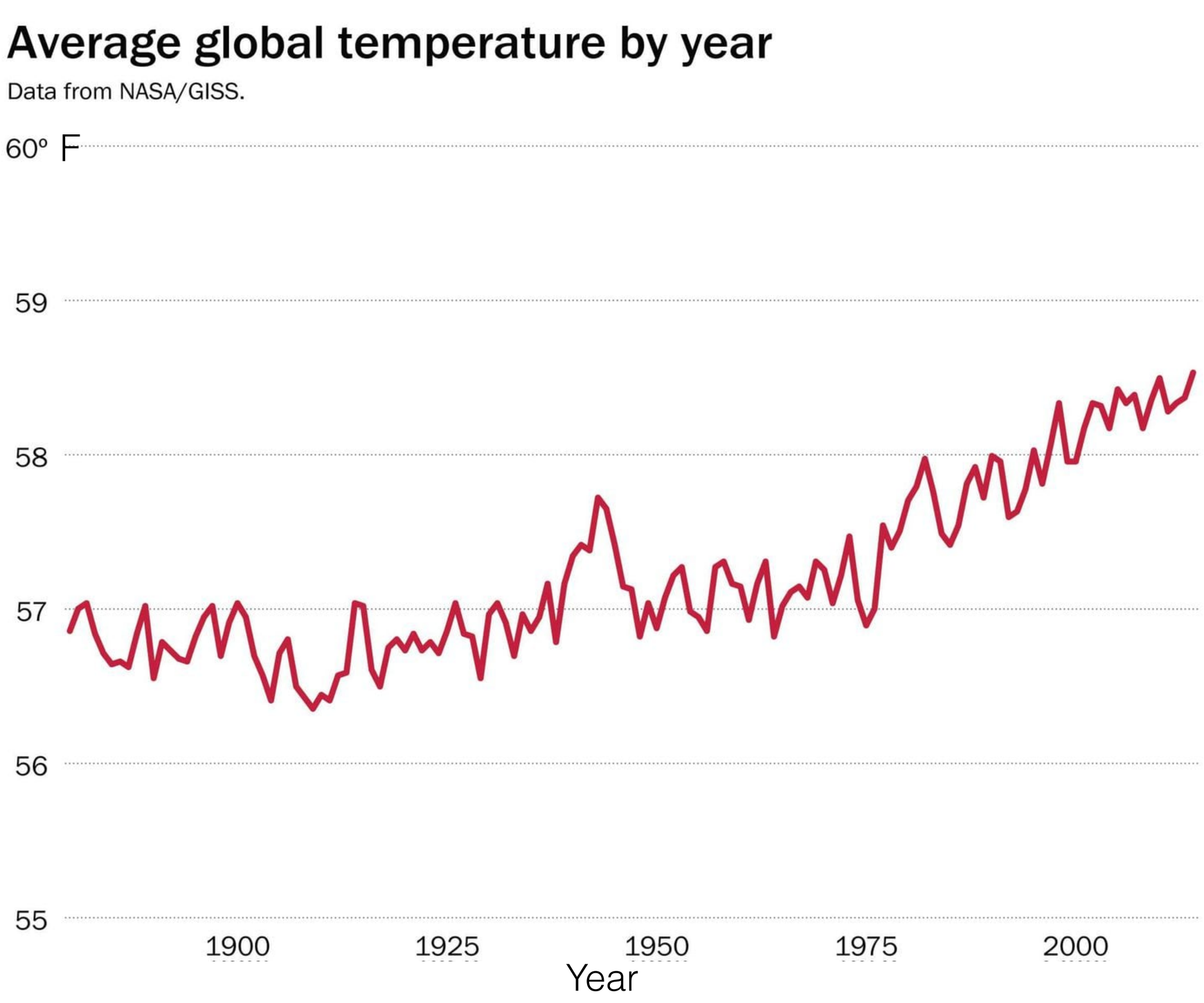
C. Not Confident

D. I need to do Machine Learning first.

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 ↓

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. Somewhat Confident

C. Not Confident

D. I need to do Machine Learning first.

Gender parity in the Olympics (100m race)

