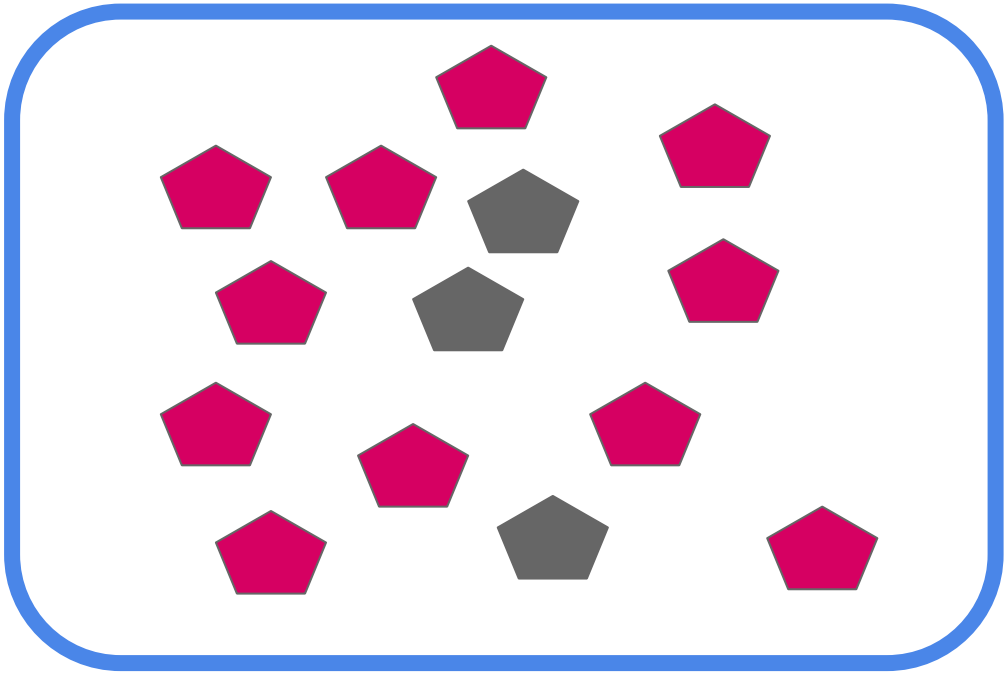
The central dogma of statistics

Jeff Leek

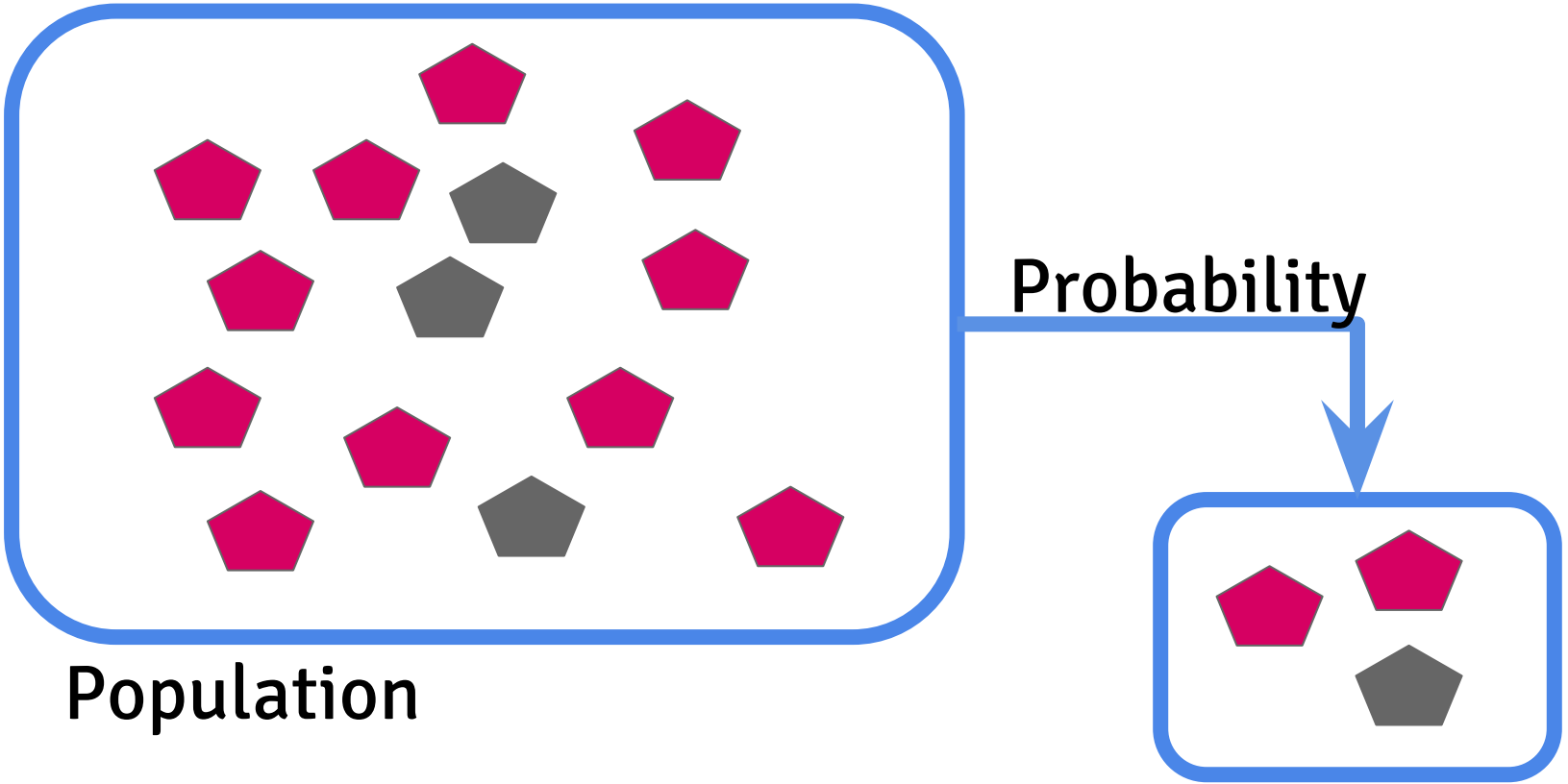
@jtleek

