Welcome to DATA 202: Data Science 2

1. Please show your *CampusClear* status and green wristband as you enter
2. We need a volunteer *Remote Student Representative*
    - Watch the Teams chat
    - Alert instructor to questions or problems
3. Fill out attendance sheet

An opening prayer for a unique semester

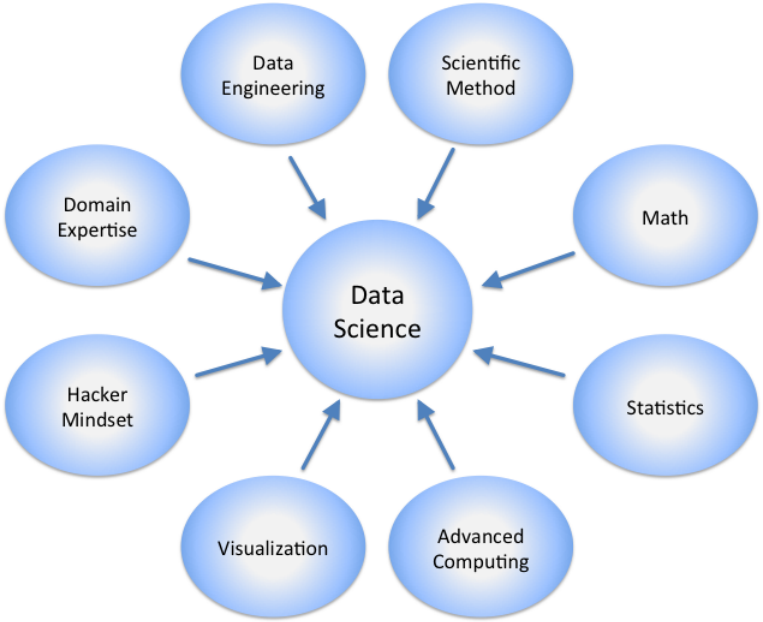
From the apostle Paul's letter to the Philippians:

> This is my prayer:
> that your love may abound more and more
> in knowledge and depth of insight,
> so that you may be able to discern what is best and may be pure
> and blameless for the day of Christ, filled with the fruit of
> righteousness that comes through Jesus Christ—to the glory and
> praise of God.

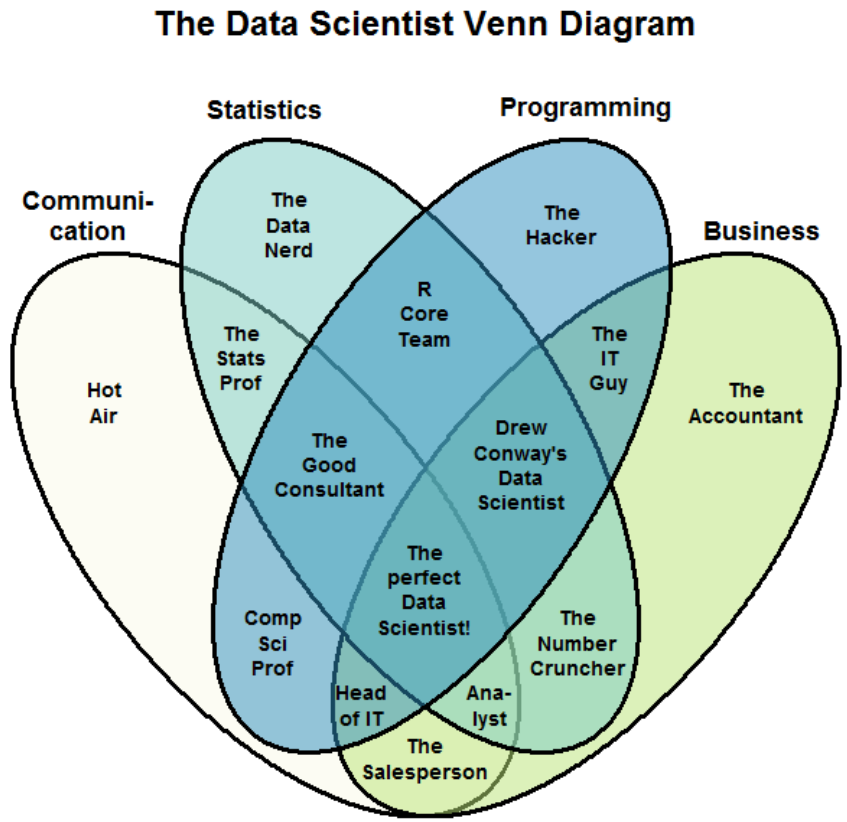
What is Data Science?