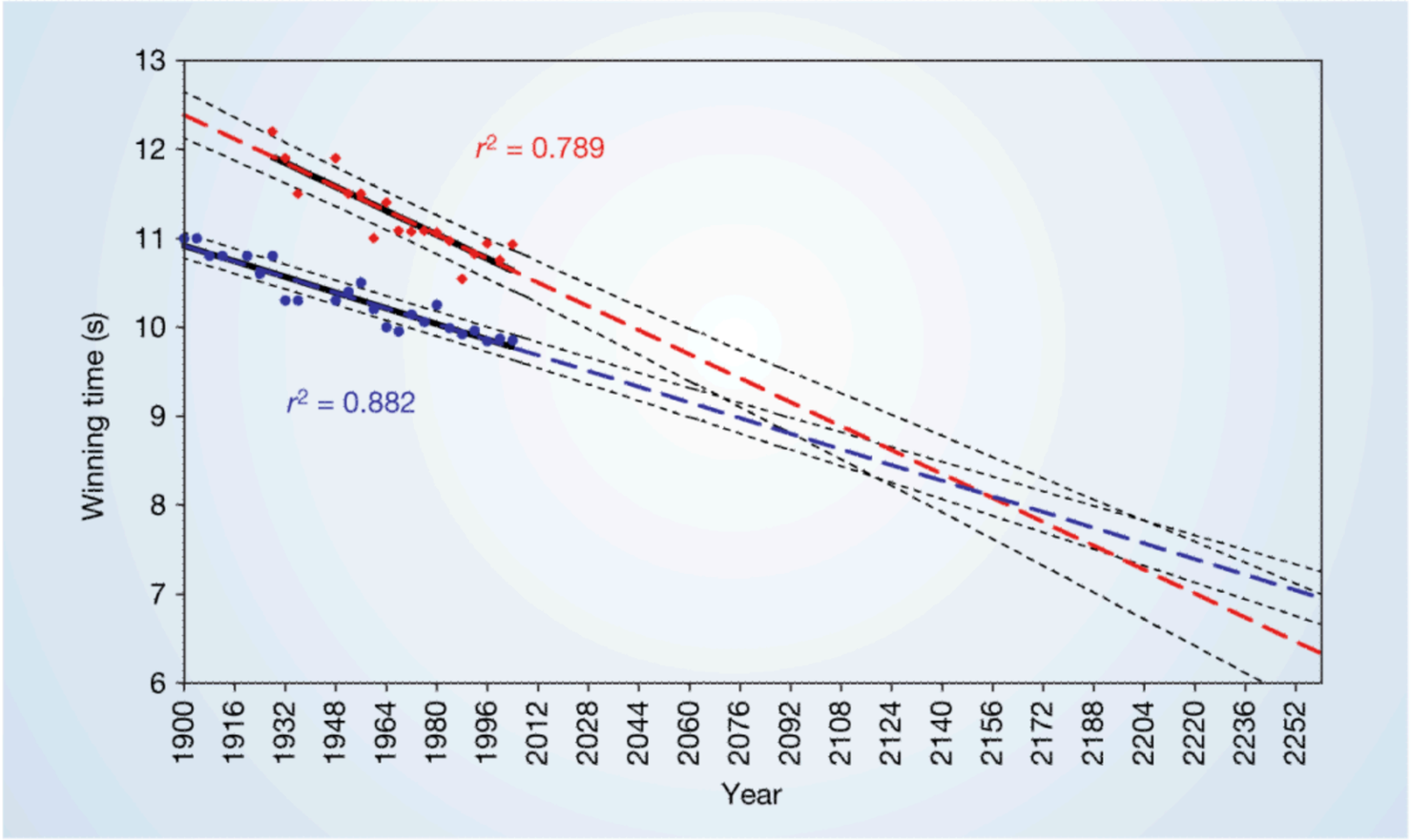
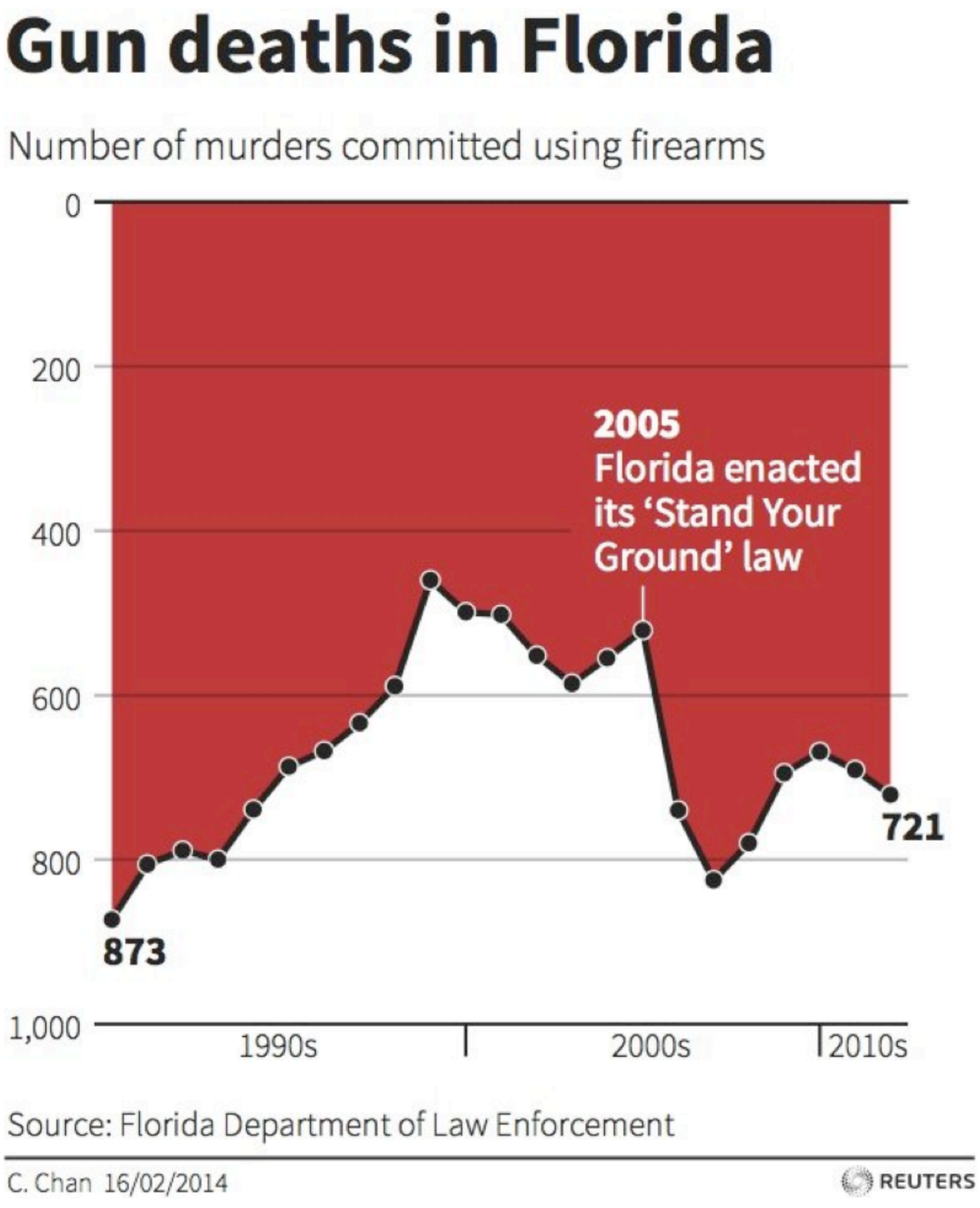