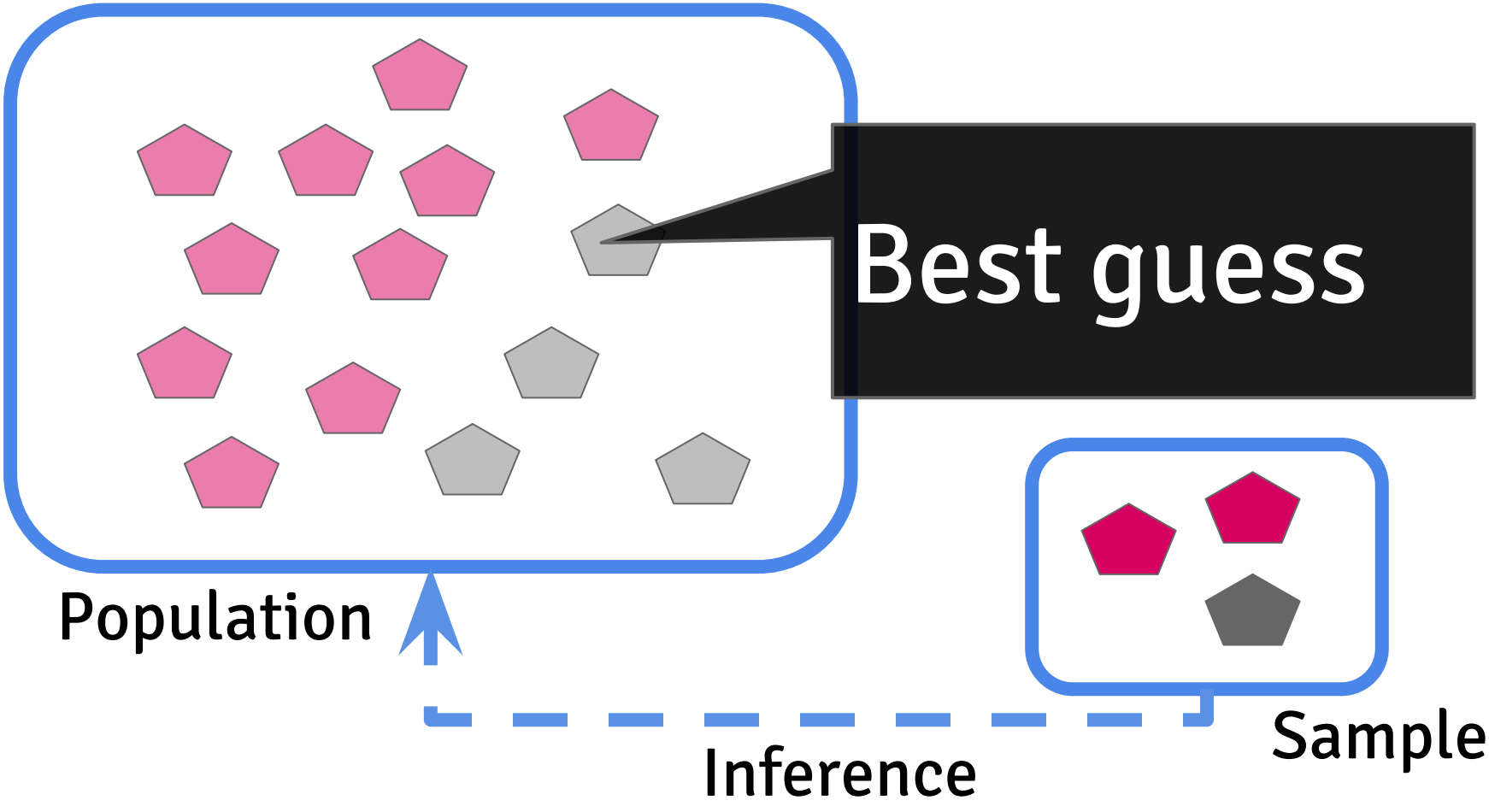
cc BY NC SA

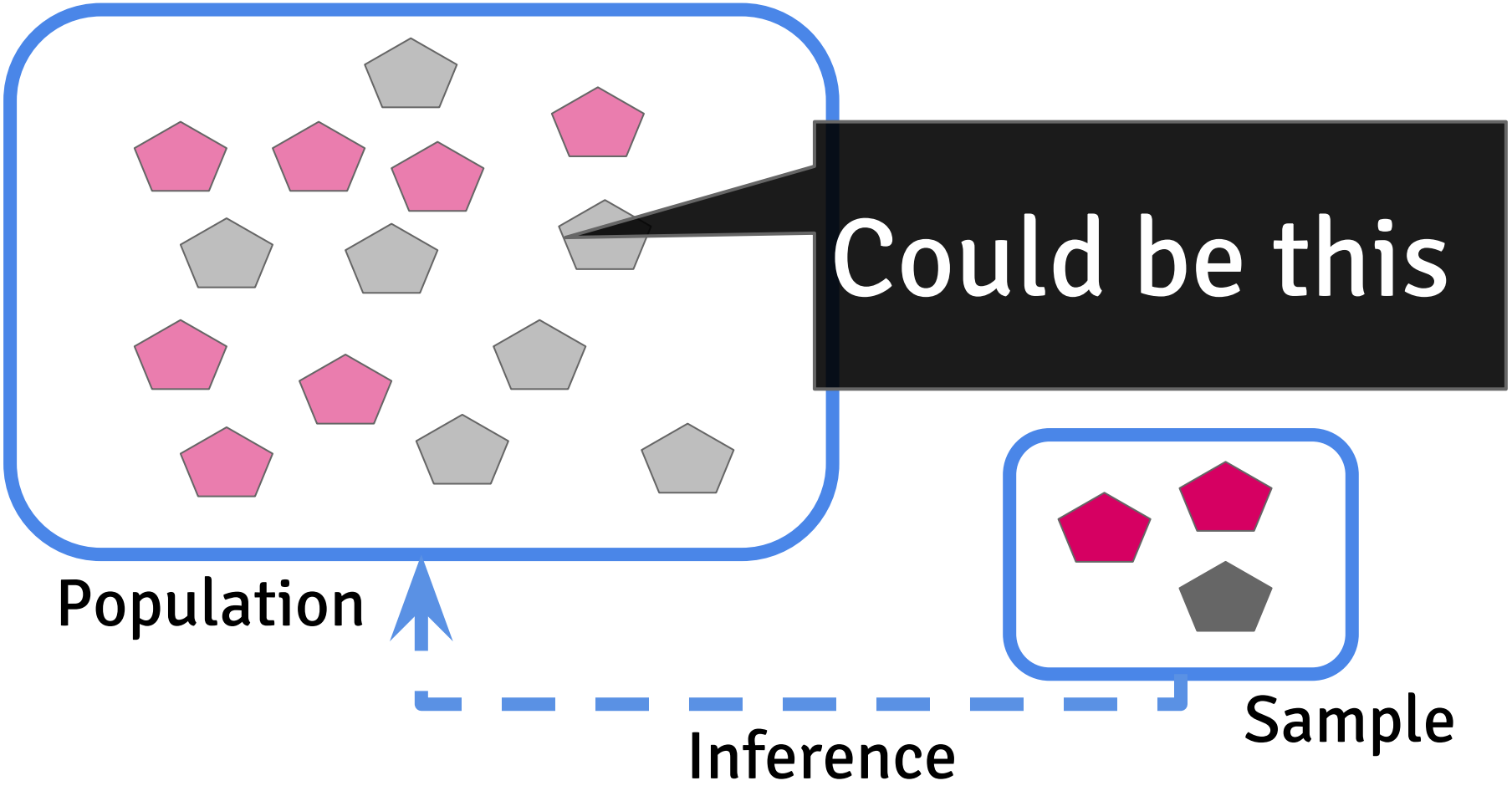
Central dogma of statistics

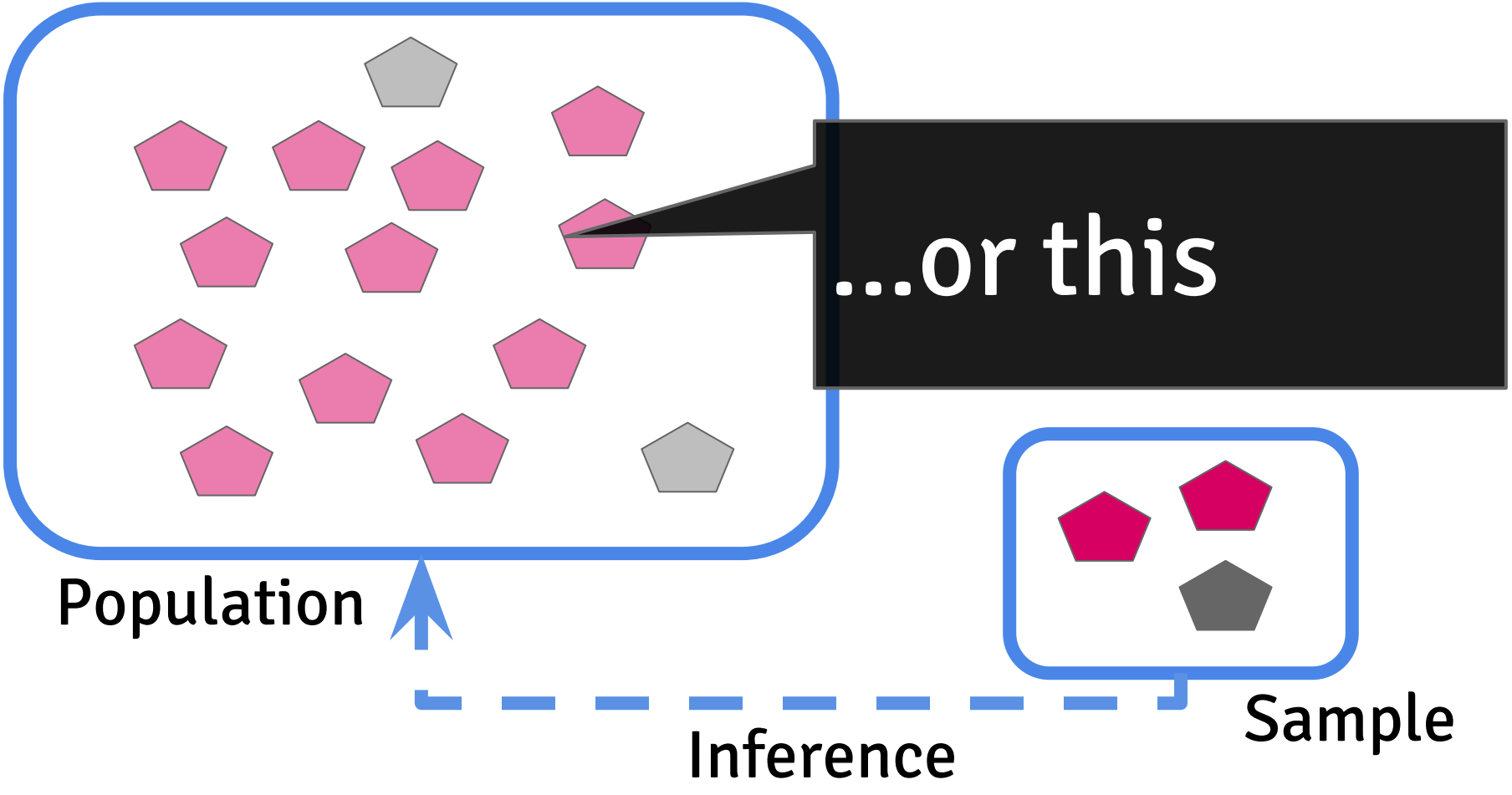