What is Data Science?

- **Data**: information, collected systematically
- **Science**: systematic study of that data



https://commons.wikimedia.org/wiki/File:DataScienceDisciplir



https://commons.wikimedia.org/wiki/File:Data_scientist_V

How does data science help you see?

How does data science help you see?

Visualization

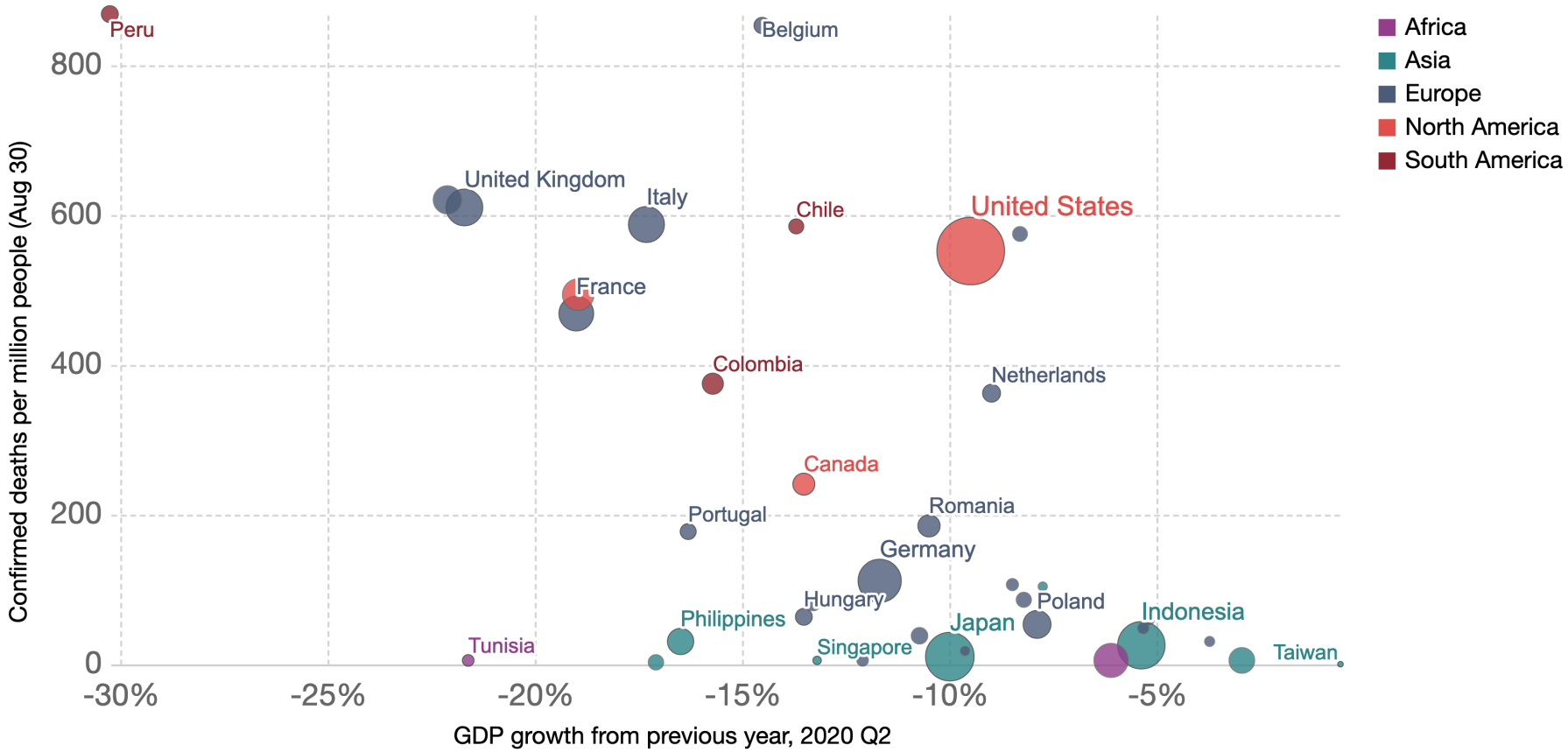