Figure 1.
The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) 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

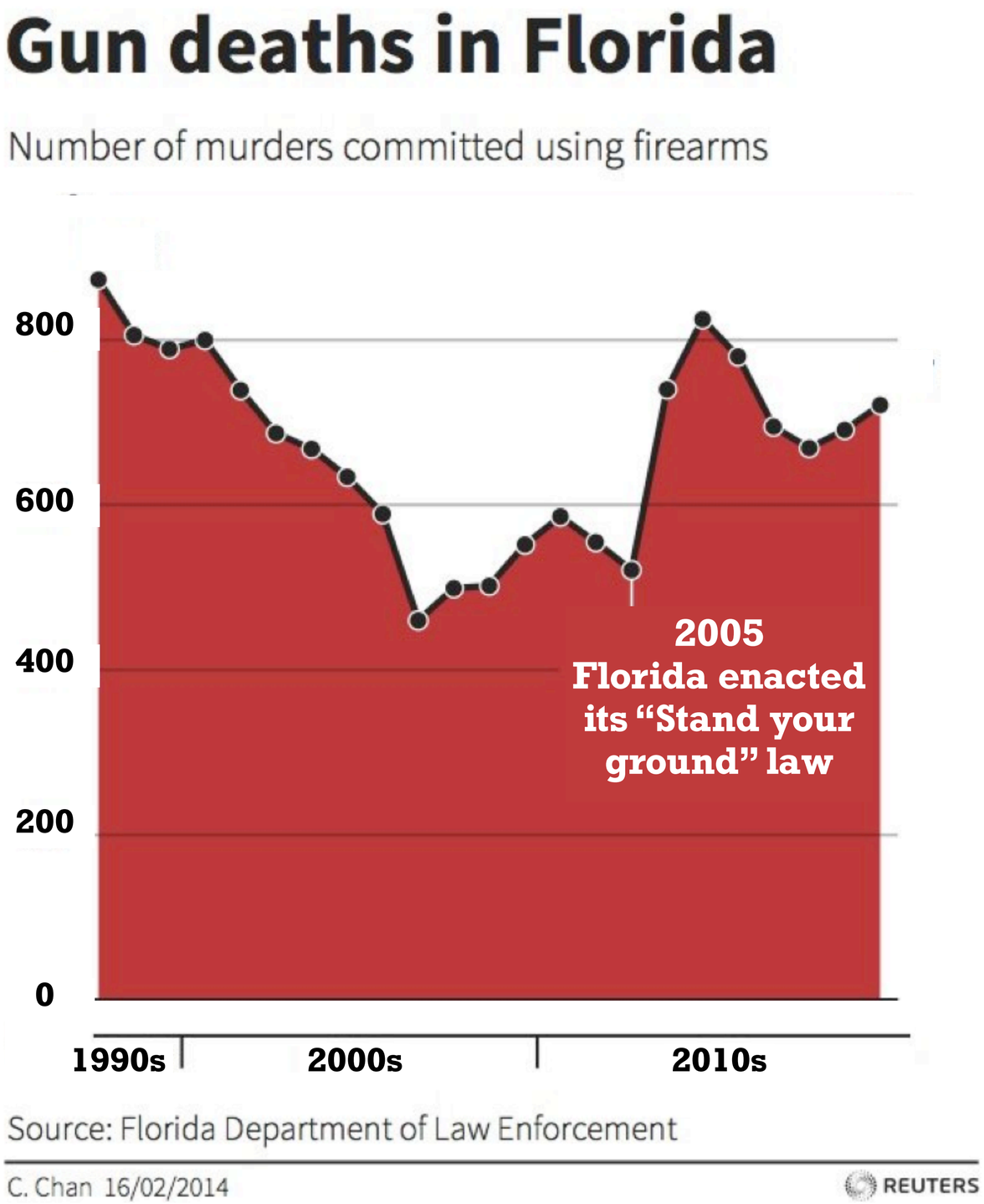
B. Somewhat Confident

C. Not Confident

D. I need to do Machine Learning first.

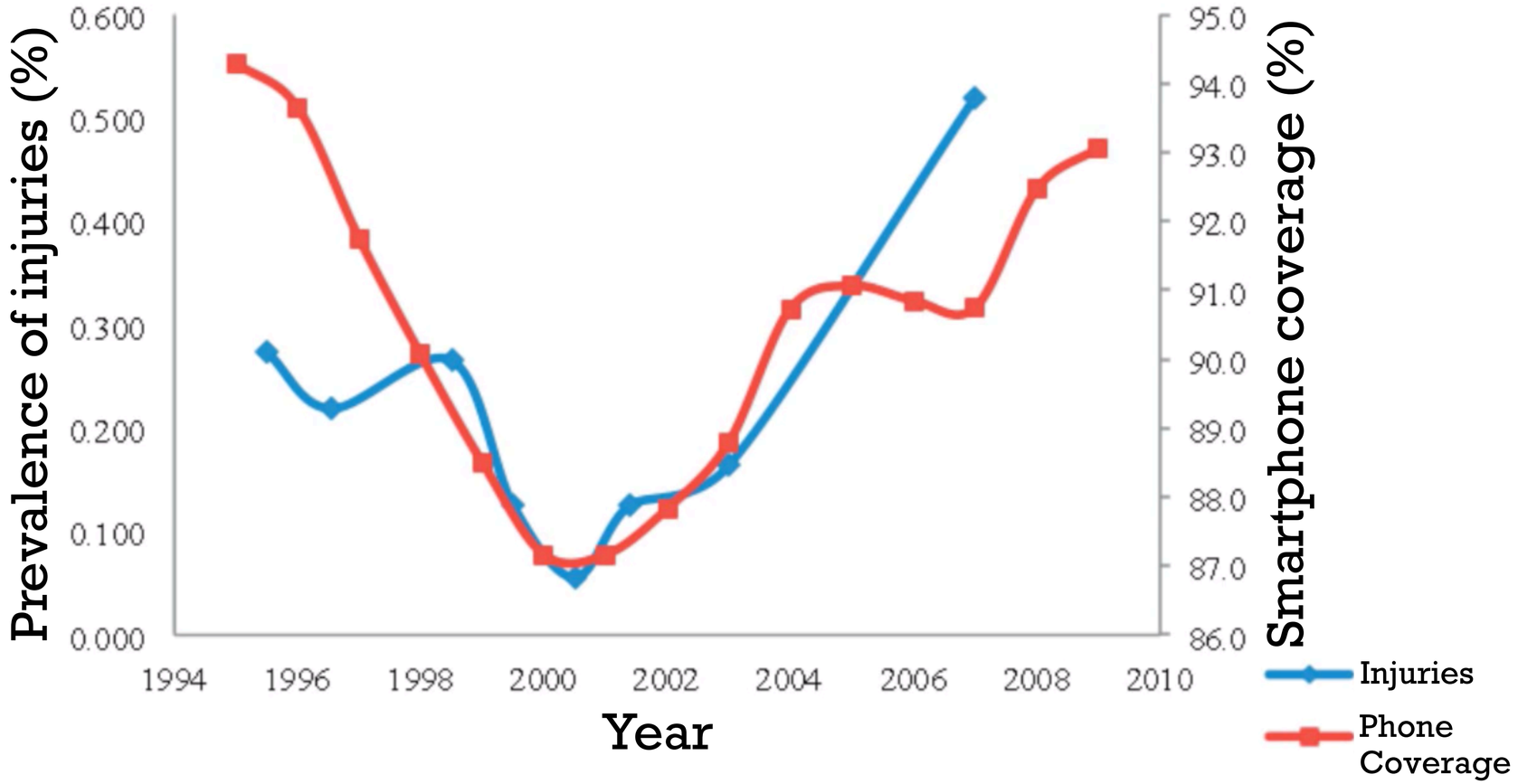
Data Source: Florida Department of Law Enforcement
Example Source: Callingbull.org & Reuters