Population

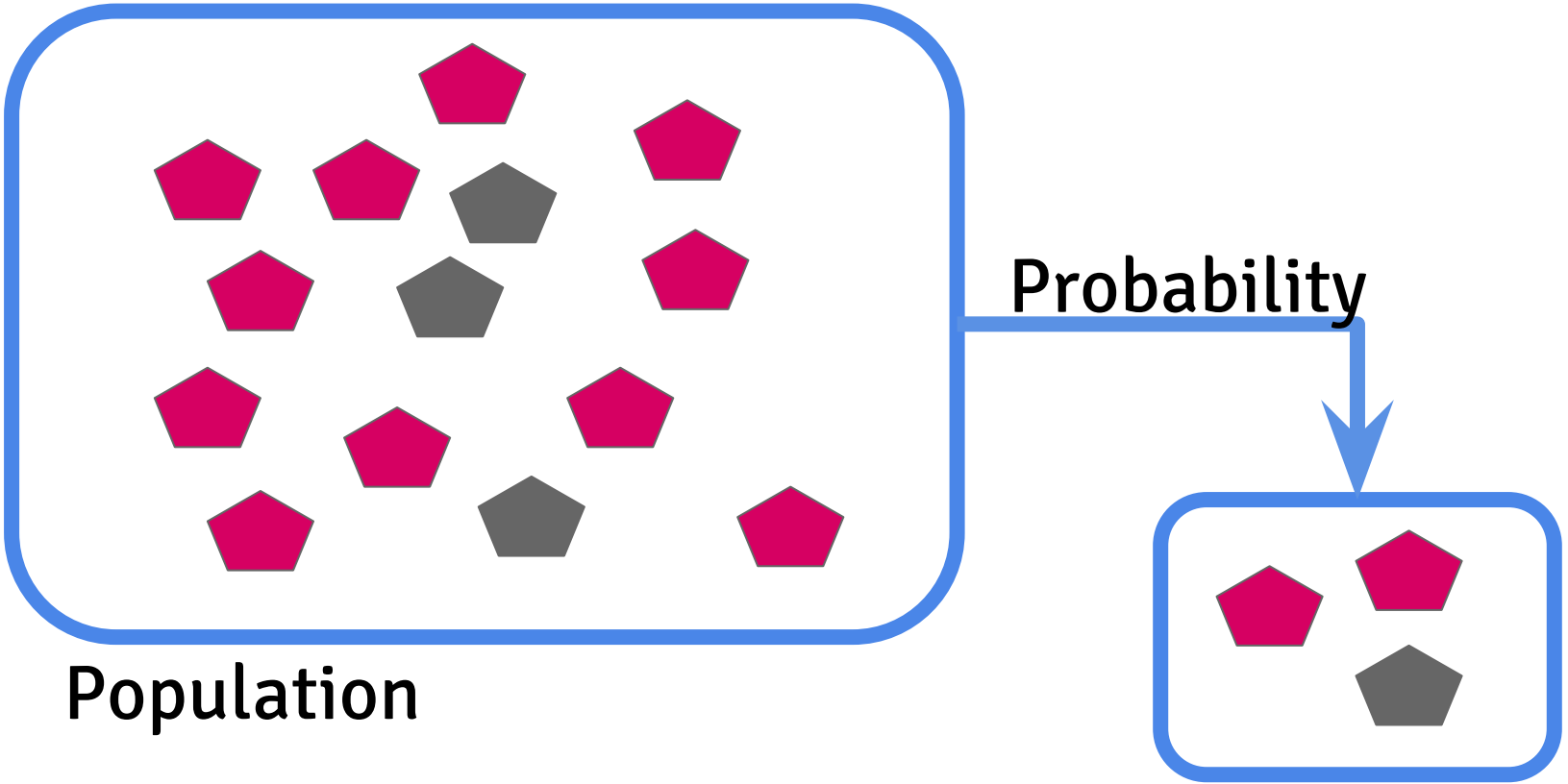

Probability
Population

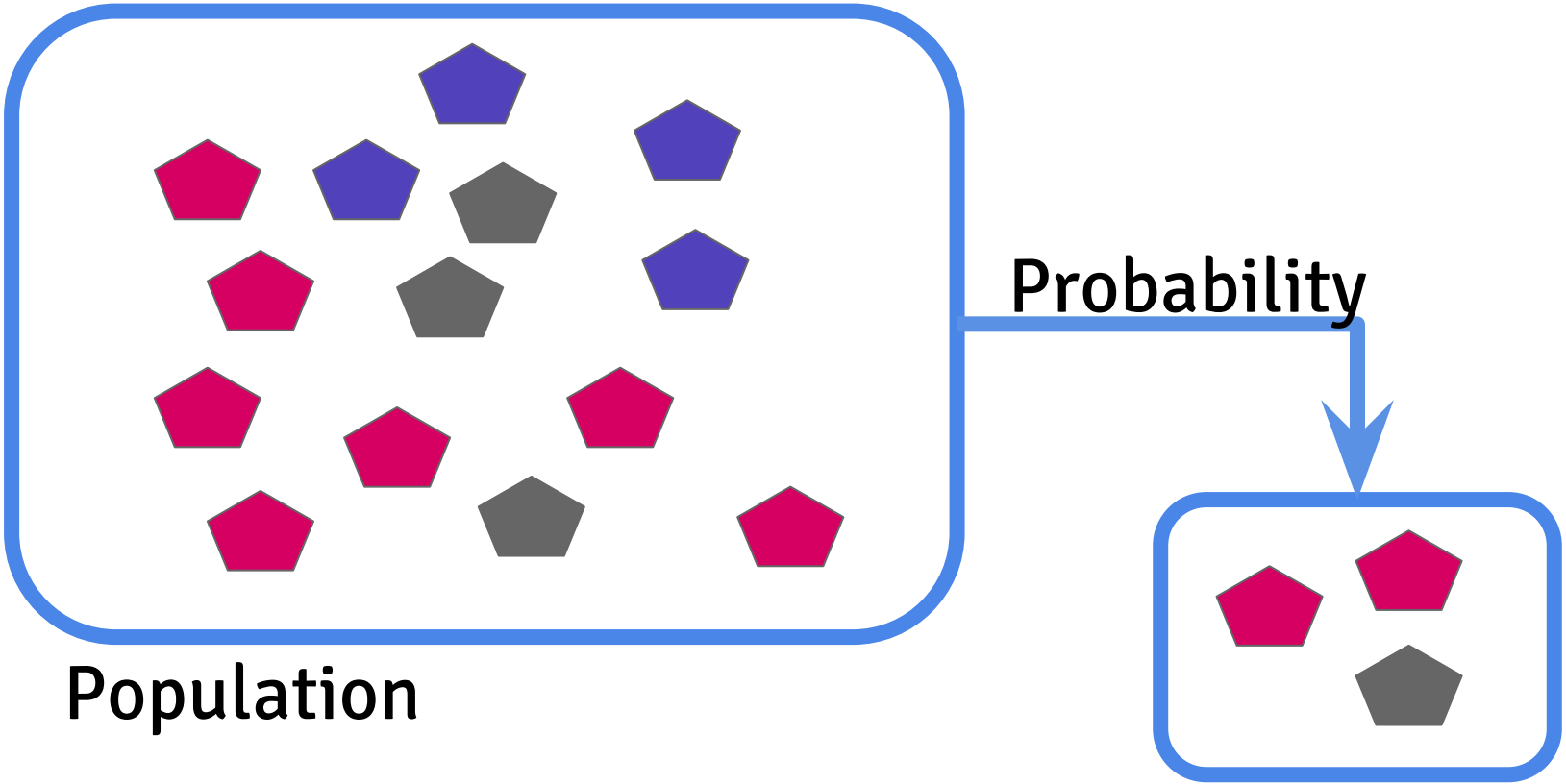

Best guess
Population
Sample