Inference

Prediction

Visualization

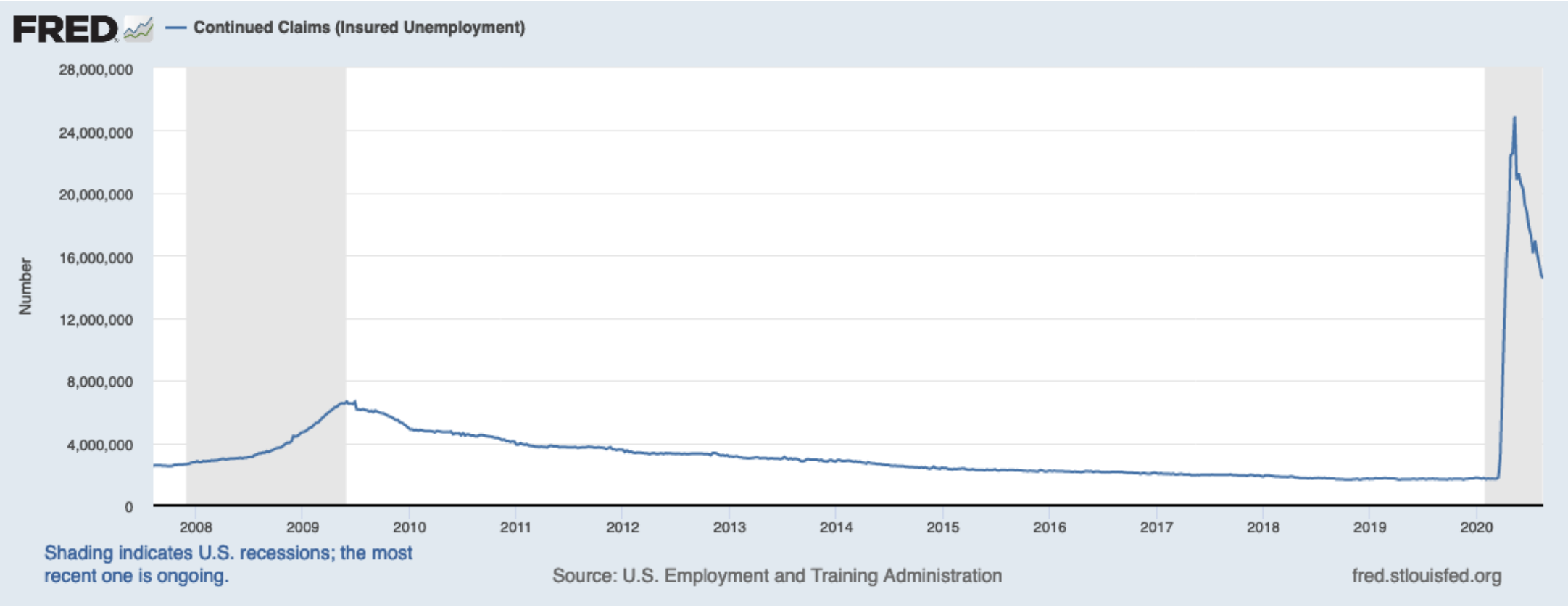
Economic decline in the second quarter of 2020 vs rate of confirmed deaths due to COVID-19

The vertical axis shows the number of COVID-19 deaths per million, as of August 30. The horizontal axis shows the percentage decline of GDP relative to the same quarter in 2019. It is adjusted for inflation.