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

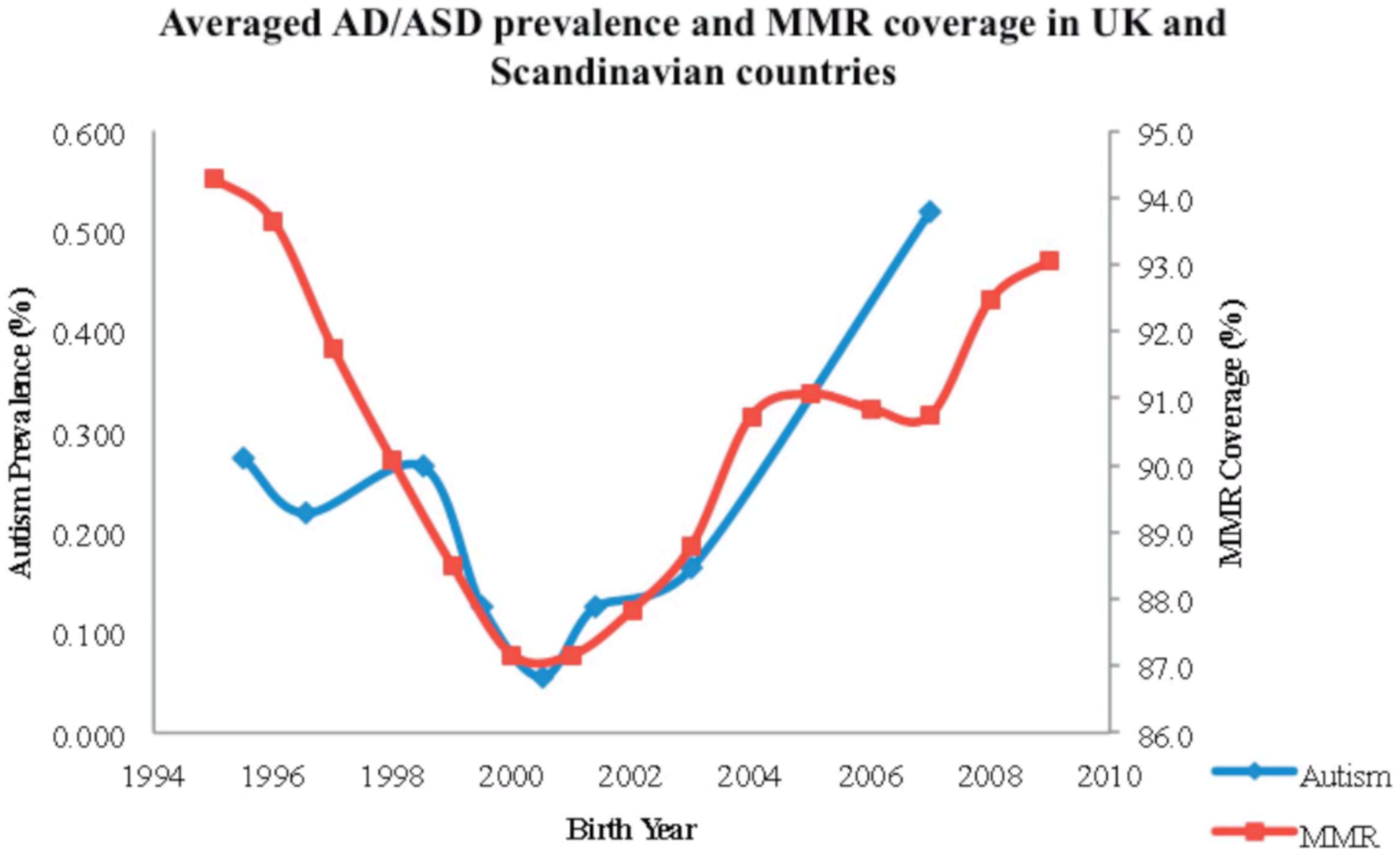
B. Somewhat Confident

C. Not Confident

D. I need to do Machine Learning first.

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