Inference

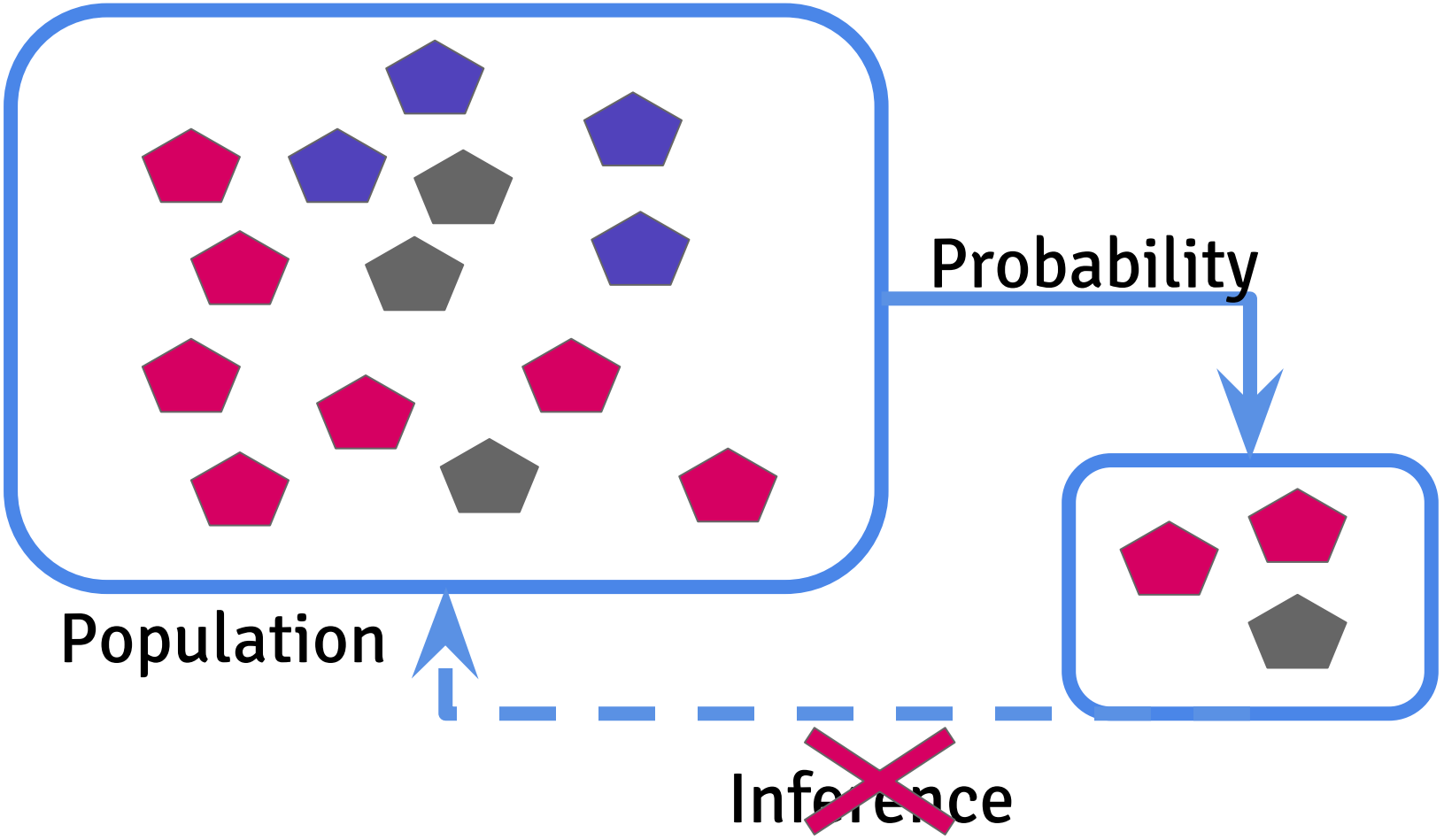

Could be this
Population
Sample
Inference


...or this
Population
Sample
Inference

Know the population