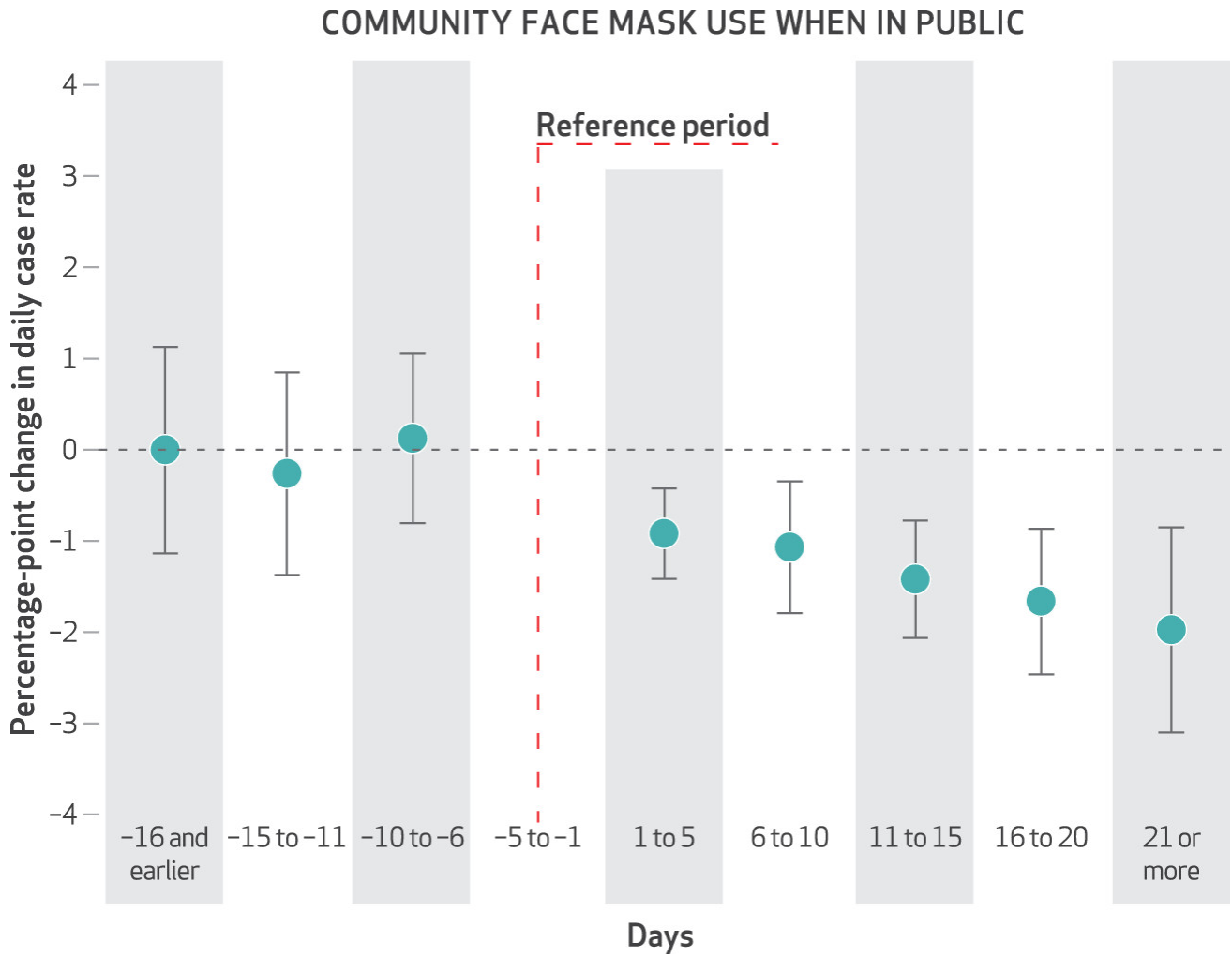
Source: European CDC, Eurostat, OECD and individual national statistics agencies

Note: Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the true number of deaths from COVID-19. Data for China is not shown given the earlier timing of its economic downturn. The country saw positive growth of 3.2% in Q2 preceded by a fall of 6.8% in Q1.



Source: https://fred.stlouisfed.org/graph/?g=v4rP

Inference



Wei Lyu and George L. Wehby. *Community Use Of Face Masks And COVID-19: Evidence From A Natural Experiment Of State Mandates In The US*. https://doi.org/10.1377/hlthaff.2020.00818

Prediction

2016 Honda Odyssey

EX-L 4dr Minivan (3.5L 6cyl 6A)