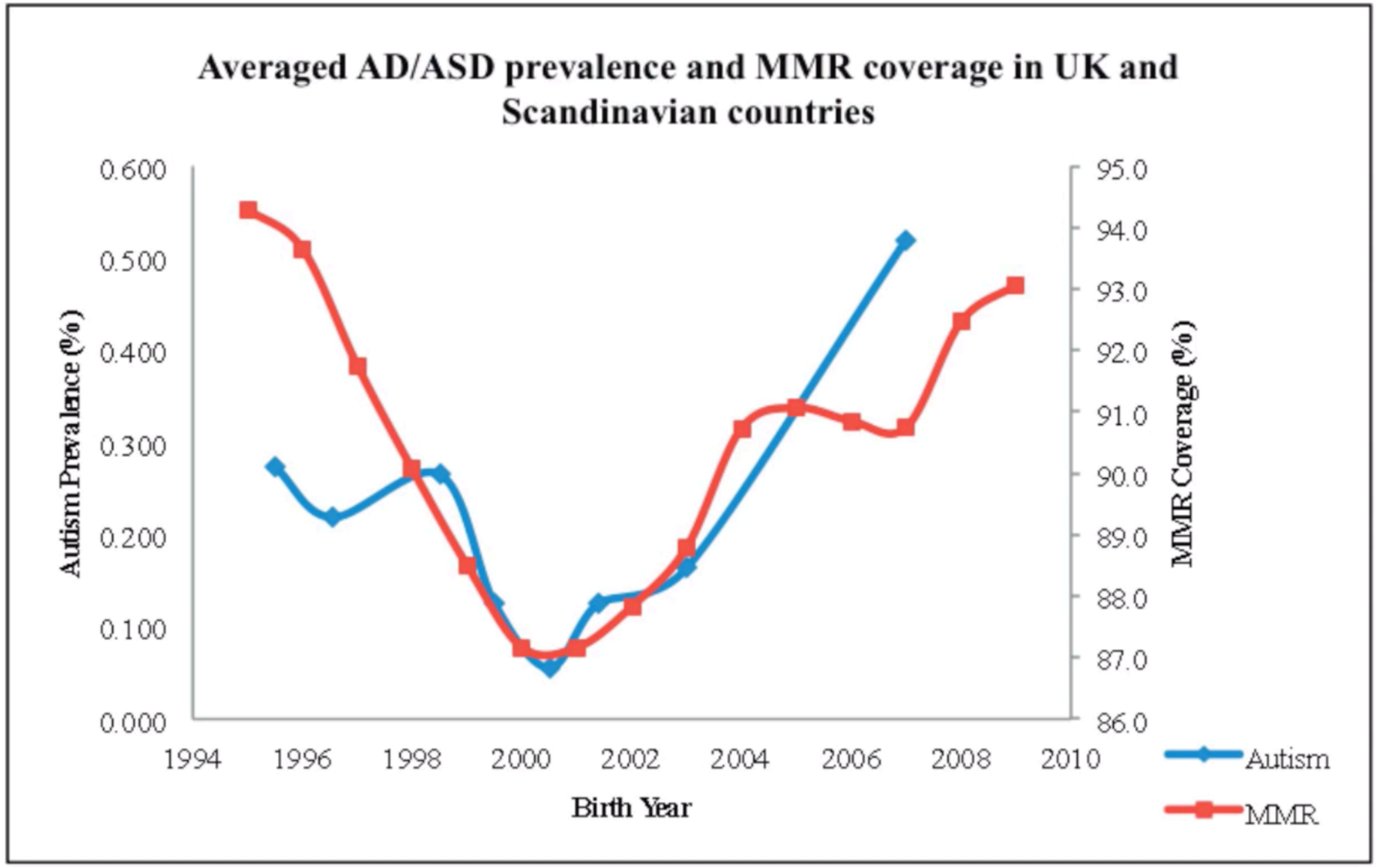


Figure 1-Averaged AD/ASD prevalence and MMR coverage in UK, Norway and Sweden. Both MMR and AD/ASD data are normalized to the maximum coverage/prevalence during the time period of this analysis.