Probability
Population


Probability
Population


Probability
Population
Inference

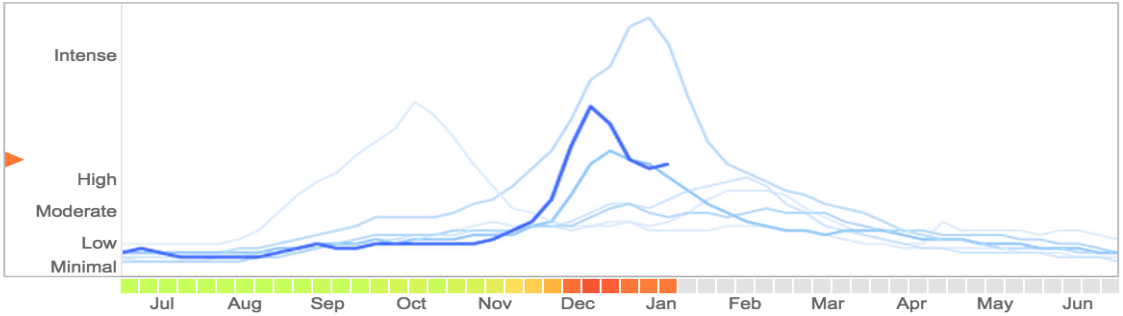
Google.org home