Mileage	37,183
Condition	Outstanding
Exterior	Modern Steel Metallic

Your appraisal As of 10/26/2019

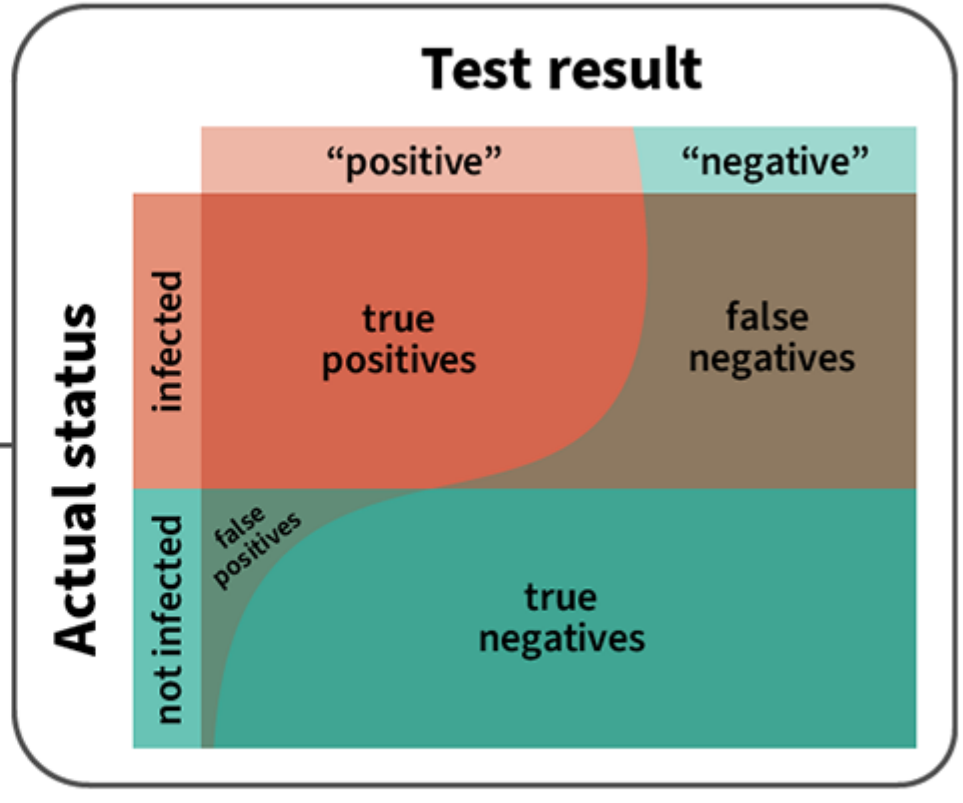
Email report

	Trade-In	Private Party	Dealer Retail
	$21,645		**$23,731**
			$25,666
National Base Price ⓘ	$19,676	$21,694	$23,484
Color Adjustment ⓘ	-$43	-$48	-$52
Regional Adjustment ⓘ	$103	$114	$123
Mileage Adjustment ⓘ	$680	$680	$680
Condition	$1,229	$1,291	$1,431