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

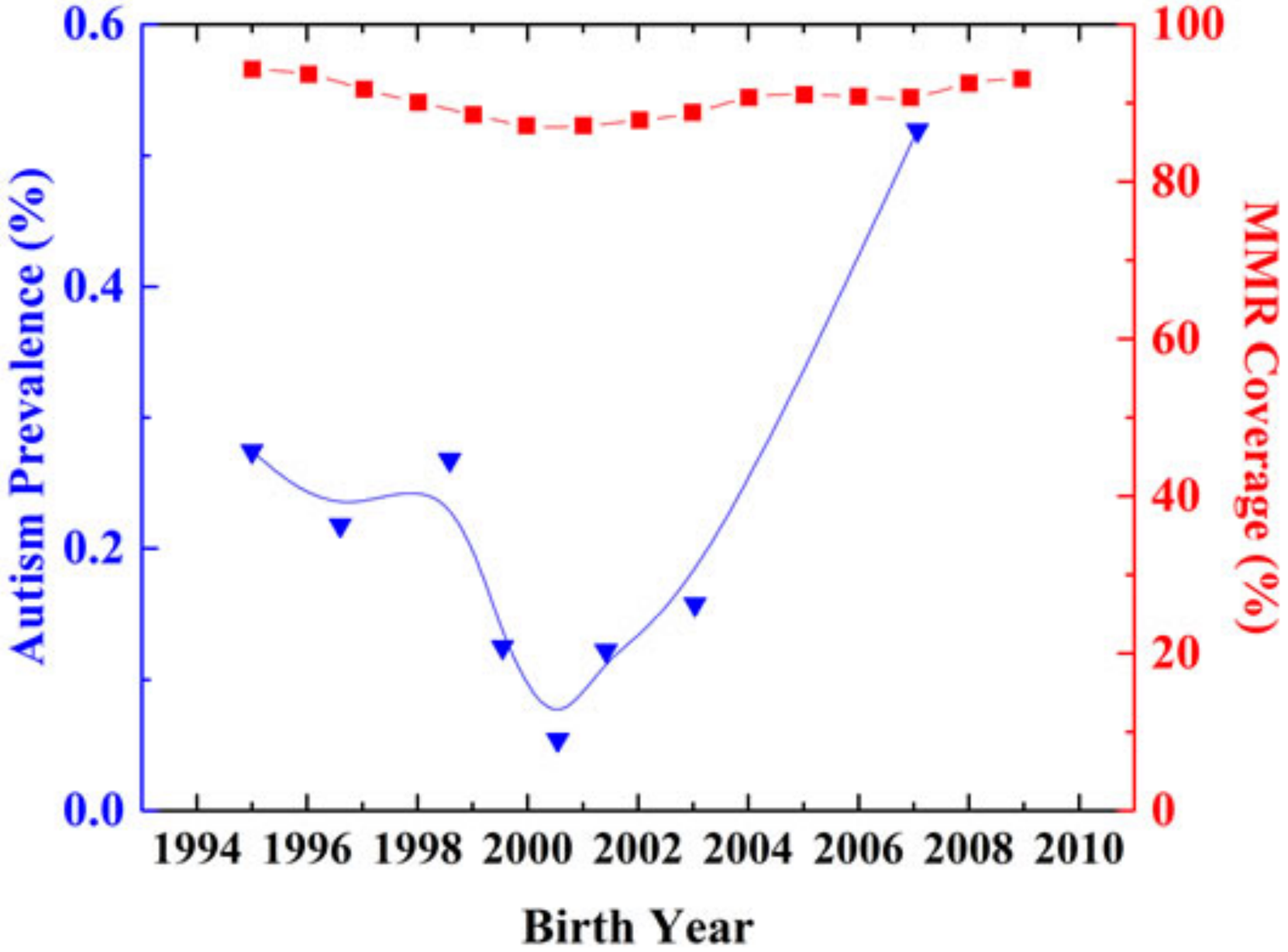
C. Not Confident

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

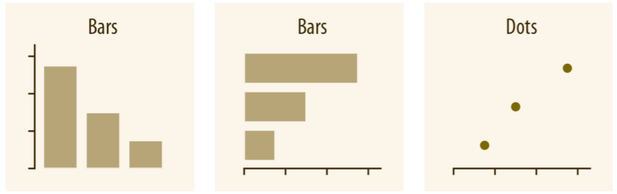
10 minutes

Part 3: Principles of Effective Visualizations

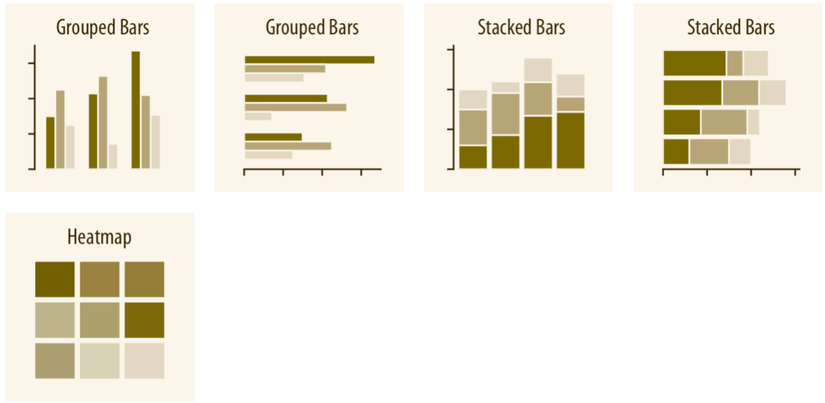
Directory of Visualizations

visualize data. It is meant both to serve as a table of contents, in case you are looking for a particular visualization whose name you may not know, and as a source of inspiration, if you need to find alternatives to the figures you routinely make.

