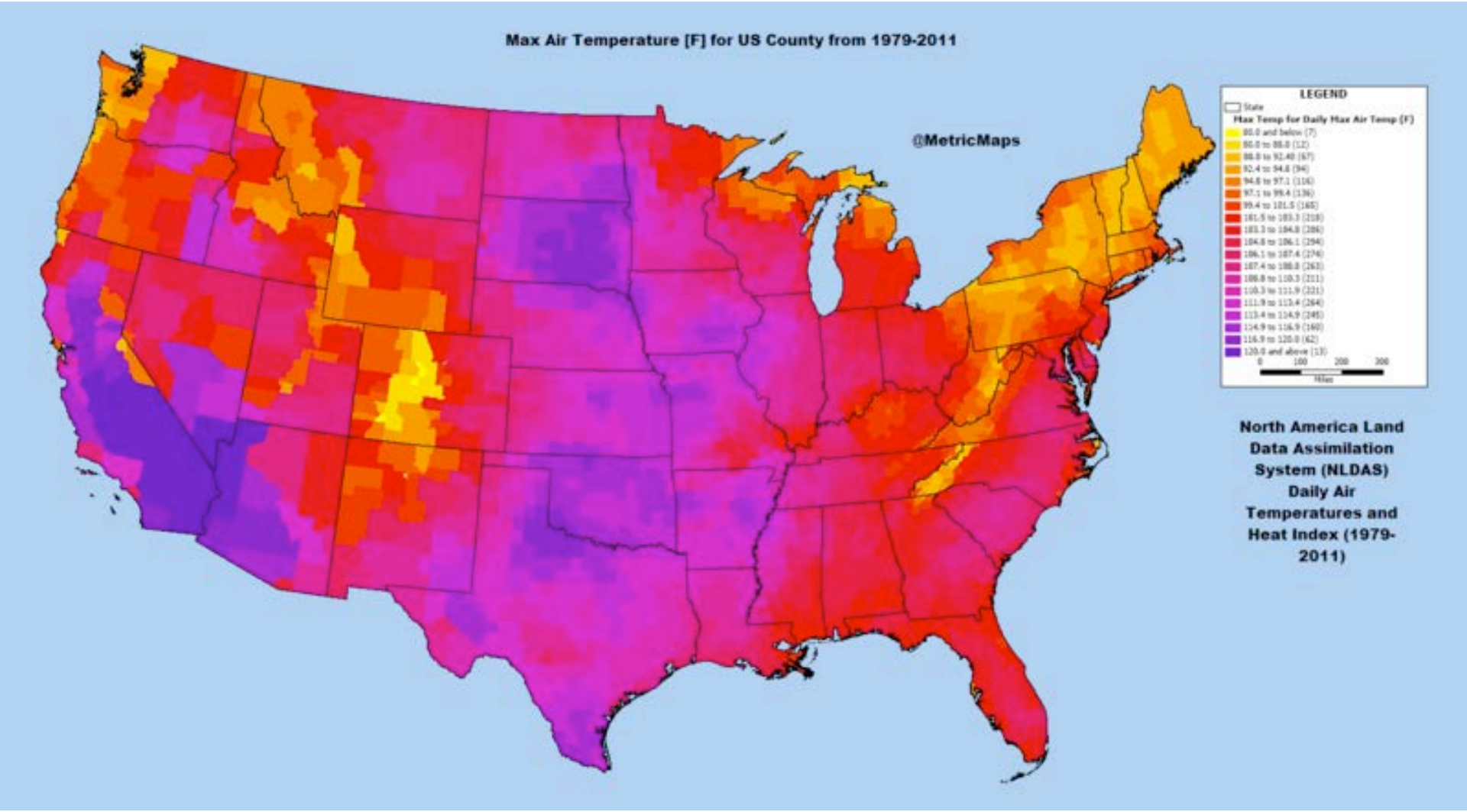
Sampling and Standard Error

To Estimate Mean from a Single Sample

- 1) Choose sample size based on estimate of skew in population
- 2) Chose a random sample from the population
- 3) Compute the mean and standard deviation of that sample
- 4) Use the standard deviation of that sample to estimate the SE
- 5) Use the estimated SE to generate confidence intervals around the sample mean

Back to Sampling Temperatures



Are 200 Samples Enough?

```
temps = getHighs()
popMean = sum(temps)/len(temps)
sampleSize = 200
numTrials = 10000
numBad = 0
for t in range(numTrials):
    sample = random.sample(temps, sampleSize)
    sampleMean = sum(sample)/sampleSize
    se = numpy.std(sample)/sampleSize**0.5
    if abs(popMean - sampleMean) > 1.96*se:
        numBad += 1
print('Fraction outside 95% confidence interval =',
      numBad/numTrials)
```

Fraction outside 95% confidence interval = 0.0511

Are 200 Samples Enough?

```
for t in range(numTrials):
    posStartingPts = range(0, len(temps) - sampleSize)
    start = random.choice(posStartingPts)
    sample = temps[start:start+sampleSize]
    sampleMean = sum(sample)/sampleSize
    se = numpy.std(sample)/sampleSize**0.5
    if abs(popMean - sampleMean) > 1.96*se:
        numBad += 1
print('Fraction outside 95% confidence interval =',
      numBad/numTrials)
```

Fraction outside 95% confidence interval = 0.9367

Has Theory Failed Us?

- No, we have violated a key assumptions
- We did not choose independent random samples
  - Data organized by city
  - Temperatures correlated with city
  - Therefore examples in sample are not independent of each other
  - Obvious here, but can be subtle
- All theoretical results incorporate some assumptions
- These must be checked before applying the theory!