Classification

The COVID-19 swab test is highly **specific** but not as **sensitive.**

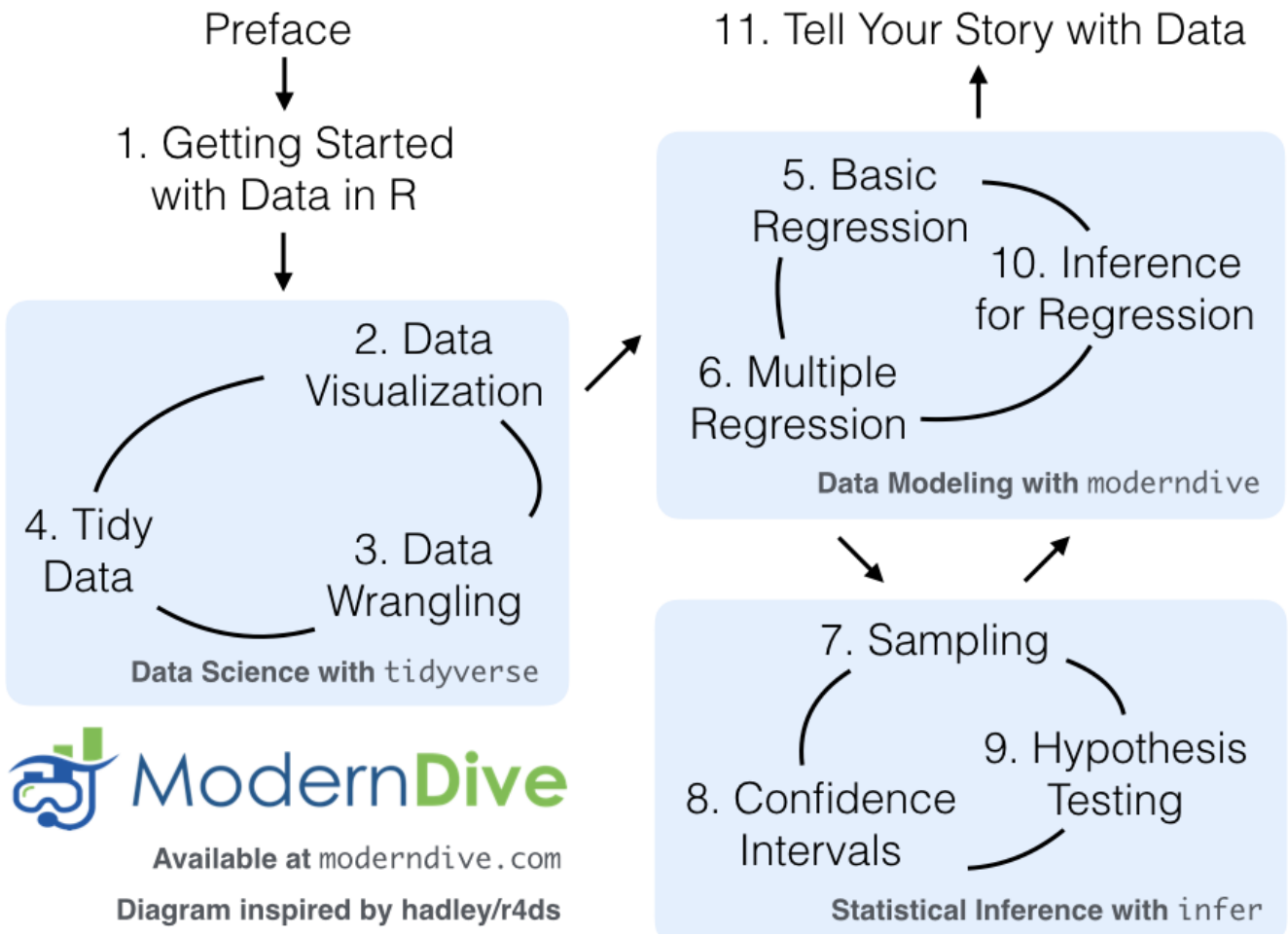
That means a positive result is almost always true, but a negative result is sometimes false.



Sensitivity = number of **true positives** / number of those tested who **really are infected** = **"how many of the infections did we find?"**

Specificity = number of **true negatives** / number of those tested who **really are not infected** = **"how many of the healthy people did we clear?"**

The DS Life Cycle



Data Science 2 focuses on...

- wrangling
- predictive modeling and validation
- visualization and communication

but touches on all of the DS lifecycle.

Uses:

- R (tidyverse, tidymodels, ggplot/plotly) the *first* time we see something
- Python (Pandas, sklearn) the *second* time we see something
- occasionally: SQL, other tools according to student interest

Our Goals

- *Skill*: how to do these things
- *Knowledge*: understanding the underlying concepts
- *Character*: wisdom in practicing these skills

Character??

DS is a lens. How can we see rightly through it? Some areas:

- Humility
- Integrity / Honesty
- Hospitality
- Compassion and justice

Humility

Challenge: data feels powerful, people listen to what you use it to say.

So we will practice:

- Citing all sources (for both data and process)
- Acknowledging limitations
- Transparent process
- Validation of results

Integrity

It’s tempting to say something that isn’t entirely true, or to manipulate the collection/analysis/reporting process to yield the answer you want

So we will practice:

- Evaluating claims that others use data to make
- Clearly articulating our analysis decisions and rationale
- Using exploratory analytics to validate data against assumptions

Hospitality

We can choose to use our tools to elucidate and clarify, rather than obscure.

So we will practice:

- Clear visual communication
- Clarity of code and process
- Writing explanations that are accessible and appropriate to audience.

Compassion and Justice

Data Science can both cause harm and reveal it.

So we will:

- Study examples of how data might cause harm
- Study examples of how harm might be mitigated or revealed

Where is this course?

tiny.cc/data202

Who am I?

- Ken Arnold
- **Office:** NH 298
- **Office hours:** TBD (fill out the poll!)
- **Email:** ka37@calvin.edu (but post course questions on Teams!)

Let's do some data science!

Go here

Why R?

R gives names to concepts.

Python:

```
data[data.column_name > value]
```

R:

```
data %>% filter(column_name > value)
```

etc.

Why git?

- Reproducibility
- Hospitality
- Everybody uses it

Acknowledgments

Much of the first weeks of this course is adapted from introds.org by the excellent Dr. Mine Çetinkaya-Rundel at University of Edinburgh.