5.1 Amounts



The most common approach to visualizing amounts (i.e., numerical values shown for some set of categories) is using bars, either vertically or horizontally arranged (Chapter 6). However, instead of using bars, we can also place dots at the location where the corresponding bar would end (Chapter 6).



If there are two or more sets of categories for which we want to show amounts, we can group or stack the bars (Chapter 6). We can also map the categories onto the *x* and *y* axis and show amounts by color, via a heatmap (Chapter 6).

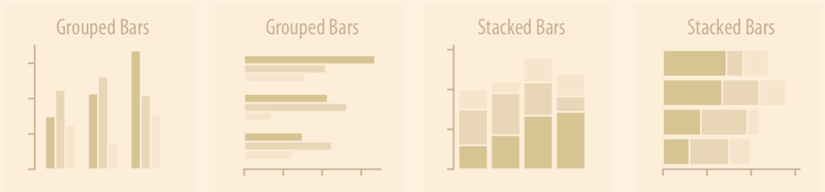
Source: Fundamentals of Data Visualization

Directory of Visualizations

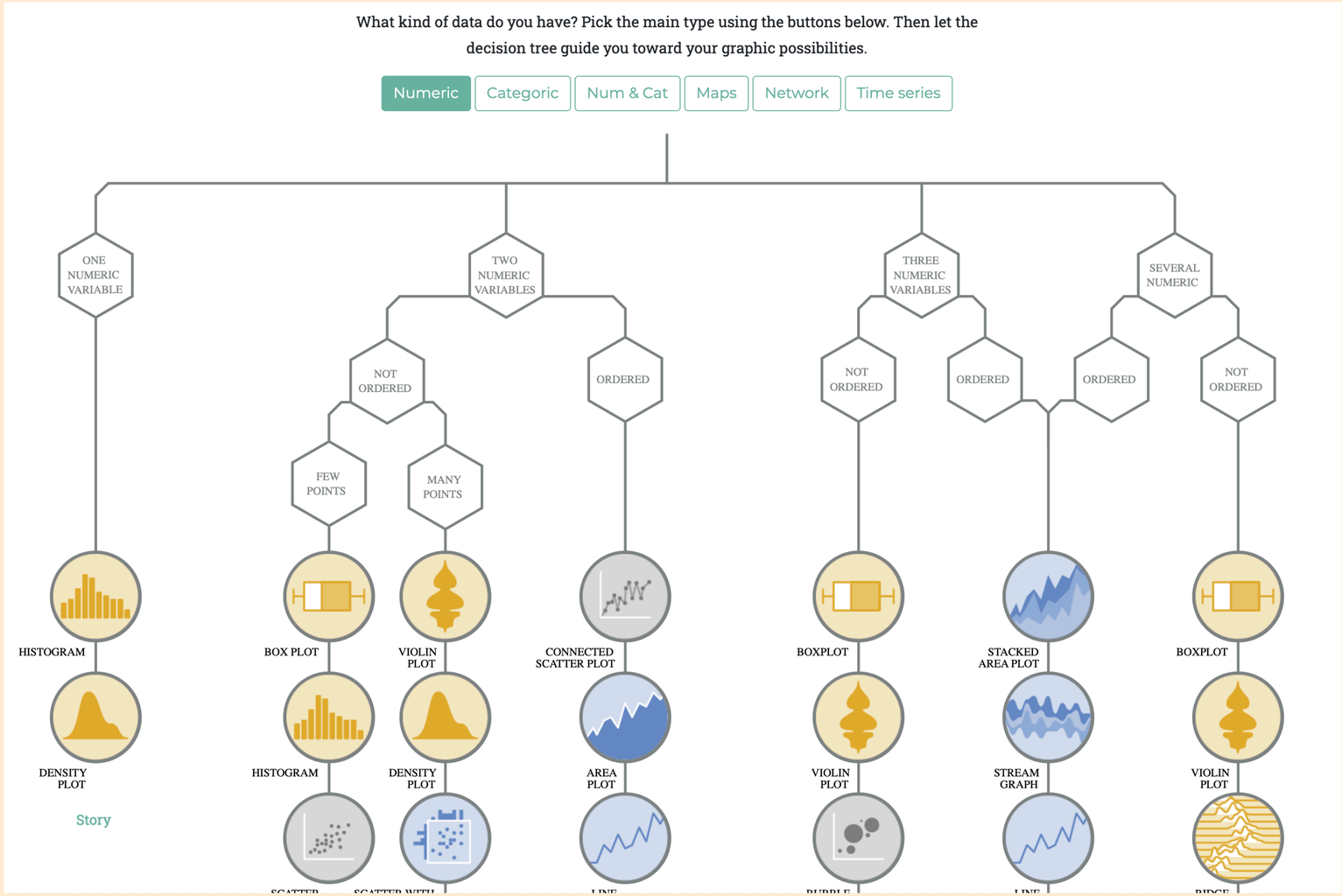
visualize data. It is meant both to serve as a table of contents, in case you are looking for a particular visualization whose name you may not know, and as a source of inspiration, if you need to find alternatives to the figures you routinely make.

5.1 Amounts

The most common approach to visualizing amounts (i.e., numerical values shown for some set of categories) is using bars, either vertically or horizontally arranged (Chapter 6). However, instead of using bars, we can also place dots at the location where the corresponding bar would end (Chapter 6).