Dengue Trends
Flu Trends
Home
United States
National
Download data
How does this work?
FAQ

Explore flu trends - United States

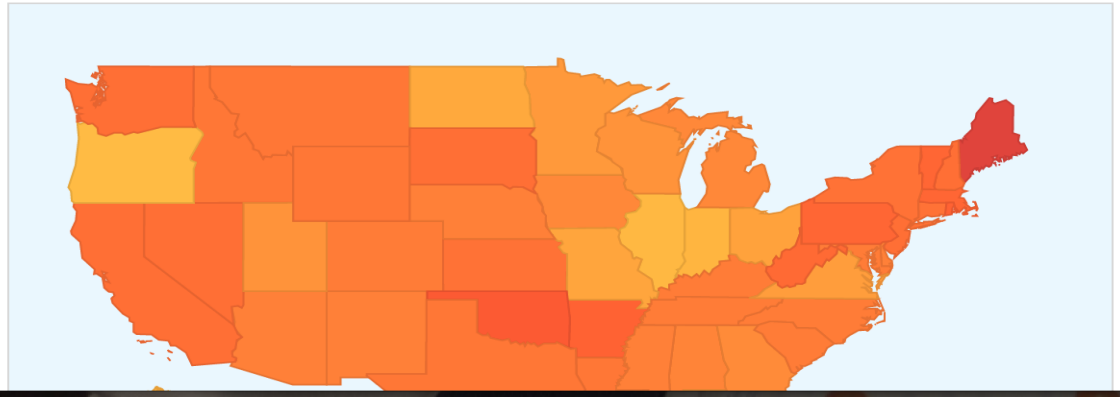
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

2014-2015 Past years



States | Cities (Experimental)



BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data

the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and

http://www.sciencemag.org/content/343/6176/1203