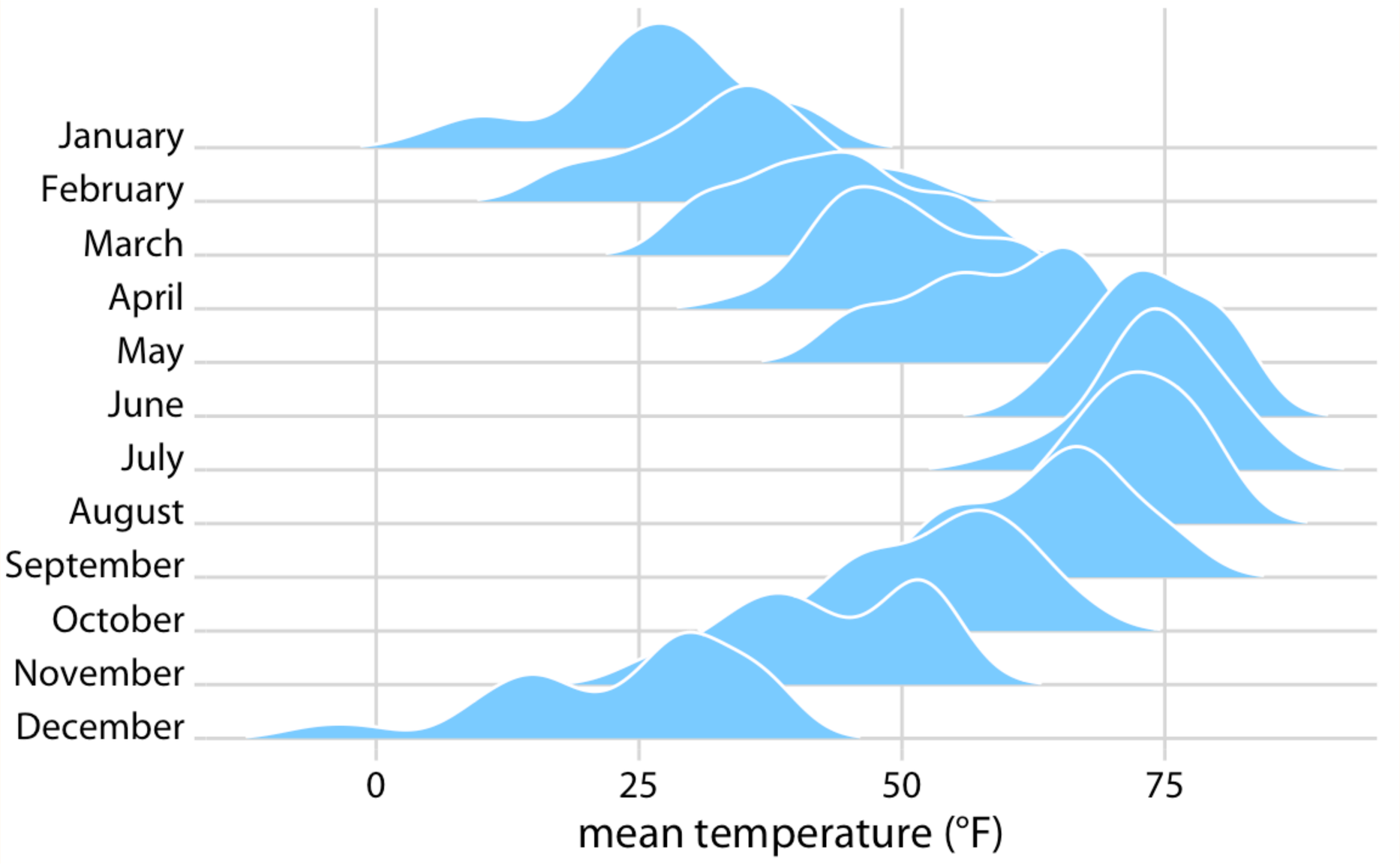


If there are two or more sets of categories for which we want to show amounts, we can group or stack the bars (Chapter 6). We can also map the categories onto the *x* and *y* axis and show amounts by color, via a heatmap (Chapter 6).



Source: Data to Viz

My favourite: Ridgeline plot



Source: Fig 9.9 of Fundamentals of Data Visualization

Principles of Effective Visualizations

Principle	Definition	Examples
• Proportional Ink	The amount of ink used to indicate a value should be proportional to the value itself.	Truncating the y-axis on a bar chart to exaggerate the difference between bars violates the principle of proportional ink.
• Data:ink ratio	Remove distracting visual elements to focus attention on the data	Lighten line weights, remove backgrounds, never use 3D or special effects, ~~remove~~ avoid unnecessary/redundant labels.
• Labels & legends	Use axes labels and titles to highlight/ communicate data	Never leave your data column names as axes labels! Generally good to add a title.
• Overplotting	With large datasets, points overlap, resulting in large clouds of data	To fix overplotting, could plot just a sample subset of the data, use alpha, and use smaller points. Or, jitter - but check if appropriate!
• Visualization choice	Must be informed by the **data** you have, the **research question** being asked and the **audience** that cares.	Pick the simplest plot that best shows most/all of the data needed to answer the research question. If you only have summary statistics, cannot show distributions. Tailor the visualization to your audience (within reason) but don't dumb it down.
• Colour & Accessibility	Colour can be used to encode information or for aesthetics/style/ design. However, colour can also be distracting if used inappropriately or poorly.	Choose a perceptually uniform colour palette; can be sequential or diverging for quantitative data. Opt for colour-blind friendly palettes. Categorical data can use qualitative colour schemes.

The most important one!

Principle	Definition	Examples
• Visualization choice	Must be informed by: 1) the **data** you have, 2) the **research question** being asked and 3) the **audience** that cares	- Summary statistics >> do not show distributions - Pick the simplest plot that best shows most/all of the data needed to answer the research question - Tailor the visualization to your audience (within reason)